Journal of Computer Science

Review Article

Human Activity Recognition at the Edge: A Systematic Review of Machine Learning Models, Deployment Constraints, and Latency Energy Accuracy Trade offs

Jose Antonio Rojas Guillén[1], Wini Ebelin Quispe Bautista[2] and Arturo Gamarra Moreno[3]

[1]*Department of Engineering, Universidad Tecnológica del Perú, Lima, Peru*
[2]*Graduate School, Universidad Continental, Huancayo, Peru*
[3]*Department of Mechanical Engineering, Universidad Nacional del Centro del Perú, Huancayo, Peru*

Article history
Received: 15-05-2026
Revised: 24-06-2026
Accepted: 29-07-2026

Corresponding Author:
Wini Ebelin Quispe Bautista
Graduate School, Universidad Continental, Huancayo, Peru
Email: 71265716@continental.edu.pe

Abstract: Human Activity Recognition (HAR) has become a key component of intelligent healthcare monitoring, assisted living, sports analytics, smart environments, and wearable computing. However, the migration of HAR models from cloud-centered architectures to edge, embedded, mobile, and wearable platforms introduces constraints related to computational capacity, memory footprint, battery consumption, inference time, and real-time responsiveness. This study presents a systematic literature review of edge-based HAR research, focusing on machine learning models, deployment constraints, and latency energy accuracy trade-offs. Following a PRISMA-based process, 78 studies published between 2021 and 2026 were analyzed from Scopus, Web of Science, IEEE Xplore, and DBLP. The findings show that accelerometers, gyroscopes, inertial measurement units, smartphones, smartwatches, and wearable sensors are the dominant data sources for HAR at the edge. CNN, LSTM, CNN-LSTM, Random Forest, Support Vector Machines, Transformer-based models, and TinyML approaches are frequently used for recognizing daily activities, clinical movements, sports actions, and context-aware behaviors. The results also indicate a gradual shift from accuracy-centered evaluation toward deployment-aware assessment, incorporating latency, inference time, energy consumption, power usage, model size, and memory footprint. Model compression, quantization, pruning, lightweight architectures, feature selection, and microcontroller-based deployment emerge as central strategies for improving edge performance. Nevertheless, the evidence remains fragmented because many studies do not report comparable hardware-level measurements. This review contributes a structured synthesis of edge-based HAR and identifies open challenges related to reproducibility, benchmark standardization, real-world validation, cross-device generalization, and balanced evaluation of latency energy accuracy trade-offs.

Keywords: Human Activity Recognition, Edge Computing, Machine Learning

Introduction

Human Activity Recognition (HAR), once seen primarily as a pattern-recognition task, has become a critical function within intelligent computing systems. Beyond identifying basic activities such as sitting, standing, walking, running, or lying down, HAR now supports healthcare monitoring, assistive living, sports analytics, home automation, human robot interaction, vehicle behavior assessment, workplace safety, and contextual awareness systems. This expansion has been enabled by the growing availability of wearable devices, smartphones, smartwatches, Inertial Measurement Units (IMUs), single-board computers, microcontrollers, IoT-enabled Wi-Fi sensing devices, and edge computing platforms. As a result, HAR now sits at the intersection of sensing technology, artificial intelligence, machine learning, embedded intelligence, and real-time decision-making (Elsts and McConville, 2021; Kim, 2026; Mekruksavanich and Jitpattanakul, 2025; Rivas-Caicedo et al., 2025).

Historically, HAR systems were commonly designed under cloud-centered or offline assumptions, where sensor data were collected locally and then transmitted to remote servers or desktop-grade environments for training, inference, or post-processing. Although these architectures enabled the use of complex machine learning and deep learning models and supported the development of large-scale benchmarks, they are not always suitable for real-world applications that require timely response, privacy preservation, reduced network dependency, continuous monitoring, and energy-efficient operation. This limitation is especially relevant in fall detection, rehabilitation tracking, clinical activity monitoring, wearable assistance, and on-device fitness feedback, where recognition must occur close to the end user with minimal delay and reasonable battery life (Fan et al., 2026a; Mani et al., 2023; Roh et al., 2021; Yun et al., 2023).

In response to these constraints, edge-based HAR has gradually shifted data processing, inference, and decision support closer to the user, moving these functions from distant cloud servers to wearable, mobile, embedded, or nearby edge devices. This transition is reflected in recent studies that have implemented or evaluated HAR systems on smartphones, smartwatches, Raspberry Pi boards, NVIDIA Jetson modules, ESP32 and ESP32S3 microcontrollers, STMicroelectronics STM32 boards, FPGA architectures, and other embedded platforms (Elsts and McConville, 2021; Karakish et al., 2025; Thottempudi et al., 2024; Ajili and Hara-Azumi, 2022; Yang et al., 2023). More than a simple change in computing location, this movement shows a broader transformation in the way HAR systems are designed: Recognition models are now expected not only to achieve high accuracy, but also to be deployable, responsive, resource-aware, and suitable for constrained hardware environments (Gomez-Carmona et al., 2022; Lalapura et al., 2024; Thottempudi et al., 2024; Yamane et al., 2025).

As a result, the central challenge is no longer limited to identifying the model with the highest classification accuracy; it also involves understanding how that accuracy is affected by practical deployment constraints such as memory footprint, inference latency, battery consumption, computational cost, and hardware availability. Traditional machine learning models, including support vector machines, random forests, k-nearest neighbors, decision trees, and multilayer perceptrons, remain relevant because they may offer lower computational cost in constrained settings (Elsts and McConville, 2021; Gomez-Carmona et al., 2022; Karakish et al., 2025). At the same time, convolutional neural networks, long short-term memory networks, gated recurrent units, CNN-LSTM hybrids, residual networks, attention mechanisms, spiking neural networks, and Transformer-based approaches have improved the recognition of spatial, temporal, and multimodal patterns from sensor data streams, although their complexity can limit deployment on wearables and embedded devices (Huang et al., 2023; Kim and Kim, 2025; Lalapura et al., 2024; Li et al., 2023; Yang et al., 2023).

The need for lightweight deployment aware HAR has motivated researchers to investigate model compression and/or quantization, pruning, knowledge distillation, feature selection, transfer learning, reduced sampling rates and hardware acceleration as methods to reduce both the memory and computational cost of a model with good recognition performance (Elsts and McConville, 2021; Fan et al., 2026b; Hayajneh et al., 2024; Ajili and Hara-Azumi, 2022; Yamane et al., 2025). Other contributions propose distributed, partitioned, or cascaded architectures in which computation is allocated across wearable devices, edge nodes, or lightweight model components to reduce unnecessary processing while maintaining satisfactory performance (Gomez-Carmona et al., 2022; Rivas-Caicedo et al., 2025). Together, these studies show that edge-based HAR requires an integrated design perspective that considers sensor modality, model architecture, deployment environment, optimization strategy, and performance metrics simultaneously.

Although this area has made considerable progress in recent years, research in this field continues to be very fragmented. There are many examples of studies that report results with an emphasis placed primarily on the accuracy of the classification system as a measure of the overall performance of their method, but little to no comparative information regarding the other important aspects of their methods such as latency, energy use, model size, memory use, hardware validation, battery life, or cross-platform reproduction (Fatima et al., 2024; Huang et al., 2023; Lalapura et al., 2024; Zaoui et al., 2025). The lack of comparability also applies to the way researchers conceptualize the feasibility of edge computing for various applications. Some researchers define edge feasibility in terms of wearable distributed recognition systems, local processing systems, adaptive models based on edge devices, low power systems inspired by recurrent neural networks, or neuromorphic networks, whereas other researchers define edge feasibility through sensor data collected via CSI, Wi-Fi, or other means of embedded sensing, and/or real-time adaptations in response to changes in the deployment environment (Mohamed and Martinez-Hernandez, 2023; Rivas-Caicedo et al., 2025; Shen et al., 2022; Singh et al., 2026).

A further difficulty concerns the diversity of sensing modalities and application domains. Edge-based HAR is no longer limited to accelerometers and gyroscopes; recent studies have used multimodal inertial sensors, air-pressure sensors, strain sensors, conductive fabric-based sensors, Wi-Fi Channel State Information, ambient sound, video input, and flexible wearable devices (Haghi et al., 2023; Kim, 2026; Lee et al., 2025; Mani et al., 2023; Viswanathuni et al., 2025). While the increased diversity in the HAR modality will allow for a greater number of potential applications of the HAR systems, this increase in the number of modalities makes comparing HAR systems difficult due to differing sampling rates, data volumes, types of noise present, pre-processing steps, privacy considerations, environmental dependencies, and hardware dependent application feasibility. For example, Wi-Fi based HAR is going to have many more hardware and environmental dependencies than IMU based models that are run on a smart watch or micro controller based Tiny ML implementations (Kim, 2026; Viswanathuni et al., 2025; Thottempudi et al., 2024).

For these reasons, a comprehensive synthesis of edge-based HAR from a deployment-aware standpoint is needed. Previous studies have provided valuable models, datasets, platforms, and optimization techniques; however, the field still lacks an integrated review focused specifically on machine learning models, deployment constraints, and latency energy accuracy trade-offs. Such a synthesis is important because the future of HAR depends not only on improving recognition performance, but also on designing systems that are lightweight, efficient, responsive, reproducible, adaptable, and feasible outside controlled experimental testbeds.

Therefore, this study systematically reviews recent research on Human Activity Recognition at the edge, with particular attention to machine learning models, deployment constraints, and latency energy accuracy trade-offs. To address the identified gap, the review analyzed 78 edge-based HAR studies published between 2021 and 2026 and retrieved from Scopus, Web of Science, IEEE Xplore, and DBLP. The analysis focused on how machine learning models, sensing configurations, edge deployment platforms, optimization strategies, and performance metrics are connected in the design and evaluation of deployable HAR systems. Accordingly, the study was guided by the following research question.

RQ: How have machine learning models been designed, optimized, and evaluated for Human Activity Recognition at the edge, and what approaches are being used to address latency energy accuracy trade-offs under deployment constraints?

This review makes three main contributions. First, it organizes recent literature on edge-based HAR by sensor type, dataset, application domain, machine learning model, hardware platform, and deployment constraint. Second, it synthesizes how recent studies evaluate performance beyond accuracy, with special emphasis on latency, inference time, energy consumption, power usage, model size, memory footprint, and computational cost. Third, it identifies open challenges related to reproducibility, benchmark standardization, real-world validation, cross-device generalizability, on-device adaptation, privacy-preserving federated learning, and consistent reporting of latency energy accuracy trade-offs.

Background and Related Work

Human Activity Recognition: Concept and Scope

HAR is defined as the process of determining what type of movement or actions an individual performs through the analysis of generated sensor data. Historically, HAR was applied primarily to determine which types of basic activities were performed including walking, sitting, standing, lying down, running, climbing stairs, etc., as well as identifying simple hand gestures. However, modern research has demonstrated that the field of HAR has expanded significantly beyond simply classifying basic daily living activities. With advancements in technology and sensing capabilities there are now applications for HAR in clinical monitoring, fall detection, rehabilitation, geriatric health and wellness services, home automation/energy management, athletic performance tracking and analysis, workplace safety, agricultural harvesting assistance, human-robotic collaboration, vehicle operation tracking and monitoring, as well as developing personalized health and wellness plans (Deeptha et al., 2024; Gomez et al., 2026; Kim and Kim, 2025; Malche et al., 2026; Mekruksavanich and Jitpattanakul, 2025; Yamane et al., 2025; Yun et al., 2023).

The ability to expand into these additional areas has been driven largely by increased access to sensing technologies. Many of the reviewed studies employed accelerometers, gyroscopes, magnetometers, Inertial Measurement Units (IMUs), mobile phones, smart watches, wearable bands, flexible sensors, radar, Wi-Fi Channel State Information (CSI) and embedded sensors within other devices (Alam et al., 2021; Guo et al., 2024; Kim, 2026; Mani et al., 2023; Viswanathuni et al., 2025; Rivas-Caicedo et al., 2025). Of particular interest among these sensing modalities are IMU's and wearable-based sensors due to their cost-effective nature, non-invasive characteristics and ability to provide user-centric data capture. For example, many researchers have employed wearable based-sensing platforms to identify various aspects of daily living activities (e.g. household chores), clinical movements (e.g. gait patterns), exercise patterns (e.g. yoga poses) and body postures (Cavalcante et al., 2023; Fan et al., 2026a; Gonçalves et al., 2024; Haghi et al., 2023; Rivas-Caicedo et al., 2025). As can be seen in this prior research HAR represents a multidisciplinary research area that incorporates machine learning algorithms, embedded systems development, pervasive computing principles, signal processing techniques and human-centered intelligent system design.

One characteristic that differentiates today's HAR approaches from those developed previously is the shift from generalized activity labeling to context-dependent recognition. While some researchers continue to develop systems designed to label users' activities generically (i.e. "sitting", "standing"), others are recognizing specific domain-related behaviors such as performing venipuncture procedures, adhering to prescribed medication regimens, identifying typical older adult movement patterns, analyzing rehabilitation exercises, monitoring drivers' behavior while driving, evaluating athletes during workouts and monitoring interactions between humans and robots in collaborative work environments (Duarte et al., 2025; Fan et al., 2026b; Mekruksavanich and Jitpattanakul, 2025; Moghbelan et al., 2024; Roh et al., 2021; Yang et al., 2023). By providing systems that recognize activities specifically related to a given application or environment (as opposed to providing generic activity recognition), it is believed that systems will possess greater utility. Additionally, however, the need to develop technically robust systems capable of operating effectively in highly variable environments with varying levels of user variability, device variability and environmental variability presents challenges do not present in earlier HAR approaches (Abuhoureyah et al., 2024; Fu et al., 2021; Mekruksavanich and Jitpattanakul, 2024).

Edge Computing and Embedded Intelligence in HAR

The move from cloud-centric HAR to edge-based HAR is perhaps the most significant shift in recent research. Most traditional HAR pipeline processes involve gathering sensor information, followed by processing that occurs in an offline or cloud environment. While this approach serves well for experimentation and developing models; it does not lend itself well to situations where real-time responses, preserving privacy, continuous monitoring, and reducing reliance on networks are necessary. Edge computing can help address these challenges by relocating the data processing and inference closer to the end-user. This relocation can occur using various means (e.g., wearable devices, smartphones, embedded boards, microcontrollers or neighboring edge nodes) (Elsts and McConville, 2021; Fra et al., 2022; Karakish et al., 2025; Thottempudi et al., 2024; Zhou et al., 2025).

Within edge-based HAR the physical location of the computing infrastructure is now part of the system design. The hardware component (s) used for collecting sensor information will now be considered part of the overall system. As noted above several studies in the examined literature have reported deployments or evaluations of HAR using microcontrollers, smart watches, smartphones, Raspberry Pi, NVIDIA Jetson based systems, embedded boards, FPGA-based architectures and hybrid edge-fog-cloud environments (Elsts and McConville, 2021; Karakish et al., 2025; Novac et al., 2021; Viswanathuni et al., 2025; Rivas-Caicedo et al., 2025; Ajili and Hara-Azumi, 2022; Zaoui et al., 2023). Thus, there appears to be an important methodological shift occurring: Rather than being concerned about if a given model performs accurately during offline evaluation HAR research is increasingly concerned with determining if a model can execute efficiently while operating within hardware constraints.

As previously stated, embedded intelligence is especially important for health-related and safety-critical applications. For example, delay in recognizing a fall could reduce the utility of a fall-detection system. Rehabilitation support systems for elderly patients and clinical activity tracking require timely recognition. Therefore, a recurring theme in many HAR studies over the past few years has been to achieve real-time or near-real-time inference capabilities (Deeptha et al., 2024; Malche et al., 2026; Moghbelan et al., 2024; Noor and Park, 2024; Thottempudi et al., 2024). Additionally, to ensure that wearable systems can provide long term usage they must conserve battery life, minimize communications overhead and limit their computation demands. All of which are factors making HAR at the edge a deployment-focused problem rather than simply an algorithm focused problem.

TinyML is an emerging area of interest that supports this transition. Through its ability to enable machine learning models to operate on microcontrollers and highly resource constrained devices TinyML provides a potential pathway toward low power and autonomously functioning HAR systems (Hayajneh et al., 2024; Lamaakal et al., 2026; Zhou et al., 2025). However, TinyML also creates a necessity for small footprint architectures, minimizing model size and optimizing for both memory considerations and specific hardware needs. Therefore, TinyML represents an additional aspect demonstrating that edge-based HAR requires designing all aspects of the system simultaneously including sensor inputs, model designs, hardware selection and optimization techniques. Recent edge-oriented HAR studies also demonstrate that deployability is affected not just by the degree of personalization, adapting to new users or context and utilizing lightweight embedded pipelines but also through on-device personalization, embedded Wi-Fi sensing and adaptive inference strategies (Alzubi et al., 2025; Patil et al., 2026; Yen et al., 2021). These methods also create the need for evaluating memory utilization, execution times, ensuring adequate protection of individual's privacy and the compatibility of the hardware components involved in a systematic manner.

Machine Learning and Deep Learning Models for HAR

Machine learning continues to play a key role in HAR since it has the capability to effectively process sensor signal noise, high-dimensional (multidimensional) nature of the sensor signals, and their temporal structure. The classical machine learning algorithms including Support Vector Machines, Random Forests, k-Nearest Neighbors, Decision Trees, and Multilayer Perceptrons still represent the majority of HAR literature reviewed above, specifically when there is an emphasis on interpretation, simplicity in terms of computational resources needed, and ease of deployment (Alam et al., 2021; Elsts and McConville, 2021; Guo et al., 2024; Mani et al., 2023; Saeed et al., 2021; Yamane et al., 2025). While these models may be competitive in cases where hand-crafted features are necessary due to limitations in training sets available, or when constraints exist due to hardware limitations, deep learning models dominate many modern studies of HAR primarily due to their potential to automatically generate complex spatial and temporal representations from raw or minimally preprocessed sensor data. CNNs are commonly employed to identify local patterns within windowed sensor data, while RNNs (LSTMs and GRUs), which allow for the modeling of temporal dependencies inherent in sequences of activity data, are also widely employed (Athota and Sumathi, 2022; Huang et al., 2023; Jameer and Syed, 2023; Lalapura et al., 2024; Zhou et al., 2025). CNN-LSTM hybrids are also frequently encountered as well since they allow for both spatial feature extraction and temporal sequence modeling of sequential activity data (Fan et al., 2026a; Jameer and Syed, 2023; Rivas-Caicedo et al., 2025; Zhou et al., 2025).

In addition to these two primary architectures mentioned previously, recent studies demonstrate an increasing number of researchers exploring alternative architectures. Spiking neural networks and other forms of neuromorphic approaches have been proposed for reducing power consumption and improving applicability for edge computing and IoT applications (Agarwal, 2025; Fra et al., 2022; Li et al., 2023; Tan et al., 2021). Additionally, transformer-based models and attention mechanisms have been proposed to enable longer range dependency modeling and improved representation learning for sensor-based HAR systems (Kim, 2026; Lamaakal et al., 2026; Li et al., 2023). Furthermore, graph neural networks, knowledge distillation techniques and cross-modal learning have been applied in various ways, particularly for multi-sensor or on-device HAR systems (Fan et al., 2026a; Gonçalves et al., 2024; Shin et al., 2025). All of these new developments suggest that the field is trending towards developing specialized architectures that attempt to strike a balance between recognition performance and deployability.

Similar tendencies are observed at the model development level as several model development related contributions propose architectures that seek to maintain recognition performance while significantly decreasing the costs associated with deploying those architectures. Examples include models that are optimized for embedded inference, adaptive or hybrid learning strategies, and compressed neural network approaches designed to operate under extremely constrained memory, processing or energy budgets (Armenta-Garcia et al., 2025; Bahador et al., 2021; Bibbò et al., 2025). Thus, the recent literature supports the notion that the choice of architecture in HAR should no longer be solely evaluated using predictive accuracy metrics; however, it should also consider the suitability of the chosen architecture for real-time and resource constrained executions.

Model complexity presents a critical trade-off. Deep models can provide better recognition performance; however, they typically require higher amounts of memory, greater computational loads, increased latency and larger amounts of energy. As such, comparisons among classical models and deep learning architectures will continue to be beneficial. For example, HAR evidence from microcontrollers suggests that simpler models can achieve comparable levels of recognition performance with much lower resource usage than deep models in certain situations (Elsts and McConville, 2021); this implies that the optimal model for edge-based HAR is not necessarily the most complex one; rather it is the one that achieves the best balance between recognition performance, efficiency and feasibility of deployment.

Deployment Constraints in Edge-Based HAR

Edge-based HAR differs from cloud-based HAR in that edge-based systems typically have limited processing power, limited memory, limited battery life, small amounts of storage space, and varying connectivity capabilities. The characteristics of these limitations will determine many aspects of both the model chosen and the sensors used (sampling rate), and will further impact how information is processed and fed back into the system (Elsts and McConville, 2021; Lattanzi et al., 2022; Suwannarat and Kurdthongmee, 2021; Yamane et al., 2025); therefore, the deployment characteristics are not just technical implementation issues; they drive both the design and the testing of the overall HAR system.

The first constraint is processing capability. Smart watches, wearable devices and microcontrollers can not run large deep learning models at high speeds. Studies which focus on micro-controllers and embedded HAR show that in addition to the model execution itself, consideration should be made to the execution time based on processing speed, amount of RAM available, the number of bits per sample, etc., as well as any applicable hardware accelerators (Chang et al., 2022; Elsts and

McConville, 2021; Novac et al., 2021). Devices like Raspberry Pi or Jetson boards offer better processing ability than those above mentioned, however these too need to be optimized for real-time operation or continuous monitoring (Debnath et al., 2025; Viswanathuni et al., 2025; Rivas-Caicedo et al., 2025; Ajili and Hara-Azumi, 2022).

The second constraint is power consumption. Systems designed for wearables should allow for extended use before requiring recharge. Consequently, energy efficiency requires considerations in addition to the model architecture including, sampling rates of sensors, communication rates, methods for compressing data locally and whether or not inference occurs locally vs. remotely (Mahdi et al., 2024; Yamane et al., 2025). Lowering the sampling rate will reduce data transmission requirements and lower energy usage. However, lowering sampling rates may potentially lower recognition accuracy due to loss of temporal patterns. Thus, there exists an inherent tradeoff between resource conservation and reliable activity recognition.

Lastly, the memory footprint and model size are also constraints on the deployability of HAR models. Models that perform well in offline evaluations may fail to meet low memory device requirements for storing/running the model. Recent studies employ techniques such as model compression, quantization, pruning, lightweight architectures, knowledge distillation, feature selection, and hardware aware optimizations to achieve acceptable performance levels using less resources to mitigate this problem (Contoli and Lattanzi, 2023; Fan et al., 2026b; Gonçalves et al., 2024; Kumari et al., 2024; Novac et al., 2021; Shin et al., 2025).

Real-time responsiveness is one final constraint. Applications where timely inference is required (e.g., fall detection, driver monitoring, patient tracking, human robot collaboration) may have a significant impact on the value/utility of the system as well as safety to users. Consequentially, inference delays are increasingly being measured and/or reported alongside traditional metrics such as classification accuracy (Fan et al., 2026a; Malche et al., 2026; Mekruksavanich and Jitpattanakul, 2025; Thottempudi et al., 2024; Yang et al., 2023). Unfortunately, the measurement practices are not consistent across all studies. While some studies measure response times at various hardware layers, other reports do little more than state what percentage of samples were correctly classified and/or make qualitative statements regarding efficiencies.

Latency Energy Accuracy Trade-Offs

Accuracy is essential for edge HAR systems as well as latency, processing time, energy consumption, power usage, memory footprint, model size and all other factors (Elsts and McConville, 2021; Gomez-Carmona et al., 2022; Lalapura et al., 2024; Zhou et al., 2025). It is possible to develop an extremely accurate model; however, a model will not be practical if it consumes too much energy or has long response times when deployed on resource limited devices (Lalapura et al., 2024; Zhou et al., 2025).

There are many examples of the trade-offs in the literature from various perspectives. Wearable distributed architectures can help to minimize local processing burdens and enhance responsiveness; however, there are several questions regarding where computations should take place and how sensor level output should be combined at each site (Rivas-Caicedo et al., 2025). Optimizing sampling frequency can minimize both energy use and amount of data collected; however, the process must be optimized to prevent loss of discriminative motion data (Yamane et al., 2025). Model compression and/or quantization methods allow for the use of large complex neural network models on limited resource devices; however, they can cause significant losses in accuracy if the model is not appropriately optimized during training (Contoli and Lattanzi, 2023; Novac et al., 2021; Shin et al., 2025). Performance can be transferred from large teacher networks to small student networks via knowledge distillation; however, the performance transfer is dependent upon the characteristics of the target hardware and the complexity of the activity recognition problem (Fan et al., 2026a; Gonçalves et al., 2024).

Another area of research addressing the latency-energy-accuracy tradeoff includes neuromorphic/spiking approaches. Event driven computing/neuromorphic computing/biologically inspired mechanisms have been used to decrease energy demands while still allowing useful recognition performance (Agarwal, 2025; Fra et al., 2022; Li et al., 2023; Mazzieri et al., 2026). Hardware acceleration via FPGAs or hybrid CPU/FPGA architectures can also improve the efficiency of inference; however, hardware acceleration can add complexity to implementation and limit portability among devices (Ajili and Hara-Azumi, 2022). The results from this body of work indicate that the latency-energy-accuracy tradeoff cannot be addressed solely by optimizing algorithms; rather, both hardware aware and deployment aware assessments are necessary.

Sensor modality is also an influence on the tradeoff. Inertial sensors typically produce compact time series data streams which are ideal for wearables. While Wi-Fi CSI and Radar based HAR can provide non-contact recognition, they create unique challenges associated with signal variability, environment dependency, placement of the device, processing complexity etc. (Abuhoureyah et al., 2024; Kim, 2026; Mazzieri et al., 2026; Viswanathuni et al., 2025); camera/video based approaches can collect rich motion data; however, they generally require significantly higher amounts of processing resources and create additional issues regarding privacy (Debnath et al., 2025; Noor and Park, 2024).

Critical Appraisal and Identified Research Gap

The most recent research papers have demonstrated an obvious progression within edge-based Human Activity Recognition (HAR). The edge-based HAR research community is transitioning from traditional sensor-based classification to more advanced, intelligent systems, including embedded intelligence, TinyML, real-time inference, multimodal sensing, model compression, neuromorphic processing, and hardware-aware deployment (Agarwal, 2025; Hayajneh et al., 2024; Lamaakal et al., 2026; Li et al., 2023; Shin et al., 2025; Zhou et al., 2025). As such, the advancement of edge-based HAR represents a technically meaningful development towards making HAR feasible in practice in health care, smart-environmental applications, mobile computing, wearable systems, and human-centric cyber-physical systems.

Although there has been a major advance in the area of edge-based HAR as noted above; however, there are some limits remaining. First, while accuracy is currently a dominant factor in the way researchers evaluate their systems; although many studies reference latency, power consumption, memory usage, or model size; these factors are not always evaluated using similar levels of rigor or presented in identical formats. Thus, there is a disconnect between the potential performance of a model versus its feasibility for deployment in an edge environment (Fan et al., 2026a; Lalapura et al., 2024; Thottempudi et al., 2024). Second, the level of hardware testing and/or validation is highly variable. While some studies develop and deploy their models on specific edge devices (e.g., microcontrollers), smartwatches, or embedded platforms; other studies only propose that they could be deployed on an edge device without conducting fully hardware-level experiments (Elsts and McConville, 2021; Novac et al., 2021; Zhou et al., 2025). Finally, one of the key barriers to reproducibility among studies is based upon differences in data sets, sensor locations, activity labels, sampling rates, pre-processing techniques, and evaluation methodologies (Alam et al., 2021; Mekruksavanich and Jitpattanakul, 2024; Novac et al., 2022; Yamane et al., 2025).

Another limitation related to generalizability exists. Many HAR systems have proven to be sensitive to variations in the user experience (i.e., device position), sensor quality, environmental conditions, and the method of performing activities. Those studies examining issues related to device-position independent sensing, location-independent sensing, transfer learning, adaptive learning, and personalizable recognition demonstrate that generalizability is one of the greatest challenges facing developers of edge-based HAR systems (Abuhoureyah et al., 2024; Fan et al., 2026b; Fu et al., 2021; Mekruksavanich and Jitpattanakul, 2024). This challenge is particularly relevant to the deployment of HAR systems in the "real world" given that there will likely be users, devices, and contextual experiences encountered when deploying a HAR system that were not experienced during the initial development phase.

Therefore, the current state-of-the-art of edge-based HAR indicates that a greater degree of integration is needed to synthesize existing knowledge. Individual studies are providing valuable technical insights into each of the various components of an edge-based HAR system. However, as previously stated the field of edge-based HAR does not possess a unified perspective regarding the interactions between machine learning algorithms used to classify human activities, sensor modalities used to collect those activities, edge platforms onto which those models will be deployed, constraints associated with deployment at the edge (such as latency and energy use) and the optimization strategies employed to reduce the impact of those constraints. The goal of this review is to organize recent research efforts toward developing a more comprehensive understanding of how to design HAR systems with consideration to deployment requirements and identify the conditions required to enable HAR models to transition from high accuracy in laboratory settings toward robust, reproducible, and real-time edge intelligence.

The reviewed literature suggests that edge-based HAR is not focused on isolated technical aspects of activity recognition. Instead, the field is structured around the interaction between sensing modalities, machine learning models, deployment platforms, optimization strategies, and deployment-aware performance metrics. For this reason, Table 1 summarizes the conceptual structure of the edge-based HAR literature and organizes the main analytical areas considered in this review.

As shown in Table 1, edge-based HAR is not limited to activity classification. The literature is organized around the interaction between sensing modalities, learning models, deployment platforms, optimization strategies, and deployment-aware evaluation metrics. This structure supports the need for a system-level synthesis rather than an accuracy-only comparison of HAR models.

HAR deployed at the edge can best be viewed as a system-wide issue (i.e., a "problem") and not simply as a single classification problem. Depending upon many factors such as the use of multiple modalities for sensors; the use of different hardware platforms; and different methods for optimizing performance, even the same model will have different characteristics. For example, a model that works well for recognition performed by a smartphone would likely be unfeasible for recognition performed by a low power micro controller. A highly accurate Deep Learning architecture is typically too large to deploy directly onto a wearable device; therefore, it is necessary to compress or quantize the model prior to its deployment. Therefore, when evaluating these types of models, latency, energy consumption, memory requirements, and accuracy must be considered as interdependent parameters and not evaluated independently.

Table 1: Conceptual Structure of Edge-Based HAR Literature

Conceptual Area	Main Focus in the Literature	Representative Evidence
Human Activity Recognition scope	Recognition of daily activities, clinical movements, elderly monitoring, sports, smart homes, driver behavior, and human–robot collaboration	(Deeptha et al., 2024; Gomez et al., 2026; Kim and Kim, 2025; Malche et al., 2026; Mekruksavanich and Jitpattanakul, 2025; Yamane et al., 2025; Yun et al., 2023)
Sensor and data modalities	Use of accelerometers, gyroscopes, IMUs, smartphones, smartwatches, wearable sensors, Wi-Fi CSI, radar, and flexible sensors	(Alam et al., 2021; Guo et al., 2024; Haghi et al., 2023; Kim, 2026; Mani et al., 2023; Viswanathuni et al., 2025; Rivas-Caicedo et al., 2025)
Edge and embedded deployment	Deployment on wearable devices, smartphones, microcontrollers, Raspberry Pi, Jetson, FPGA, and edge computing platforms	(Elsts and McConville, 2021; Gonçalves et al., 2024; Karakish et al., 2025; Novac et al., 2021; Viswanathuni et al., 2025; Ajili and Hara-Azumi, 2022; Zaoui et al., 2023)
ML/DL models	Use of classical ML, CNN, LSTM, GRU, CNN-LSTM, Transformers, SNNs, TinyML, and compressed models	(Agarwal, 2025; Athota and Sumathi, 2022; Fan et al., 2026a; Gonçalves et al., 2024; Lalapura et al., 2024; Li et al., 2023; Shin et al., 2025; Zhou et al., 2025)
Deployment constraints	Computational capacity, memory footprint, energy consumption, battery life, inference time, and hardware limitations	(Chang et al., 2022; Elsts and McConville, 2021; Lattanzi et al., 2022; Mahdi et al., 2024; Novac et al., 2021; Suwannarat and Kurdthongmee, 2021; Yamane et al., 2025)
Optimization strategies	Quantization, pruning, compression, knowledge distillation, feature selection, lightweight design, hardware acceleration	(Contoli and Lattanzi, 2023; Fan et al., 2026b; Gonçalves et al., 2024; Kumari et al., 2024; Novac et al., 2021; Shin et al., 2025; Trabelsi Ajili and Hara-Azumi, 2022)
Trade-off problem	Balance between accuracy, latency, energy consumption, model size, memory usage, and deployment feasibility	(Agarwal, 2025; Elsts and McConville, 2021; Gomez-Carmona et al., 2022; Lalapura et al., 2024; Rivas-Caicedo et al., 2025; Yamane et al., 2025; Zhou et al., 2025)

Materials and Methods

Research Design

This study employed a Systematic Literature Review (SLR) to synthesize contemporary empirical research on Human Activity Recognition (HAR) at the edge. The review focused on machine learning models, deployment constraints, and latency energy accuracy trade-offs in HAR studies involving edge, embedded, wearable, mobile, IoT, TinyML, or other resource-constrained environments.

The systematic review was conducted following the PRISMA guidelines, which structured the identification, screening, eligibility assessment, and inclusion of studies. To strengthen transparency, reproducibility, and methodological consistency, the review also incorporated principles from evidence-based software engineering in the definition of the search strategy, eligibility criteria, data extraction process, quality assessment, and synthesis procedure. This methodological design was suitable because edge-based HAR is a technologically diverse field, involving different sensor modalities, machine learning architectures, hardware platforms, optimization strategies, validation protocols, and deployment-aware metrics. The study was therefore guided by the following research question.

RQ: How have machine learning models been designed, optimized, and evaluated for Human Activity Recognition at the edge, and how do current studies address latency energy accuracy trade-offs under deployment constraints?

To answer this question, the review examined four complementary dimensions: The scope and application domains of HAR; the machine learning and deep learning models used for activity recognition; the edge, embedded, wearable, mobile, IoT, or TinyML environments in which these models were evaluated or deployed; and the deployment-aware performance indicators reported by the studies, including accuracy, latency, inference time, energy consumption, power usage, memory footprint, model size, computational cost, and hardware feasibility.

Information Sources and Search Strategy

The literature search was conducted in four scientific databases: Scopus, Web of Science, IEEE Xplore, and DBLP. These sources were selected because they provide broad and complementary coverage of the fields connected to edge-based HAR, including computer science, artificial intelligence, machine learning, embedded systems, edge computing, wearable sensing, Internet of Things, and human activity recognition. Scopus and Web of Science were used as multidisciplinary citation databases, IEEE Xplore was included because of its strong coverage of engineering, embedded systems, IoT, and edge computing, and DBLP was used as a specialized computer science bibliographic source to complement retrieval from indexed and conference-oriented venues.

The search covered studies published between 2021 and 2026. This period was selected because it captures recent advances in edge computing, TinyML, wearable intelligence, embedded machine learning, hardware-aware optimization, and real-time HAR. The search strategy was organized around four conceptual blocks: Human Activity Recognition, edge or embedded deployment, machine learning or deep learning models, and deployment-performance metrics.

The general search logic was structured as follows: ("human activity recognition" OR "sensor-based human activity recognition" OR "wearable human activity recognition") AND ("edge computing" OR "edge intelligence" OR "embedded device" OR "embedded system" OR "wearable device" OR "TinyML" OR "microcontroller") AND ("machine learning" OR "deep learning" OR "neural network" OR "CNN" OR "LSTM" OR "transformer" OR "model compression" OR "quantization" OR "pruning") AND ("latency" OR "inference time" OR "energy consumption" OR "power consumption" OR "memory footprint" OR "computational cost" OR "accuracy").

The search strings were adapted to the syntax and search capabilities of each database. Scopus was searched using TITLE-ABS-KEY, Web of Science was searched using TS =, IEEE Xplore was searched through metadata fields, and DBLP was searched using simplified keyword combinations because its search interface is primarily bibliographic and does not support the same field-based structure as Scopus or Web of Science. The final database-specific strategies are shown in Table 2.

Table 2 shows that the search strategy preserved the same conceptual structure across all databases while adapting the syntax to each source. This ensured methodological consistency between Scopus, Web of Science, IEEE Xplore, and DBLP, while allowing each database to be searched according to its own indexing and query structure.

The searches included only English-language publication since 2021 and up to 2026. The searches on Scopus and Web of Science included only research articles. All the search results from IEEE Xplore and DBLP were reviewed for inclusion as either an original research article or a complete research paper (including adequate methodology and technical details). During both retrieval and screening phases, it was confirmed that full-text access was available. All items that were initially identified by the searches in DBLP were further checked against the relevant journal's website or indexed full-text source prior to determining whether they met study inclusion criteria.

PRISMA Procedure

The study selection process followed the PRISMA stages of identification, screening, eligibility, and inclusion. During the identification stage, 4,297 records were retrieved from four scientific databases: Scopus (n = 2,395), Web of Science (n = 1,104), IEEE Xplore (n = 767), and DBLP (n = 31). Before screening, 1,148 duplicate records were removed based on DOI, title, author, and year matching. No records were marked as ineligible by automation tools, and no additional records were removed for other reasons.

Table 2: Search Strategy by Database

Database	Search Strategy
Scopus	TITLE-ABS-KEY (("human activity recognition" OR "sensor-based human activity recognition" OR "wearable human activity recognition") AND ("edge computing" OR "edge intelligence" OR "embedded device*" OR "embedded system*" OR "wearable device*" OR "TinyML" OR "microcontroller*") AND ("machine learning" OR "deep learning" OR "neural network*" OR "convolutional neural network*" OR "CNN" OR "long short-term memory" OR "LSTM" OR "transformer*" OR "model compression" OR "model optimization" OR "quantization" OR "pruning") AND ("latency" OR "inference time" OR "energy consumption" OR "power consumption" OR "memory footprint" OR "computational cost" OR "accuracy")) AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND DOCTYPE(ar) AND LANGUAGE(english)
Web of Science	TS=(("human activity recognition" OR "sensor-based human activity recognition" OR "wearable human activity recognition") AND ("edge computing" OR "edge intelligence" OR "embedded device*" OR "embedded system*" OR "wearable device*" OR "TinyML" OR "microcontroller*") AND ("machine learning" OR "deep learning" OR "neural network*" OR "convolutional neural network*" OR "CNN" OR "long short-term memory" OR "LSTM" OR "transformer*" OR "model compression" OR "model optimization" OR "quantization" OR "pruning") AND ("latency" OR "inference time" OR "energy consumption" OR "power consumption" OR "memory footprint" OR "computational cost" OR "accuracy"))
IEEE Xplore	("All Metadata":"human activity recognition" OR "All Metadata":"sensor-based human activity recognition" OR "All Metadata":"wearable human activity recognition" OR "All Metadata":"sensor-based activity recognition" OR "All Metadata":"wearable activity recognition" OR "All Metadata":"HAR") AND ("All Metadata":"edge computing" OR "All Metadata":"edge intelligence" OR "All Metadata":"edge device" OR "All Metadata":"embedded device" OR "All Metadata":"embedded system" OR "All Metadata":"wearable device" OR "All Metadata":"internet of things" OR "All Metadata":"IoT" OR "All Metadata":"TinyML" OR "All Metadata":"microcontroller") AND ("All Metadata":"machine learning" OR "All Metadata":"deep learning" OR "All Metadata":"neural network" OR "All Metadata":"CNN" OR "All Metadata":"LSTM" OR "All Metadata":"transformer" OR "All Metadata":"model compression" OR "All Metadata":"model optimization" OR "All Metadata":"quantization" OR "All Metadata":"pruning") AND ("All Metadata":"latency" OR "All Metadata":"inference time" OR "All Metadata":"energy consumption" OR "All Metadata":"power consumption" OR "All Metadata":"memory footprint" OR "All Metadata":"accuracy")
DBLP	"human activity recognition" AND ("edge computing" OR "embedded" OR "wearable" OR "TinyML" OR "microcontroller") AND ("machine learning" OR "deep learning" OR "CNN" OR "LSTM")

DOI: 10.3844/jcssp.2026.2590.2616

After duplicate removal, 3,149 records were screened by title and abstract. At this stage, 2,914 records were excluded because of conceptual irrelevance, wrong context, inappropriate methodological focus, or lack of alignment with the research questions. Consequently, 235 reports were sought for full-text retrieval. Of these, 9 reports could not be retrieved because the full text was unavailable, leaving 226 reports for eligibility assessment.

During the eligibility stage, 148 reports were excluded for predefined reasons. These included reports not written in English (n = 12), ineligible publication types after full-text verification (n = 26), papers below the minimum length criterion of 10 pages (n = 17), studies outside the review scope after full-text assessment (n = 72), and reports with insufficient methodological or empirical information (n = 21). Ultimately, 78 studies that satisfied all inclusion requirements were selected for inclusion in the systematic review. A detailed representation of the entire process from identifying study records to determining which studies meet inclusion criteria can be viewed as the flow diagram shown in Fig. 1. In the flow diagram it shows how the initial number of studies were first removed by eliminating duplicates; then studies were narrowed down based on the titles and abstracts of the studies; finally studies were narrowed further to those eligible for full-text retrieval; and ultimately studies were narrowed one last time (final selection) to determine which studies would be included.

As shown in Fig. 1, the identification stage retrieved 4,297 records from Scopus, Web of Science, IEEE Xplore, and DBLP. After removing 1,148 duplicate records, 3,149 records were screened by title and abstract, and 235 reports were sought for retrieval. After excluding 9 unavailable reports and 148 reports that did not meet the eligibility criteria, the final corpus consisted of 78 studies included in the systematic review.

Eligibility Criteria

Eligible study characteristics were established prior to the beginning of a full-text review process. The use of pre-established eligibility criteria was used to minimize variation (selection bias) in the screening process. Studies that applied Machine Learning or Deep Learning Models for Human Activity Recognition (HAR) were eligible for inclusion. Additionally, studies needed to have reported metrics associated with either recognition performance or deployment (e.g., Accuracy, Precision, Recall, F1-Score, Latency, Inference Time, Energy Consumption, Power Usage, Memory Footprint, Model Size, Computational Cost).

Excluded from this systematic review would be those studies outside the publication date range; written in languages other than English; not empirical research (i.e., reviews, surveys, bibliometric studies, etc.); and not directly applicable to Edge Computing (EC), Embedded Systems (ES), Wearable Devices (WD), Mobile Applications (MA), Internet of Things (IoT), Tiny Machine Learning (TinyML), or Resource-Constrained Environments.

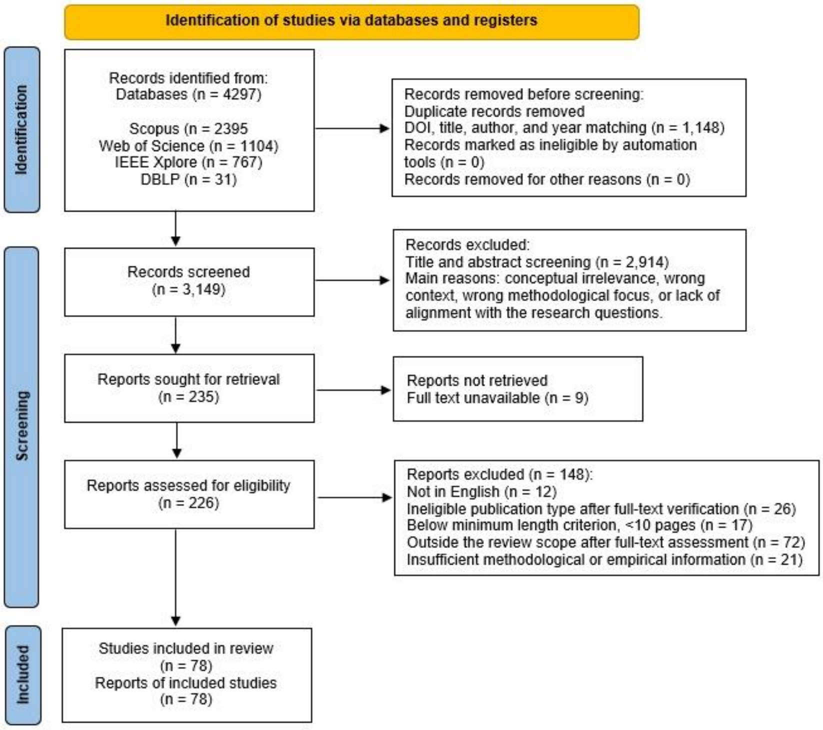

Fig. 1: PRISMA flow diagram of the study selection process

Table 3: Inclusion and Exclusion Criteria

Criterion	Inclusion	Exclusion
Publication year	Studies published between 2021 and 2026	Studies published before 2021
Language	English	Non-English documents
Publication type	Original research articles or full research papers with sufficient methodological detail	Reviews, surveys, bibliometric studies, editorials, book chapters, theses, posters, abstracts, notes, and non-empirical documents
Length	Full papers with sufficient technical detail	Short papers with fewer than 10 pages
Topic relevance	Studies focused on Human Activity Recognition	Studies not focused on HAR
Technical scope	HAR studies using machine learning or deep learning models	Studies without ML/DL-based recognition models
Deployment scope	Edge, embedded, wearable, mobile, IoT, TinyML, or resource-constrained deployment	Cloud-only, offline-only, or purely conceptual studies without edge or deployment relevance
Metrics	Studies reporting recognition or deployment metrics such as accuracy, F1-score, latency, inference time, energy consumption, memory usage, model size, or computational cost	Studies without sufficient performance or deployment-related evidence
Accessibility	Full text available for analysis	Full text not retrievable

Also excluded from this review would be short manuscripts with less than ten pages of content since these typically lack sufficient detail regarding methodology, experiments and/or deployment specifics to support systematic data extraction.

The inclusion and exclusion criteria used to determine the suitability of the studies are presented in Table 3. These criteria were established prior to the full-text evaluation to ensure consistency in the selection process and to ensure that the final corpus aligned with the review's focus on aspects related to deployment.

As shown in Table 3, this study used predefined eligibility criteria to ensure that retained studies addressed HAR with machine learning or deep learning models in edge, embedded, wearable, mobile, IoT, TinyML, or other resource-constrained environments.

Studies were excluded when they were cloud-only, offline-only, purely conceptual, non-empirical, too short to provide sufficient technical detail, inaccessible, or unrelated to deployment-aware HAR. Therefore, the final corpus contained studies with sufficient technical and empirical information on HAR deployment rather than a general collection of HAR literature.

Treatment of Hybrid Cloud Edge and Simulation-Based Studies

Additional screening rules were applied to distinguish edge-relevant HAR studies from cloud-only, offline-only, or purely conceptual studies. Hybrid cloud–edge, fog–edge, and cloud fog edge studies were retained only when the edge layer performed a clear functional role, such as local sensing, preprocessing, local inference, communication reduction, latency reduction, privacy preservation, or deployment-aware evaluation. Studies were excluded when model training, inference, and evaluation were conducted exclusively in remote cloud, offline, or desktop-grade environments without explicit relevance to edge, embedded, wearable, mobile, IoT, TinyML, or resource-constrained deployment.

Simulation-based research was not always excluded. The simulation had to have constraints on deployment (i.e., memory limitations; inference/processing speed; communication costs; inference time; energy/power requirements; model size/feasibility of deployment on a device) for it to be included. Studies which used pure concept or cloud level simulations and/or did not provide empirical or relevant data about how well they performed in terms of deployment-related constraints were also excluded. The end result is that the studies in the final corpus are those related to Har studies which specifically focus on the deployments at the edge where resources can be severely limited compared to those found under typical offline/desktop/cloud computing environments.

Data Extraction Process

A structured extraction matrix was used for extracting data. Every study that was included in the systematic review received an ID based on its position in the ordered reference list. The extraction matrix contained information regarding:

(1) Bibliographical details
(2) Technical specifications
(3) Methodology
(4) Deployment
(5) Performance indicators
(6) Optimization techniques
(7) Findings
(8) Limitations
(9) Future research avenues

Information that was extracted from each study were: Bibliographical data, HAR scope, sensors/devices, dataset (s), models, deployment environment, deployment restrictions, performance indicator (s); optimization strategy (ies); primary findings; limitation (s); and future research avenue (s).

Use of this structure provided a basis for comparing studies along technical and methodological aspects instead of treating them as independent entities. Extraction of data also allowed researchers to identify repeated use of sensors, selected models, deployed hardware, optimization technique (s), and results obtained with regards to latency- energy- accuracy evaluations.

Quality Assessment and Reliability

The methodological quality of the included studies was assessed using a six-item appraisal rubric designed for a systematic literature review involving methodologically heterogeneous HAR studies. The purpose of this assessment was to determine whether each study provided sufficient methodological, conceptual, technical, and empirical detail for reliable data extraction and thematic synthesis.

Six quality assessment criteria were applied:

QA1 : Clear objective, research problem, or technical contribution
QA2 : Methodological transparency regarding dataset, model, experimental design, or evaluation procedure
QA3 : Relevance to Human Activity Recognition or a closely related activity-recognition task
QA4 : Relevance to edge, embedded, wearable, mobile, IoT, TinyML, or resource-constrained deployment
QA5 : Reporting of recognition-performance metrics such as accuracy, precision, recall, or F1-score
QA6 : Reporting or discussion of deployment-aware metrics such as latency, inference time, energy consumption, power usage, memory footprint, model size, computational cost, or hardware constraints

Each criterion was scored using a binary scale, where 1 indicated that the criterion was sufficiently addressed and 0 indicated that the criterion was absent or insufficiently addressed. The total quality score for each study therefore ranged from 0 to 6 points and was calculated as follows:

$$QS_{\mathrm{i}} = \sum_{j=1}^{6} QA_{\mathrm{ij}} \quad (1)$$

Where QS_i represents the total quality score of study i, and QA_{ij} represents the score assigned to criterion j. Since six criteria were assessed using a binary scale, the maximum possible score was 6 points.

To facilitate interpretation, a normalized Methodological Quality Index was calculated using the following expression:

$$MQI_{\mathrm{i}} = \left(\frac{QS_{\mathrm{i}}}{6}\right) \times 100 \quad (2)$$

Thus, studies scoring 6 obtained an MQI of 100.00%, studies scoring 5 obtained an MQI of 83.33%, and studies scoring 4 obtained an MQI of 66.67%. Studies scoring 6 were classified as highly relevant and methodologically strong. Studies scoring 5 were classified as strong but with minor reporting limitations. Studies scoring 4 were considered eligible for inclusion only when they remained directly relevant to the review scope but provided less complete deployment-level information. Studies scoring below 4 were excluded from the final synthesis.

Consistency in the screening and quality evaluation process was increased by applying pre-defined eligibility standards and a six-point quality assessment rubric, with data from two investigators being cross-checked. Following this, the first classifications, extracted data, and quality scores were reviewed as a team. When conflicts or ambiguity were found, the two investigators went back into the full text of the original study and compared the evidence that supported every score. They solved disagreements via collaboration and consensus. The method of comparing the two investigators' interpretations of the same material, provided greater reliability, transparency and accountability for all decisions on the studies included and their quality.

Table 4 presents the quality score distribution of the 78 included studies after consensus deliberation. The results show that all included studies exceeded the minimum methodological quality threshold. Specifically, 71 studies obtained the maximum score of 6 points, equivalent to an MQI of 100.00%, while 7 studies obtained 5 points, equivalent to an MQI of 83.33%. No included study scored 4 points or lower.

Data Synthesis

A thematic synthesis approach was used because the included studies were highly heterogeneous in their methodological and technical characteristics. The studies differed in terms of datasets, sensor locations, sampling rates, activity labels, model architectures, hardware platforms, experimental conditions, validation protocols, and deployment metrics.

Table 4: Quality Score Distribution of Included Studies

Quality score	MQI (%)	Interpretation	Number of studies	Percentage
6	100.00	Highly relevant and methodologically strong	71	91.00%
5	83.33	Strong, with minor reporting limitations	7	9.00%
4	66.67	Eligible threshold, but not observed among included studies	0	0.00%
Total	—	Included studies	78	100.00%

For this reason, a quantitative meta-analysis was not appropriate, since the studies did not report comparable effect sizes or standardized measurements of latency, energy consumption, memory footprint, model size, or hardware performance.

The synthesis was organized around nine analytical categories that captured the main technical and methodological dimensions of edge-based HAR: Application domains and target activities; sensor modalities and data sources; datasets used for training and evaluation; machine learning and deep learning models; edge, embedded, wearable, mobile, and IoT deployment platforms; deployment constraints and hardware limitations; performance metrics; optimization strategies; and reported research gaps or future directions.

This organization allowed the review to move beyond a descriptive summary of the included studies. Rather than simply listing models, sensors, datasets, or accuracy values, the synthesis provided an integrated interpretation of how edge-based HAR systems are designed, evaluated, and constrained in real deployment contexts. By connecting technical, methodological, and deployment-related evidence, the review identified patterns that would not be visible through accuracy-based comparison alone. In particular, the synthesis examined how recognition performance interacts with hardware feasibility, latency, energy consumption, memory requirements, model size, and optimization strategies. In this way, the review contributes a deployment-aware understanding of HAR at the edge and reinforces the need to evaluate future HAR models as complete edge-intelligence systems rather than as isolated classification algorithms.

Results

Study Selection and Temporal Profile of the Included Studies

The systematic selection process resulted in a final corpus of 78 studies published between 2021 and 2026. The included studies were retained because they addressed Human Activity Recognition (HAR) using machine learning or deep learning approaches and provided evidence related to edge, embedded, wearable, mobile, IoT, TinyML, or resource-constrained deployment.

Table 5 presents the temporal profile of the included studies. In addition to reporting the number and percentage of studies per year, the table identifies the main deployment-oriented emphasis observed in each period and provides representative evidence from the reviewed corpus. This helps characterize not only the temporal concentration of the evidence, but also the progressive movement of edge-based HAR toward hardware-aware, energy-aware, adaptive, and deployment-aware evaluation.

Table 5: Temporal Profile and Deployment-Oriented Evolution of Included Studies

Year	Number of studies	Percentage (%)	Cumulative percentage (%)	Main deployment-oriented emphasis	Representative evidence
2021	13	16.7	16.7	Early edge-HAR feasibility, smartphones, wearable sensing, microcontroller readiness, and low-power implementation	Elsts and McConville (2021); Knez et al. (2021); Roh et al. (2021); Suwannarat and Kurdthongmee (2021); Yen et al. (2021)
2022	8	10.3	27	Embedded and hardware-aware HAR, tiny wearable devices, FPGA acceleration, and neuromorphic/low-power approaches	Chang et al. (2022); Fra et al. (2022); Lattanzi et al. (2022); Ajili and Hara-Azumi (2022)
2023	14	17.9	44.9	Compression, edge-assisted deep learning, flexible wearables, smart-home HAR, and real-time sensor-embedded inference	Contoli and Lattanzi (2023); Haghi et al. (2023); Huang et al. (2023); Jameer and Syed (2023); Shakerian et al. (2023)
2024	14	17.9	62.8	Deployment-aware evaluation, device-position independence, IoT-based HAR, sampling-rate effects, and edge-intelligence constraints	Deeptha et al. (2024); Guo et al. (2024); Lalapura et al. (2024); Mekruksavanich and Jitpattanakul (2024); Noor and Park (2024)
2025	20	25.6	88.5	Expansion toward personalized HAR, embedded systems, lightweight architectures, Wi-Fi sensing, edge video processing, and federated or adaptive strategies	Alzubi et al. (2025); Karakish et al. (2025); Lee et al. (2025); Viswanathuni et al. (2025); Shin et al. (2025); Zaoui et al. (2025)
2026	9	11.5	100	Recent emphasis on online incremental learning, TinyML, radar-based sensing, wearable real-time fall detection, and efficient edge deployment	Fan et al. (2026a); Gomez et al. (2026); Kim (2026); Lamaakal et al. (2026); Malche et al. (2026); Mazzieri et al. (2026)
Total	78	100	100	—	—

As shown in Table 5, the reviewed literature is concentrated in the most recent years of the review period. Studies published between 2023 and 2026 account for 57 of the 78 included studies, representing 73.1% of the final corpus. The highest concentration appears in 2025, with 20 studies, equivalent to 25.6% of the corpus. This temporal pattern indicates that edge-based HAR has gained momentum alongside the growth of wearable intelligence, embedded AI, TinyML, hardware-aware optimization, and deployment-aware evaluation.

The temporal profile also shows a progressive change in emphasis. Earlier studies were more strongly associated with feasibility, wearable sensing, smartphones, microcontrollers, and low-power implementation. More recent studies increasingly address adaptive learning, personalized HAR, Wi-Fi and radar-based sensing, TinyML, online incremental learning, federated learning, and real-time deployment under constrained hardware conditions. This supports the relevance of reviewing the 2021–2026 period as a contemporary window for understanding how HAR is shifting from accuracy-centered evaluation toward deployment-aware edge intelligence.

Application Domains, Sensors and Datasets

The research examined in this review demonstrates that the scope of edge based HAR has moved beyond merely classifying physical activities. While recognizing daily living activities continues to be a primary focus for edge-based HAR, it is now being used in applications such as healthcare; rehabilitation; fall detection; elderly monitoring; sports analytics; smart home automation; automotive driver monitoring; industrial processes; and human-robot interactions. Edge-based HAR is relevant to many of these application areas due to their requirements for real time responses, continued data collection (monitoring), localized processing (inference) and both preserving user's privacy while conserving power.

Accelerometer; gyroscope; Inertial Measurement Unit (IMU); smartphone; smartwatch; and wearable devices have been the most common types of sensing technologies used with edge-based HAR. Accelerometers, Gyroscopes, IMUs, Smartphones, Smartwatches and Wearable Devices all support body worn and mobile sensing technologies that support edge-based HAR. However, the study also included other sensing modalities including wireless fidelity/channel state information (Wi-Fi/CSI); radar; cameras/videos; pressure sensors; audio; flexible sensors; textile sensors and strain-based sensing systems.

Table 6 summarizes the application domains, sensing modalities, and dataset types reported in the included studies. This table is useful for identifying the empirical contexts in which edge-based HAR has been most frequently applied, as well as the dominant data sources and dataset types used for model development and evaluation. Because several studies addressed more than one application domain or used more than one sensor type, the categories are not mutually exclusive.

As shown in Table 6, wearable and inertial sensing form the main empirical basis of edge-based HAR. Accelerometers, gyroscopes, IMUs, smartphones, smartwatches, and wearable sensors are frequently used because they are portable, widely available, relatively low-cost, and suitable for continuous body-motion monitoring.

Table 6: Application Domains, Sensors and Datasets in Edge-Based HAR

Dimension	Main Categories	Number of Studies
Application domain	Healthcare, clinical monitoring, rehabilitation, elderly care, or fall detection	66
	Daily living/basic activity recognition	58
	Smart home or ambient intelligence	45
	Sports, exercise, fitness, or training	32
	Human–robot, industrial, or workplace monitoring	10
	Driver or transport-related activity recognition	3
	Agriculture-related activity recognition	1
Sensors/devices/data sources	Accelerometer	59
	Gyroscope	49
	Wearable device or wearable sensor	59
	Smartphone or mobile phone	31
	IMU or inertial sensors	22
	Magnetometer	17
	Smartwatch	13
	Wi-Fi/CSI	7
	Camera/video/vision	6
	Pressure sensor or barometer	5
	Audio/microphone	4
	Radar	3
	Flexible, textile, or strain sensors	16
Dataset type	Public or benchmark dataset	67
	Private or self-collected dataset	39
	Simulated dataset	9

Note: Categories are not mutually exclusive; one study may be counted in more than one application domain, sensing modality, or dataset type

This pattern shows that most edge-based HAR research continues to rely on body-centered sensing, especially in applications related to daily activity recognition, fall detection, rehabilitation, clinical monitoring, and wearable healthcare (Athota and Sumathi, 2022; Elsts and McConville, 2021; Malche et al., 2026; Rivas-Caicedo et al., 2025; Yamane et al., 2025).

At the same time, the reviewed studies suggest that HAR at the edge is gradually moving beyond conventional inertial sensing.

The use of Wi-Fi/CSI, radar, vision, pressure, audio, flexible sensors, textile sensors, and strain-based systems reflects a broader interest in context-aware, non-contact, and body-integrated recognition. Wi-Fi/CSI and radar are especially useful in smart-home and ambient-intelligence scenarios because they can support continuous recognition without requiring users to wear devices. Flexible and textile sensors, on the other hand, make it possible to integrate sensing more naturally into wearable systems while preserving comfort and mobility (Guo et al., 2024; Haghi et al., 2023; Mani et al., 2023; Mazzieri et al., 2026; Viswanathuni et al., 2025). This expansion increases the practical scope of HAR, but it also makes comparison across studies more difficult. Each sensing modality involves different requirements in terms of sampling rate, preprocessing, environmental robustness, privacy, data volume, computational cost, and hardware feasibility. For this reason, findings obtained with one sensor configuration cannot be directly transferred to another deployment context without considering these technical differences. Future HAR studies should therefore report sensor configuration, sensor placement, sampling conditions, preprocessing steps, and target hardware more consistently.

The dataset distribution also reveals a relevant tension in the field. Public and benchmark datasets remain important because they support reproducibility and allow comparison across models. However, the frequent use of private or self-collected data sets shows that many edge-based HAR applications require context-specific data that public benchmarks do not fully capture. This is particularly important in clinical monitoring, rehabilitation, elderly care, wearable sensing, and domain-specific recognition tasks, where user behavior, sensor placement, and environmental conditions may differ substantially from controlled benchmark settings (Fan et al., 2026a; Malche et al., 2026; Mekruksavanich and Jitpattanakul, 2025; Scrugli et al., 2023). Consequently, future research should combine benchmark-based evaluation with real-world and context-specific validation to improve both comparability and deployment relevance.

Machine Learning Models and Edge Deployment Platforms

The included studies reported a broad range of machine learning and deep learning models, reflecting the methodological diversity of edge-based HAR. CNN and LSTM architectures were the most frequent model families, which confirms the importance of both spatial and temporal representation learning in activity recognition. CNNs were commonly used to extract local patterns from segmented sensor windows, whereas LSTM and GRU models were used to capture temporal dependencies in motion sequences. Hybrid CNN-LSTM architectures were also frequently reported because they combine spatial feature extraction with sequential modeling, making them suitable for recognizing dynamic and transitional human activities (Athota and Sumathi, 2022; Contoli and Lattanzi, 2023; Fan et al., 2026a; Rivas-Caicedo et al., 2025; Zhou et al., 2025).

However, the results also show that edge-based HAR does not depend exclusively on deep learning. Classical machine learning models such as SVM, Random Forest, kNN, Decision Tree, MLP, and ANN continue to be relevant, especially when computational cost, interpretability, memory efficiency, and deployment feasibility are important. This is a significant finding because the most complex architecture is not always the most practical option for edge environments. In several deployment contexts, simpler models can provide competitive recognition performance while requiring lower inference cost, less memory, and better compatibility with microcontrollers, wearable devices, or mobile platforms (Elsts and McConville, 2021; Mekruksavanich and Jitpattanakul, 2024; Yamane et al., 2025; Zaoui et al., 2023).

This tendency is also reflected in studies that use compact recurrent models, cascade strategies, lightweight architectures, and knowledge-distillation-based approaches. These methods suggest that reducing model complexity should not be understood as a weakness, but as a deliberate strategy to make HAR models more deployable. In wearable, mobile, and embedded contexts, a smaller or compressed model may be more useful than a larger architecture when it maintains acceptable accuracy while reducing memory use, inference time, energy consumption, or hardware requirements (Gad et al., 2025; Zhao, 2025).

The reviewed studies also show that edge-based HAR is evaluated across diverse deployment platforms, including wearable devices, smartphones, smartwatches, microcontrollers, embedded systems, Raspberry Pi boards, NVIDIA Jetson platforms, IoT environments, and cloud/fog/edge hybrid architectures. This diversity reinforces the idea that edge-based HAR should be approached as a hardware-aware system design problem rather than as a purely algorithmic classification task. The practical value of a model depends not only on its predictive performance, but also on its ability to operate

DOI: 10.3844/jcssp.2026.2590.2616

efficiently on the target device under real constraints of processing capacity, memory, battery life, latency, and connectivity.

Table 7 presents the model families and deployment platforms identified in the reviewed studies. The table helps connect algorithmic choices with hardware-oriented deployment contexts and shows that model selection in edge-based HAR is closely related to the conditions under which the system is expected to operate. Because several studies compared multiple models or evaluated more than one deployment environment, the categories are not mutually exclusive.

Table 7: Machine Learning Models and Edge Deployment Platforms

Dimension	Main Categories	Number of Studies
Model family	CNN	52
	LSTM	37
	SVM	20
	Random Forest	15
	MLP/ANN	13
	GRU	13
	Transformer or attention-based models	9
	kNN	9
	Decision Tree	7
	Spiking or neuromorphic models	7
	TinyML-based models	6
	Autoencoder-based models	2
Edge/deployment platform	Wearable device or wearable environment	61
	Edge computing platform or edge inference setting	42
	Microcontroller/MCU-based platform	32
	Smartphone or mobile device	28
	Embedded system or embedded board	28
	IoT/AIoT environment	18
	Raspberry Pi	17
	Smartwatch	13
	NVIDIA Jetson platform	10
	Cloud/fog/hybrid environment	12

Note: Categories are not mutually exclusive; one study may report more than one model family or deployment platform. CNN = Convolutional Neural Network; LSTM = Long Short-Term Memory; GRU = Gated Recurrent Unit; SVM = Support Vector Machine; kNN = k-Nearest Neighbor; ANN = Artificial Neural Network; MLP = Multilayer Perceptron; MCU = Microcontroller Unit; AIoT = Artificial Intelligence of Things

Table 9 summarizes the completeness of deployment-aware reporting across the 78 included studies. The table distinguishes between studies that reported recognition performance only and those that also provided hardware, latency, energy, power, memory, or model-size information.

As shown in Table 9, all included studies reported recognition performance, but deployment-aware reporting was less consistent. Most studies identified the target hardware or platform, and many reported latency, energy, power, memory, or model-size indicators. However, only 28 studies provided a complete deployment-aware profile combining recognition performance, hardware/platform specification, latency or inference time, energy or power indicators, and memory or model-size information. This confirms that edge-based HAR has moved beyond accuracy-only evaluation, while also showing that reporting practices remain fragmented and limit direct comparison across models, platforms, datasets, and deployment contexts.

As shown in Table 7, CNN and LSTM dominate the modeling landscape, but it also reveals the emergence of newer deployment-aware approaches such as TinyML, spiking neural networks, neuromorphic models, and Transformer/attention-based architectures. These models are relevant because the field is not only seeking higher accuracy, but also more efficient inference under hardware and energy constraints (Agarwal, 2025; Fra et al., 2022; Lamaakal et al., 2026; Li et al., 2023; Zhou et al., 2025).

The deployment results also show that wearable platforms are the most common execution context. This is expected because HAR usually depends on sensors placed near or on the human body. However, the presence of microcontrollers, embedded boards, Raspberry Pi, and Jetson platforms indicates that recent studies increasingly evaluate whether HAR models can operate under actual edge conditions. This is crucial because a model that performs well in offline evaluation may not be suitable for deployment if it exceeds memory, battery, or latency constraints (Elsts and McConville, 2021; Malche et al., 2026; Ajili and Hara-Azumi, 2022; Zhou et al., 2025).

Performance Metrics and Latency Energy Accuracy Trade-Offs

The most important result of the review is that edge-based HAR research is gradually moving from an accuracy-centered evaluation paradigm toward a deployment-aware evaluation paradigm. Accuracy was reported in all included studies, but a substantial number of studies also considered latency, inference time, energy consumption, power usage, battery-related indicators, memory footprint, model size, computational cost, execution cost, or throughput.

This shift is central to the contribution of the review. In cloud-based or offline HAR, high accuracy is often treated as the main indicator of model success. In edge-based HAR, however, recognition performance must be interpreted together with deployment feasibility. A model may achieve high accuracy under benchmark conditions but still be unsuitable for real-time wearable, mobile, or embedded deployment if it consumes excessive energy, requires too much memory, depends on unavailable hardware resources, or produces delayed predictions. Therefore, edge-based HAR evaluation requires a broader set of indicators that connect predictive performance with practical execution constraints. Table 8 summarizes the performance metrics and optimization strategies reported across the included studies.

This table is important because it shows the extent to which recent HAR research has moved beyond accuracy-centered evaluation and incorporated deployment-aware indicators such as latency, inference time, energy consumption, power usage, memory footprint, model size, computational cost, and throughput. It also identifies the main optimization strategies used to make HAR models more suitable for edge deployment, including lightweight architectures, hardware-aware implementation, quantization, model compression, feature reduction, TinyML, pruning, distributed inference, and knowledge distillation. Because several studies reported multiple metrics or applied more than one optimization strategy, the categories are not mutually exclusive.

Optimization Results

These results show that the two most commonly used methods for optimizing edge-based HAR are lightweight architectural design and implementing HAR using hardware-specific implementations. Methods such as quantization, compression, reducing features, pruning, TinyML, and distributing inference also frequently occur in the literature. Additionally, these various methods indicate that HAR deployed in an edge environment will be optimized across multiple levels including: Sensor data acquisition, how the acquired data represents features, what architecture the model has, how precise (numerically) the model is implemented, where it will be deployed, and the inferencing pipeline (Alessandrini et al., 2021; Alotaibi et al., 2023; Bollampally et al., 2024; Knez et al., 2021; Sakka et al., 2025).

In addition to these findings other research indicated that adaptive compression, implementing HAR with TinyML in mind, and deploying HAR models in a way that takes advantage of hardware specific characteristics can make them more deployable in real world applications. When this occurs, the contributions associated with each study are generally described by both improved accuracy as well as reductions in: Total energy consumed, required memory size, average time to inference, and/or adaptability to resource-constrained mobile and embedded systems. Therefore, optimizing strategies should be viewed as part of an entire end-to-end deployment process instead of as separate post-processing steps.

Trade-Off Patterns

From examining the trade-offs presented above it is evident that there is no "best" model for all edge-based HAR. What constitutes the "best" solution is dependent upon interactions between the particular application domain (s), sensor modalities being utilized, hardware platforms available, and desired performance priorities.

Table 8: Performance Metrics and Optimization Strategies in Edge-Based HAR

Dimension	Main Categories or Patterns	Number of Studies / Evidence
Performance metric	Accuracy	78
	Computational cost, execution cost, throughput, or related measure	69
	Energy, power, or battery-related metric	64
	Latency or inference time	58
	Memory usage, memory footprint, or model size	54
	F1-score	47
	Precision	42
	Recall	38
Optimization strategy	Lightweight or compact architecture design	45
	Hardware acceleration or hardware-aware implementation	44
	Quantization	28
	Model compression	28
	Feature selection, feature reduction, or sampling reduction	19
	TinyML or TensorFlow Lite-based optimization	14
	Pruning or sparsity	12
	Distributed, partitioned, cascaded, or edge-cloud inference	12
	Knowledge distillation	6

Table 9: Reporting Completeness of Deployment-Aware Evidence in Edge-Based HAR Studies

Reporting category	Operational definition	Number of studies	Percentage	Representative evidence
Accuracy reported	The study reported accuracy or equivalent recognition-performance results.	78	100.00%	Athota and Sumathi (2022); Elsts and McConville (2021); Zhou et al. (2025)
Hardware/platform specified	The study reported the target hardware, device, board, wearable, smartphone, microcontroller, Raspberry Pi, Jetson, FPGA, embedded board, or equivalent platform.	77	98.70%	Elsts and McConville (2021); Contoli and Lattanzi (2023); Scrugli et al. (2023); Rivas-Caicedo et al. (2025)
Latency/inference time reported	The study reported latency, inference time, response time, execution time, runtime, or equivalent temporal-performance indicator.	58	74.40%	Fan et al. (2026a); Elsts and McConville (2021); Malche et al. (2026); Rivas-Caicedo et al. (2025)
Energy/power/battery reported	The study reported energy consumption, power consumption, battery use, battery life, low-power operation, or equivalent energy-related indicator.	64	82.10%	Contoli and Lattanzi (2023); Shen et al. (2022); Scrugli et al. (2023); Rivas-Caicedo et al. (2025)
Memory/model size reported	The study reported memory usage, memory footprint, RAM/Flash use, storage, model size, parameter count, or equivalent size-related indicator.	54	69.20%	Elsts and McConville (2021); Novac et al. (2021); Lattanzi et al. (2022); Malche et al. (2026)
Accuracy-dominant reporting	The study mainly emphasized recognition performance and provided limited or no deployment-aware measurements such as latency, energy, power, memory, or model size.	2	2.60%	Athota and Sumathi (2022); Saeed et al. (2021)
Full deployment-aware reporting	The study reported accuracy together with hardware/platform information, latency/inference time, energy/power/battery indicators, and memory/model size information.	28	35.90%	Elsts and McConville (2021); Contoli and Lattanzi (2023); Scrugli et al. (2023); Rivas-Caicedo et al. (2025); Malche et al. (2026)

For instance, in order to achieve timely fall detection and clinical monitoring, low-latency recognition is typically a requirement; however, when utilizing long term wearables, there tends to be greater focus on battery life and memory efficiency. Furthermore, since HAR models running on microcontrollers tend to be limited by strict model-size requirements compared to those deployed on smartphones, while wireless connectivity (e.g. Wi-Fi) or radar-based sensing may reduce the need for body worn sensors they do add signal processing complexities (Guo et al., 2024; Mazzieri et al., 2026; Viswanathuni et al., 2025; Shakerian et al., 2023).

Summary of Results

Overall, the results show that edge-based HAR is technically mature but methodologically fragmented. The reviewed literature is dominated by wearable and inertial sensing, CNN/LSTM-based architectures, and increasing use of TinyML, model compression, quantization, lightweight design, and microcontroller-based deployment. The most relevant finding is that evaluation has expanded beyond accuracy, although reporting remains inconsistent across latency, energy, memory, model size, and hardware-level evidence. This fragmentation limits direct comparison and reinforces the need for standardized deployment-aware reporting in future HAR studies.

Discussion

Main Trends in Edge-Based HAR

The research demonstrates that Human Activity Recognition (HAR) at the Edge is evolving beyond an activity classification issue to being a deployment aware computing challenge. Research now examines not just what activities are recognized by HAR models based on sensor readings, but whether HAR models can function with the constraints of wearable devices, Smartphones, Embedded Boards, Micro Controllers, and Edge Platforms. The above change in emphasis is evident when looking at research that evaluate HAR models as opposed to simply testing them in an off-line or Cloud Based Experimental Setting (Elsts and McConville, 2021; Novac et al., 2021; Zhou et al., 2025).

The transition described above is important due to the fact that there are several HAR Applications that need continuous and instant recognition. Many HAR applications include healthcare monitoring, fall detection, rehabilitation, elderly care, sports tracking, smart home applications, driver monitoring, and human-robot interaction. These applications will be less useful if the systems have a delay or if it takes too long to recognize human activity. Systems that take too much time to recognize human activity will also consume a lot of power which would make them impractical to use in the field (Deeptha et al., 2024; Malche et al., 2026; Mekruksavanich and Jitpattanakul, 2025).

The dominance of wearable and inertial sensing also confirms that body-centered data acquisition remains the foundation of edge-based HAR. Accelerometers, gyroscopes, IMUs, smartphones, and wearable sensors are widely used because they are portable, relatively low-cost, and suitable for continuous motion monitoring (Athota and Sumathi, 2022; Mekruksavanich and Jitpattanakul, 2024; Rivas-Caicedo et al., 2025). At the same time, the field is diversifying toward non-inertial and context-aware modalities. Wi-Fi/CSI and radar-based approaches support non-contact recognition, while flexible and textile sensors enable more body-integrated HAR systems (Ankita et al., 2021; Haghi et al., 2023; Mazzieri et al., 2026; Viswanathuni et al., 2025).

This diversification expands the scope of HAR but also increases methodological complexity. Each sensing modality introduces different requirements in terms of sampling frequency, signal preprocessing, privacy, environmental robustness, data volume, and computational load. Therefore, edge-based HAR should be understood as a system-level problem in which sensors, models, hardware, and evaluation metrics interact.

From Accuracy-Centered Evaluation to Deployment-Aware Evaluation

One of the most important findings of this review is a shift from an emphasis on accuracy in evaluations to an emphasis on deployment awareness in evaluations. While accuracy will always be essential since models need to accurately recognize human activity, accuracy by itself does not suffice for the purposes of HAR in an Edge environment. For example, a model could have achieved excellent performance during bench testing for recognizing human activities, yet not be suitable for actual deployment due to its large memory requirements, long latency, very high energy consumption, or dependence on specific hardware resources that do not exist in wearables or embedded systems (Elsts and McConville, 2021; Lalapura et al., 2024; Suwannarat and Kurdthongmee, 2021).

Some studies demonstrate the extended reasoning for evaluating HAR as described above. Using distributed wearable recognition reduces both local processing/communication overheads. Using frequency sampling rate reduction lowers the required energy and data volume if recognition quality is maintained (Rivas-Caicedo et al., 2025; Yamane et al., 2025). In addition, studies on using microcontrollers for HAR also demonstrate that deploying a model into the embedded domain is dependent upon the execution time of the model, the amount of memory used, and whether the model has been made feasible for embedded use, rather than just based on how well the model classified (Elsts and McConville, 2021).

Despite this progress, reporting practices remain inconsistent. Some studies provide detailed hardware-level measurements, including latency, inference time, memory footprint, power consumption, or model size. Others describe efficiency more generally without reporting comparable deployment metrics. This limits direct comparison across models, datasets, devices, and application domains. For this reason, future HAR studies should report accuracy or F1-score together with latency, inference time, energy or power consumption, memory footprint, model size, and the exact hardware platform used for inference.

Model Complexity and Hardware Feasibility

CNN's and LSTMs (and related) models have dominated much of the reviewed literature due to their ability to learn both spatial and temporal relationships within a sensor stream. CNN's are often utilized for extracting spatial features from segmented sensor window areas; whereas LSTM and Gated Recurrent Units (GRU) are utilized for modeling sequential dependences present within human movement data. As hybrid CNN-LSTMs (a combination of spatial feature extraction with temporal modeling), they provide an advantage for modeling transitional and dynamic activity patterns (Alsaadi et al., 2025).

Although model complexity increases as it relates to improved accuracy in controlled experimentation settings; increased model complexity typically equates to increased computational costs, memory usage, latency and power consumption. Therefore, when deploying these types of models on a smart-watch, wearable node, micro-controller or other low-power embedded devices; it will likely create problems (Elsts and McConville, 2021; Lattanzi et al., 2022; Novac et al., 2021).

As previously stated, classical machine learning algorithms still hold relevance in this area. Models including Support Vector Machine (SVM), Random Forest, k-Nearest Neighbor (kNN), Decision Trees, Multilayer Perceptron (MLP) and Artificial Neural Networks (ANN) still exist as viable options since they potentially provide adequate accuracy at reduced computational costs. In addition, where constraints limit the feasibility of utilizing high-performance architectures with respect to memory availability, battery life, latency etc., using simpler models with slightly less accurate results could be more feasible than trying to deploy a complex architecture that has limited capability in meeting its required specifications (Chang et al., 2022; Elsts and McConville, 2021; Saeed et al., 2021).

This indicates that model selection in edge-based Human Activity Recognition (HAR) systems should be based upon deployment conditions. Therefore, the "best" model is not always the one that achieves the highest accuracy offline, but the one that represents the most optimal trade-off between recognition performance, inference speed, memory utilization, energy consumption and hardware compatibility. This further emphasizes the necessity of designing hardware aware models versus selecting models solely based upon accuracy alone.

Latency Energy Accuracy Trade-offs

The Latency Energy Accuracy (LEA) trade-off is one of the central technical challenges in edge-based HAR. Improving one dimension may negatively affect another. For example, reducing the sampling rate can lower energy consumption and computational cost, but it may also remove motion patterns that are important for accurate recognition. Similarly, compression or quantization can reduce model size and inference time, but overly aggressive optimization may degrade recognition performance (Contoli and Lattanzi, 2023; Zhou et al., 2025).

Additionally, the LEA trade-off will depend heavily upon the target device on which the model is deployed. A model that is deployable on a smartphone or raspberry pi might not be able to run effectively on a microcontroller. On the other hand, models that have been optimized for use with Jetson or FPGA-based systems may have very efficient inference capabilities; however, there may be additional costs associated with implementing such solutions and potentially limited portability to other types of devices (Noor and Park, 2024; Ajili and Hara-Azumi, 2022). Thus, when making statements regarding "edge deployment", we must clearly indicate the intended deployment hardware, as different types of edge hardware vary greatly in terms of available computing resources, memory space and available energy (Ajili and Hara-Azumi, 2022; Zhou et al., 2025).

TinyML provides a valuable solution to address many of these challenges. This is due to its ability to provide ML inferencing on extremely resource-constrained devices. While TinyML can provide some degree of relief from the LEA trade-off by providing tighter controls on memory usage, model size, inference cost and precision of numeric values used during training and inference (Hayajneh et al., 2024; Zhou et al., 2025); it does so in an even more deliberate manner. That is to say, decisions made by developers using TinyML to develop a HAR system do so under a much more explicitly defined constraint set than traditional development environments.

Another potential direction that holds promise for developing HAR systems that consume less energy include neuromorphic and spiking neural network techniques. Such networks seek to reduce power consumption through event-driven or biological computations while still retaining reasonable amounts of useful recognition capability (Fra et al., 2022; Li et al., 2023). While not as prevalent as CNNs or LSTMs, they are worth mentioning here as continuous sensor streams combined with always-on recognition scenarios require architectures that can function efficiently for extended durations.

Reproducibility, Benchmarking, and Real-World Validation

The study identifies several other issues with reproducibility and benchmarking. Datasets used in public research settings allow researchers to compare their results on the same types of data and provide a basis for future work (methodological continuity). Unfortunately, however, benchmarks rarely represent typical real-world usage since different studies have employed sensors at varying locations, sampled at varying rates, were conducted using different devices, had varying degrees of user variability, and took place under differing environmental conditions (Elsts and McConville, 2021; Zhou et al., 2025).

Datasets collected by private entities and/or through self-collection are valuable in that they typically contain information about how the technology will be utilized in a particular way (e.g., clinical activities, rehabilitation, smart home monitoring, flexible wearable sensing, or personalized recognition). That said, these datasets pose significant barriers to achieving reproducibility in that they are often not made available to the public, and/or the details regarding the methodology employed in developing them are not provided (Fan et al., 2026b; Haghi et al., 2023; Mekruksavanich and Jitpattanakul, 2025). There exists an inherent tradeoff here; if the data being developed reflects "real world" use cases then the results will likely be very applicable in practice unfortunately the lack of transparency into the development process will make it difficult to independently verify those results.

HAR systems remain generally difficult to generalize. HAR algorithms can be sensitive to aspects of the user (e.g., sensor position, device orientation, activity execution style), and changes to the environment. Those studies that attempt to develop methods for addressing some of these generalizability concerns (device-position independence, transfer learning, personalization, or adaptive sensing) demonstrate that there is still much to be done before edge-based HAR is validly generalized across users, devices, and contexts (Abuhoureyah et al., 2024; Fu et al., 2021; Mekruksavanich and Jitpattanakul, 2024).

This is important because in contrast to most lab-based datasets, edge-based HAR systems operate in a variety of non-controlled environmental conditions.

Validation designs can dramatically impact reported performance of HAR models. Studies comparing various validation methodologies have demonstrated that standard data splitting techniques can potentially lead to overestimation of recognition accuracy when both training and test sets consist of largely the same subjects or patterns. As such, deployment-aware HAR needs to pay greater attention to designing subject-independent, cross-user and cross-device validation protocols especially in applications where end-users interact with the system (s) in uncontrolled environments.

Therefore, future studies need to expand upon current accuracy measures based solely on within-dataset accuracy and include cross-user, cross-device, cross-position and cross-environment evaluations wherever feasible. Otherwise, without sufficient validation it is unclear whether a HAR model has achieved sufficient robustness to support deployment beyond experimental/controlled settings.

Implications for Researchers and Practitioners

Findings for researchers point toward the need for additional detail within research on edge-based HAR; specifically, that studies include an expanded format of their reporting. A minimum requirement for such reporting would include:

- (i) The data set utilized
- (ii) The type of sensors used as input into the model
- (iii) The sampling rate of those sensors
- (iv) The architecture of the models being trained
- (v) The target device (s) that will be deployed
- (vi) Metrics of performance (accuracy or F1 score)
- (vii) Latency or time required to make predictions
- (viii) Memory requirements or model size
- (ix) Energy or power usage by the model

As these details are reported in all other fields of study, it is reasonable to expect they be included in reports of edge-based HAR research so that readers can assess whether proposed models are actually deployable at the edge (Elsts and McConville, 2021; Novac et al., 2021; Zhou et al., 2025).

Similarly, findings for practitioners imply that selection of a suitable HAR model depends upon specific contexts. In applications where safety is critical or clinical, for example, low-latency prediction capabilities and high reliability may be more valuable than small improvements in predictive accuracy. Likewise, if an application involves long-term monitoring using wearables, energy usage per unit time and/or memory usage per unit time may have a significant impact on usability. Finally, when designing edge-based HAR systems intended for use in smart homes or ambient intelligence environments, concerns about privacy, no intrusiveness and environmental robustness may take precedence over many other considerations (Deeptha et al., 2024; Malche et al., 2026; Viswanathuni et al., 2025).

Research Gaps and Future Directions

The first major gap concerns the lack of consistent and systematic reporting of latency–energy–accuracy trade-offs. Although all included studies reported recognition performance, only one subset provided comparable evidence on latency, inference time, energy consumption, power usage, memory footprint, model size, and hardware platform. This limits direct comparison across models, datasets, devices, and application domains. Future studies should report accuracy or F1-score together with latency, energy or power consumption, memory or model-size indicators, and the exact hardware used for inference, so that deployment feasibility can be evaluated more consistently (Lalapura et al., 2024; Suwannarat and Kurdthongmee, 2021; Thottempudi et al., 2024).

The second gap concerns real-world validation. Many current studies rely on public datasets, benchmark datasets, or experimentally controlled laboratory settings. These settings are useful for model development and initial comparison, but they do not fully represent the variability of real deployment environments. Edge-based HAR systems must operate under changing user behavior, heterogeneous devices, different sensor placements, variable sampling conditions, and environmental noise. Therefore, future studies should move toward longer monitoring periods, more diverse user groups, multiple device configurations, varied sensor positions, and less controlled deployment scenarios. This would make it possible to evaluate whether HAR models remain reliable when they are used outside benchmark conditions (Fan et al., 2026a; Fu et al., 2021; Mekruksavanich and Jitpattanakul, 2025).

The third gap concerns generalization, personalization, and adaptive learning. HAR systems must support users with different movement patterns, body characteristics, activity styles, sensor placements, and device configurations. This is particularly important for edge-based HAR because models are expected to operate continuously on wearable, mobile, or embedded devices under changing conditions. Potential avenues for improving generalization and personalization include transfer learning, incremental learning, adaptive sensing, federated learning, and personalized model orchestration. However, these approaches should be validated under constrained edge environments rather than only under offline or benchmark conditions (Abuhoureyah et al., 2024; Fu et al., 2021; Zaoui et al., 2025).

Within this gap, an important emerging direction is the transition from static model deployment toward on-device continuous learning and privacy-preserving federated learning. In conventional HAR pipelines, models are usually trained offline and then deployed to wearable, mobile, or embedded devices. However, real-world activity recognition is affected by changes in users, sensor placement, device orientation, movement patterns, and environmental conditions. For this reason, edge HAR systems may benefit from local adaptation mechanisms that allow models to update incrementally from new user-specific data without requiring complete retraining in the cloud (Fan et al., 2026a; Lee et al., 2025; Zaoui et al., 2025; Zhao, 2025).

On-device continuous learning is especially relevant for personalized HAR because it allows models to adapt to new movement patterns while preserving low-latency inference. However, this approach also introduces important technical challenges. Continuous adaptation must be achieved under limited memory, restricted processing capacity, and battery constraints, while also avoiding catastrophic forgetting and distinguishing useful adaptation signals from noisy sensor data. In this context, federated learning offers a complementary privacy-preserving strategy because it enables multiple devices or clients to improve HAR models collaboratively without sharing raw sensor data. This is particularly relevant in health monitoring, rehabilitation, elderly care, and smart-home applications, where activity data may reveal sensitive behavioral or physiological information.

The fourth gap concerns the limited integration between model optimization and hardware-aware design. Techniques such as quantization, pruning, compression, TinyML, FPGA acceleration, knowledge distillation, and neuromorphic computing are promising for improving edge feasibility. Nevertheless, many studies still evaluate these techniques separately, without comparing them under common datasets, platforms, and deployment metrics. Future research should therefore assess optimization strategies using shared benchmarks, comparable hardware platforms, and standardized reporting of accuracy, latency, energy consumption, memory footprint, and model-size indicators (Fra et al., 2022; Gonçalves et al., 2024; Ajili and Hara-Azumi, 2022).

In summary, future progress in edge-based HAR depends on moving from isolated model-performance evaluation toward system-level deployment evaluation. The most suitable HAR systems will not necessarily be those with the highest offline accuracy, but those that maintain reliable recognition under the combined constraints of latency, energy consumption, memory availability, hardware resources, user variability, privacy requirements, and real-world deployment conditions. For this reason, the next stage of edge-based HAR should move toward adaptive, privacy-aware, and resource-conscious learning pipelines capable of maintaining performance under changing operational contexts.

Conclusion

Edge-based HAR is a key research area for the development of intelligent, low-latency, energy-efficient, and resource-aware activity recognition systems. The evidence shows that the field has moved beyond accuracy-centered evaluation toward deployment-aware design. The main challenge is no longer only to improve recognition performance, but to achieve a practical balance between accuracy, latency, energy consumption, memory demand, and hardware feasibility. Future progress will depend on evaluating HAR systems as integrated edge-intelligence solutions rather than as isolated classification models.

Summary of Findings

This systematic literature review analyzed 78 studies published between 2021 and 2026 on Human Activity Recognition at the edge. The findings show that edge-based HAR has evolved from a conventional activity-classification task into a deployment-aware computing problem. In this context, accuracy remains necessary, but it is no longer sufficient as the main criterion for evaluating HAR systems. Real-world implementation also depends on latency, inference time, energy consumption, power usage, memory footprint, model size, computational cost, and hardware feasibility.

Two main conclusions emerge from the review. First, wearable, mobile, embedded, and IoT-based HAR systems increasingly rely on lightweight architectures, model compression, quantization, TinyML, hardware-aware implementation, and distributed or hybrid inference strategies. These approaches reflect a broader transition toward HAR systems that are not only accurate, but also efficient, portable, and feasible for constrained devices. Second, reporting practices remain fragmented. Although most studies report recognition performance, fewer provide complete deployment-aware evidence combining hardware specifications, latency, energy or power consumption, memory usage, and model-size indicators.

Overall, the review shows that future progress in edge-based HAR will depend on standardized reporting, real-world validation, cross-device generalization, privacy-preserving federated learning, and on-device continuous adaptation. These directions are essential for moving HAR systems from benchmark-level performance toward robust, reproducible, adaptive, and deployable edge intelligence.

Theoretical and Technical Contributions

This study will help develop current knowledge on edge-based HAR research through contributions in three primary areas. First, it organizes recent edge-based HAR research according to sensors, datasets, application domains, machine learning models, deployment platforms, optimization strategies, and deployment-aware metrics. Second, it highlights the latency energy accuracy trade-off as a central analytical problem in edge-based HAR. Third, it identifies the methodological distance between high recognition performance under offline or benchmark conditions and actual deployability under constrained edge, embedded, wearable, mobile, IoT, and TinyML environments.

This study emphasizes the need to consider HAR at the edge as a system level design problem. The model architecture; sensor modality; sampling rate; optimization technique; hardware platform; and application domain are all important factors that affect how feasible it would be to achieve successful real-world deployments.

Practical Implications

The findings provide practical guidance for researchers and practitioners designing edge-based HAR systems. For safety-critical and clinical applications, such as fall detection, rehabilitation, and patient monitoring, low-latency recognition and reliable performance may be more important than marginal improvements in accuracy. For long-term wearable monitoring, energy efficiency, memory footprint, and model size become essential for continuous operation. For smart home and ambient intelligence applications, non-intrusive sensing, privacy, environmental robustness, and hardware feasibility should be prioritized.

Therefore, developers of edge-based HAR systems should report both recognition metrics, such as accuracy, precision, recall, and F1-score, and deployment metrics, such as latency, inference time, energy consumption, power usage, memory footprint, model size, computational cost, and the target hardware platform. Without these metrics, it is difficult to determine whether a proposed HAR model is suitable for real-world edge deployment.

Limitations

This review has several limitations. First, the search was restricted to studies published between 2021 and 2026 and written in English. Therefore, relevant studies published before 2021 or in other languages may not have been captured. Second, the review focused on original research articles and full research papers with sufficient methodological and technical detail. As a result, reviews, surveys, bibliometric studies, book chapters, theses, abstracts, posters, short papers, and other non-empirical documents were excluded.

Third, although 78 studies were included in the final corpus, the heterogeneity of datasets, sensors, activity labels, sampling rates, model architectures, deployment platforms, validation protocols, and reported metrics prevented the application of a quantitative meta-analysis. Consequently, the review followed a thematic synthesis approach. Despite these limitations, the eligibility criteria helped ensure that the included studies provided sufficient empirical and technical evidence to support a deployment-aware analysis of HAR at the edge.

Acknowledgment

The authors acknowledge the scientific contributions of the researchers whose published studies formed the basis of this systematic literature review.

Funding Information

This research received no external funding.

Author's Contributions

Jose Antonio Rojas Guillén: Led the conceptualization of the study, defined the review scope, structured the systematic literature review protocol, coordinated the search strategy, organized the extraction matrix, conducted literature analysis, synthesized the main findings, drafted substantial sections of the manuscript, integrated methodological and editorial revisions, and approved the final version of the manuscript.

Wini Ebelin Quispe Bautista: Contributed to the research design, refinement of the review objective and research question, support in database search organization, PRISMA-based screening, eligibility verification, data extraction, cross-checking of selected studies to reduce selection bias, synthesis of findings, writing-original draft, writing-review and editing, and approved the final version of the manuscript.

Arturo Gamarra Moreno: Supervised the technical coherence of the manuscript from an engineering perspective, reviewed the interpretation of edge, embedded, wearable, and hardware-constrained deployment environments, validated the discussion of latency energy accuracy trade-offs, contributed to the analysis of computational, mechanical, and resource-related deployment constraints, reviewed the technical terminology, writing-review and editing, and approved the final version of the manuscript.

Ethics

This study used secondary data obtained from published scientific literature and did not involve human participants, animals, or sensitive personal data. All sources were analyzed and cited according to academic and ethical standards. This manuscript is original and has not been previously published or submitted elsewhere.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Abuhoureyah, F., Sim, K. S., & Wong, Y. C. (2024). Multi-User Human Activity Recognition Through Adaptive Location-Independent WiFi Signal Characteristics. *IEEE Access*, *12*, 112008–112024. https://doi.org/10.1109/access.2024.3438871

Agarwal, P. (2025). niCNN: A Novel Neuromorphic Approach to Energy-Efficient and Lightweight Human Activity Recognition on Edge Devices. *International Journal of Advanced Computer Science and Applications*, *16*(7), 0713. https://doi.org/10.14569/ijacsa.2025.0160713

Ajili, M. T., & Hara-Azumi, Y. (2022). Multimodal Neural Network Acceleration on a Hybrid CPU-FPGA Architecture: A Case Study. *IEEE Access*, *10*, 9603–9617. https://doi.org/10.1109/access.2022.3144977

Alam, A., Das, A., Tasjid, Md. S., & Marouf, A. A. (2021). Leveraging Sensor Fusion and Sensor-Body Position for Activity Recognition for Wearable Mobile Technologies. *International Journal of Interactive Mobile Technologies (IJIM)*, *15*(17), 141. https://doi.org/10.3991/ijim.v15i17.25197

Alessandrini, M., Biagetti, G., Crippa, P., Falaschetti, L., & Turchetti, C. (2021). Recurrent Neural Network for Human Activity Recognition in Embedded Systems Using PPG and Accelerometer Data. *Electronics*, *10*(14), 1715. https://doi.org/10.3390/electronics10141715

Alotaibi, F., Alnfiai, M. M., Al-Wesabi, F. N., Alduhayyem, M., Hilal, A. M., & Hamza, M. A. (2023). Internet of Things-driven Human Activity Recognition of Elderly and Disabled People Using Arithmetic Optimization Algorithm with LSTM Autoencoder. *Journal of Disability Research*, *2*(3), 136–146. https://doi.org/10.57197/jdr-2023-0038

Alsaadi, M., Keshta, I., Ramesh, J. V. N., Nimma, D., Shabaz, M., pathak, N., Singh, P. P., Kiyosov, S., & Soni, M. (2025). Logical reasoning for human activity recognition based on multisource data from wearable device. *Scientific Reports*, *15*(1), 380. https://doi.org/10.1038/s41598-024-84532-8

Alzubi, A., Lin, D., Reimann, J., & Alsaleem, F. (2025). G-CTRNN: A Trainable Low-Power Continuous-Time Neural Network for Human Activity Recognition in Healthcare Applications. *Applied Sciences*,*15*(13), 7508. https://doi.org/10.3390/app15137508

Ankita, Rani, S., Babbar, H., Coleman, S., Singh, A., & Aljahdali, H. M. (2021). An Efficient and Lightweight Deep Learning Model for Human Activity Recognition Using Smartphones. *Sensors*, *21*(11), 3845. https://doi.org/10.3390/s21113845

Armenta-Garcia, J. A., Gonzalez-Navarro, F. F., Caro-Gutierrez, J., & Garcia-Reyes, C. I. (2025). Tools and Methods for Achieving Wi-Fi Sensing in Embedded Devices. *Sensors*, *25*(19), 6220. https://doi.org/10.3390/s25196220

Athota, R. K., & Sumathi, D. (2022). Human activity recognition based on hybrid learning algorithm for wearable sensor data. *Measurement: Sensors*, *24*, 100512. https://doi.org/10.1016/j.measen.2022.100512

Bahador, N., Ferreira, D., Tamminen, S., & Kortelainen, J. (2021). Deep Learning–Based Multimodal Data Fusion: Case Study in Food Intake Episodes Detection Using Wearable Sensors. *JMIR MHealth and UHealth*, *9*(1), e21926. https://doi.org/10.2196/21926

Bibbò, L., Angiulli, G., Laganà, F., Pratticò, D., Cotroneo, F., La Foresta, F., & Versaci, M. (2025). MEMS and IoT in HAR: Effective Monitoring for the Health of Older People. *Applied Sciences*, *15*(8), 4306. https://doi.org/10.3390/app15084306

Bollampally, A., Kavitha, J., Sumanya, P., Rajesh, D., Jaffar, A. Y., Eid, W. N., Albarakati, H. M., Aldosari, F. M., & Alharbi, A. A. (2024). Optimizing Edge Computing for Activity Recognition: A Bidirectional LSTM Approach on the PAMAP2 Dataset. *Engineering, Technology & Applied Science Research*, *14*(6), 18086–18093. https://doi.org/10.48084/etasr.8861

Cavalcante, A. F., Kunst, V. H. de L., Chaves, T. de M., de Souza, J. D. T., Ribeiro, I. M., Quintino, J. P., da Silva, F. Q. B., Santos, A. L. M., Teichrieb, V., & da Gama, A. E. F. (2023). Deep Learning in the Recognition of Activities of Daily Living Using Smartwatch Data. *Sensors*, *23*(17), 7493. https://doi.org/10.3390/s23177493

Chang, J., Kang, M., & Park, D. (2022). Low-Power On-Chip Implementation of Enhanced SVM Algorithm for Sensors Fusion-Based Activity Classification in Lightweighted Edge Devices. *Electronics*, *11*(1), 139. https://doi.org/10.3390/electronics11010139

Contoli, C., & Lattanzi, E. (2023). A Study on the Application of TensorFlow Compression Techniques to Human Activity Recognition. *IEEE Access*, *11*, 48046–48058. https://doi.org/10.1109/access.2023.3276438

Debnath, A., Kim, Y.-W., & Byun, Y.-C. (2025). LightSTATE: A Generalized Framework for Real-Time Human Activity Detection Using Edge-Based Video Processing and Vision Language Models. *IEEE Access*, *13*, 97609–97627. https://doi.org/10.1109/access.2025.3574659

Deeptha, R., Ramkumar, K., Venkateswaran, S., Hassan, M. M., Hassan, Md. R., Noori, F. M., & Uddin, Md. Z. (2024). Enhancing human activity recognition for the elderly and individuals with disabilities through optimized Internet-of-Things and artificial intelligence integration with advanced neural networks. *Frontiers in Neuroinformatics*, *18*, 4583. https://doi.org/10.3389/fninf.2024.1454583

Duarte, O. dos S., Jacinto, G., Véstias, M., & Duarte, R. P. (2025). Intelligent Sports Weights. *Sensors*, *25*(12), 3808. https://doi.org/10.3390/s25123808

Elsts, A., & McConville, R. (2021). Are Microcontrollers Ready for Deep Learning-Based Human Activity Recognition? *Electronics*, *10*(21), 2640. https://doi.org/10.3390/electronics10212640

Fan, L., Peng, H., Xiao, L., Shi, L., Zhou, S., Song, Y., & Fan, H. (2026a). Agile human activity recognition for wearable devices based on online incremental learning. *Frontiers in Public Health*, *14*, 388. https://doi.org/10.3389/fpubh.2026.1727388

Fan, Y., Phaphan, W., Jitpattanakul, A., & Mekruksavanich, S. (2026b). Knowledge Distillation for Efficient Human Activity Recognition on Resource-Constrained Smartwatches. *IEEE Access*, *14*, 25854–25870. https://doi.org/10.1109/access.2026.3664977

Fatima, I., Farhan, A. A., Tamoor, M., Rehman, S. ur, Alhulayyil, H. A., & Tariq, F. (2024). DiscHAR: A Discrete Approach to Enhance Human Activity Recognition in Cyber Physical Systems: Smart Homes. *Computers*, *13*(11), 300. https://doi.org/10.3390/computers13110300

Fra, V., Forno, E., Pignari, R., Stewart, T. C., Macii, E., & Urgese, G. (2022). Human activity recognition: suitability of a neuromorphic approach for on-edge AIoT applications. *Neuromorphic Computing and Engineering*, *2*(1), 014006. https://doi.org/10.1088/2634-4386/ac4c38

Fu, Z., He, X., Wang, E., Huo, J., Huang, J., & Wu, D. (2021). Personalized Human Activity Recognition Based on Integrated Wearable Sensor and Transfer Learning. *Sensors*, *21*(3), 885. https://doi.org/10.3390/s21030885

Gad, M. M., Gad, W., Abdelkader, T., & Naik, K. (2025). Personalized Smart Home Automation Using Machine Learning: Predicting User Activities. *Sensors*, *25*(19), 6082. https://doi.org/10.3390/s25196082

Gomez, J. M. S., Ahmed, S., Souissi, S., & Alouini, M.-S. (2026). Edge-Aided Radar-Based Exercise Form Classification Using Lightweight Ensemble Learning for Personalized Healthcare. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *34*, 1405–1416. https://doi.org/10.1109/tnsre.2026.3669061

Gomez-Carmona, O., Casado-Mansilla, D., Lopez-de-Ipina, D., & Garcia-Zubia, J. (2022). Optimizing Computational Resources for Edge Intelligence Through Model Cascade Strategies. *IEEE Internet of Things Journal*, *9*(10), 7404–7417. https://doi.org/10.1109/jiot.2021.3118845

Gonçalves, P. H. N., Bragança, H., & Souto, E. (2024). Efficient Human Activity Recognition on Wearable Devices Using Knowledge Distillation Techniques. *Electronics*, *13*(18), 3612. https://doi.org/10.3390/electronics13183612

Guo, W., Yamagishi, S., & Jing, L. (2024). Human Activity Recognition via Wi-Fi and Inertial Sensors With Machine Learning. *IEEE Access*, *12*, 18821–18836.https://doi.org/10.1109/access.2024.3360490

Haghi, M., Ershadi, A., & Deserno, T. M. (2023). Recognizing Human Activity of Daily Living Using a Flexible Wearable for 3D Spine Pose Tracking. *Sensors*, *23*(4), 2066. https://doi.org/10.3390/s23042066

Hayajneh, A. M., Hafeez, M., Zaidi, S. A. R., & McLernon, D. (2024). TinyML Empowered Transfer Learning on the Edge. *IEEE Open Journal of the Communications Society*, *5*, 1656–1672. https://doi.org/10.1109/ojcoms.2024.3373177

Huang, X., Yuan, Y., Chang, C., Gao, Y., Zheng, C., & Yan, L. (2023). Human Activity Recognition Method Based on Edge Computing-Assisted and GRU Deep Learning Network. *Applied Sciences*, *13*(16), 9059. https://doi.org/10.3390/app13169059

Jameer, S., & Syed, H. (2023). A DCNN-LSTM based human activity recognition by mobile and wearable sensor networks. *Alexandria Engineering Journal*, *80*, 542–552. https://doi.org/10.1016/j.aej.2023.09.013

Karakish, M., Elsawaf, A., & Fouz, M. A. (2025). Gait classification and trajectory generation for different daily activities using deep learning on an embedded system. *Discover Applied Sciences*, *7*(12), 1421. https://doi.org/10.1007/s42452-025-07637-3

Kim, S.-U., & Kim, J.-Y. (2025). Improving Human Activity Recognition Through 1D-ResNet: A Wearable Wristband for 14 Workout Movements. *Processes*, *13*(1), 207. https://doi.org/10.3390/pr13010207

Kim, T. (2026). CSI-Chain: A Complete End-to-End Framework for WiFi CSI Sensing. *KSII Transactions on Internet and Information Systems*, *20*(3). https://doi.org/10.3837/tiis.2026.03.008

Knez, T., Machidon, O., & Pejović, V. (2021). Self-Adaptive Approximate Mobile Deep Learning. *Electronics*, *10*(23), 2958. https://doi.org/10.3390/electronics10232958

Kumari, N., Yadagani, A., Behera, B., Semwal, V. B., & Mohanty, S. (2024). Human motion activity recognition and pattern analysis using compressed deep neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *12*(1). https://doi.org/10.1080/21681163.2024.2331052

Lalapura, V. S., Bhimavarapu, V. R., Amudha, J., & Satheesh, H. S. (2024). A Systematic Evaluation of Recurrent Neural Network Models for Edge Intelligence and Human Activity Recognition Applications. *Algorithms*, *17*(3), 104. https://doi.org/10.3390/a17030104

Lamaakal, I., Yahyati, C., Maleh, Y., El Makkaoui, K., & Ouahbi, I. (2026). Explainable Kolmogorov–Arnold Networks for Zero-Shot Human Activity Recognition on TinyML Edge Devices. *Machine Learning and Knowledge Extraction*, *8*(3), 55. https://doi.org/10.3390/make8030055

Lattanzi, E., Donati, M., & Freschi, V. (2022). Exploring Artificial Neural Networks Efficiency in Tiny Wearable Devices for Human Activity Recognition. *Sensors*, *22*(7), 2637. https://doi.org/10.3390/s22072637

Lee, C., Kang, H., & Ju, K. S. (2025). Location-Aware Ambient Sound Model Adaptation for On-Device Human Activity Recognition in Living Spaces. *IEEE Access*, *13*, 84911–84924. https://doi.org/10.1109/access.2025.3568828

Li, Y., Yin, R., Kim, Y., & Panda, P. (2023). Efficient human activity recognition with spatio-temporal spiking neural networks. *Frontiers in Neuroscience*, *17*, 3037. https://doi.org/10.3389/fnins.2023.1233037

Mahdi, H. F., Choudhury, T., Yousif, Y. K., Al-Jameel, M., & Abushiba, W. (2024). LSTM Autoencoders for Internet of Things Data Compression and Battery Conservation. *Journal of Soft Computing and Data Mining*, *5*(2), 151–160. https://doi.org/10.30880/jscdm.2024.05.02.011

Malche, T., Upadhyay, G. M., Tharewal, S., Balyan, V., Mishra, V. K., Gupta, G., & Soni, P. K. (2026). A Wearable System for Real-Time Fall Detection on Resource-Constrained Devices. *Future Internet*, *18*(4), 211. https://doi.org/10.3390/fi18040211

Mani, N., Haridoss, P., & George, B. (2023). Evaluation of a Combined Conductive Fabric-Based Suspender System and Machine Learning Approach for Human Activity Recognition. *IEEE Open Journal of Instrumentation and Measurement*, *2*, 1–10. https://doi.org/10.1109/ojim.2023.3301861

Mazzieri, R., Cicciarella, E., Pegoraro, J., Corradi, F., & Rossi, M. (2026). Spatiotemporal radar gesture recognition with hybrid spiking neural networks: balancing accuracy and efficiency. *Neuromorphic Computing and Engineering*, *6*(1), 014014. https://doi.org/10.1088/2634-4386/ae4648

Mekruksavanich, S., & Jitpattanakul, A. (2024). Device Position-Independent Human Activity Recognition with Wearable Sensors Using Deep Neural Networks. *Applied Sciences*, *14*(5), 2107. https://doi.org/10.3390/app14052107

Mekruksavanich, S., & Jitpattanakul, A. (2025). Efficient Wearable Sensor-Based Activity Recognition for Human–Robot Collaboration in Agricultural Environments. *Informatics*, *12*(4), 115. https://doi.org/10.3390/informatics12040115

Moghbelan, Y., Esposito, A., Zyrianoff, I., Spaletta, G., Borgo, S., Masolo, C., Ballarin, F., Seidita, V., Toni, R., Barbaro, F., Di Conza, G., Quartulli, F. P., & Di Felice, M. (2024). A Smart Motor Rehabilitation System Based on the Internet of Things and Humanoid Robotics. *Applied Sciences*, *14*(24), 11489. https://doi.org/10.3390/app142411489

Mohamed, S. A., & Martinez-Hernandez, U. (2023). A Light-Weight Artificial Neural Network for Recognition of Activities of Daily Living. *Sensors*, *23*(13), 5854. https://doi.org/10.3390/s23135854

Noor, N., & Park, I. K. (2024). Factorized 3D-CNN for Real-Time Fall Detection and Action Recognition on Embedded System. *IEEE Access*, *12*, 112852–112863. https://doi.org/10.1109/access.2024.3443618

Novac, P.-E., Boukli Hacene, G., Pegatoquet, A., Miramond, B., & Gripon, V. (2021). Quantization and Deployment of Deep Neural Networks on Microcontrollers. *Sensors*, *21*(9), 2984. https://doi.org/10.3390/s21092984

Novac, P.-E., Pegatoquet, A., Miramond, B., & Caquineau, C. (2022). UCA-EHAR: A Dataset for Human Activity Recognition with Embedded AI on Smart Glasses. *Applied Sciences*, *12*(8), 3849. https://doi.org/10.3390/app12083849

Patil, M. S., Sarode, H. J., Banubakode, A., Patil, P. T., Patil, N., Varadarajan, V., & Dewi, D. A. (2026). Inquisitive biometric feature analysis and implementation for recognition tasks using camouflaged segmentation with AI and IoT. *International Journal of Reconfigurable and Embedded Systems (IJRES)*, *15*(1), 119. https://doi.org/10.11591/ijres.v15.i1.pp119-129

Rivas-Caicedo, J. L., Niño-Tejada, K., Saldaña-Aristizabal, L., & Patarroyo-Montenegro, J. F. (2025). A Distributed Wearable Computing Framework for Human Activity Classification. *Electronics*, *14*(16), 3203. https://doi.org/10.3390/electronics14163203

Roh, H., Shin, S., Han, J., & Lim, S. (2021). A deep learning-based medication behavior monitoring system. *Mathematical Biosciences and Engineering, 18*(2), 1513–1528. https://doi.org/10.3934/mbe.2021078

Saeed, U., Kumar, K., Laghari, A. A., Ahmed Khuhro, M., Islam, N., Ghulam, M. S., Hussain, F., & Ahmed Shaikh, A. (2021). A comparative analysis of classification techniques for human activity recognition using wearable sensors and smart-phones. *EAI Endorsed Transactions on Pervasive Health and Technology*, *8*(30), e1. https://doi.org/10.4108/eai.2-11-2021.171752

Sakka, M., Orabi, O., Alnakri, Y., & Bahrami, M. R. (2025). PWMF-ResMiniNet: A Novel Lightweight Deep Learning and Fusion Framework Combining Precision-Weighted Integration of Multiple CNNs for Wearable Sensor-Based Human Activity Recognition. *IEEE Access*, *13*, 209534–209555. https://doi.org/10.1109/access.2025.3642079

Scrugli, M. A., Blažica, B., Raffo, L., & Meloni, P. (2023). A Microcontroller-Based Platform for Cognitive Tracking of Sensorimotor Training. *IEEE Access*, *11*, 70778–70794. https://doi.org/10.1109/access.2023.3294097

Shakerian, A., Douet, V., Nejati, A. S., & Landry, R. (2023). Real-Time Sensor-Embedded Neural Network for Human Activity Recognition. *Sensors, 23*(19), 8127. https://doi.org/10.3390/s23198127

Shen, Z., Howard, N., & Nunez-Yanez, J. (2022). Big–Little Adaptive Neural Networks on Low-Power Near-Subthreshold Processors. *Journal of Low Power Electronics and Applications*, *12*(2), 28. https://doi.org/10.3390/jlpea12020028

Shin, K., Lim, J., & Baek, Y. (2025). Graph Neural Network Compression via Cross-Modal Distillation for On-Device mmWave Human Activity Recognition. *IEEE Access*, *13*, 205148–205160. https://doi.org/10.1109/access.2025.3635398

Singh, H., Singh, H., Sawhney, R. S., Singh, K., & Singh, K. (2026). Eco-Friendly Corn Silk-Based Triboelectric Nanogenerator Sensor for Automated Human Motion Recognition using Boosting-Driven Machine Learning. *Precision Chemistry*, *4*(4), 477–495. https://doi.org/10.1021/prechem.5c00180

Suwannarat, K., & Kurdthongmee, W. (2021). Optimization of deep neural network-based human activity recognition for a wearable device. *Heliyon*, *7*(8), e07797. https://doi.org/10.1016/j.heliyon.2021.e07797

Tan, Y., Wong, Y., & Radzi, S. (2021). Brain-inspired spiking neural networks for Wi-Fi based human activity recognition. *Jordanian Journal of Computers and Information Technology*, *7*(4), 1. https://doi.org/10.5455/jjcit.71-1629096728

Thottempudi, P., Acharya, B., & Moreira, F. (2024). High-Performance Real-Time Human Activity Recognition Using Machine Learning. *Mathematics*, *12*(22), 3622. https://doi.org/10.3390/math12223622

Viswanathuni, V. R. S., Yakkati, R. R., Yeduri, S. R., & Cenkeramaddi, L. R. (2025). WiCNNAct: Wi-Fi-Based Human Activity Recognition Utilizing Deep Learning on the Edge Computing Devices. *IEEE Access*, *13*, 104173–104183. https://doi.org/10.1109/access.2025.3579649

Yamane, T., Kimura, M., & Morita, M. (2025). Effects of Sampling Frequency on Human Activity Recognition with Machine Learning Aiming at Clinical Applications. *Sensors*, *25*(12), 3780. https://doi.org/10.3390/s25123780

Yang, L., Du, W., & Zhao, Y. (2023). A lightweight temporal attention-based convolution neural network for driver's activity recognition in edge. *Computers and Electrical Engineering, 110*, 108861. https://doi.org/10.1016/j.compeleceng.2023.108861

Yen, C.-T., Liao, J.-X., & Huang, Y.-K. (2021). Feature Fusion of a Deep-Learning Algorithm into Wearable Sensor Devices for Human Activity Recognition. *Sensors*, *21*(24), 8294. https://doi.org/10.3390/s21248294

Yun, S.-H., Kim, H.-J., Ryu, J.-K., & Kim, S.-C. (2023). Fine-Grained Motion Recognition in At-Home Fitness Monitoring with Smartwatch: A Comparative Analysis of Explainable Deep Neural Networks. *Healthcare*, *11*(7), 940. https://doi.org/10.3390/healthcare11070940

Zaoui, C., Aharrane, N., Benabbou, F., Zaouch, A., & Bouaine, C. (2025). Model-Oriented Orchestration for Federated Learning: Personalization Through Intelligent Client Matching. *IEEE Access*, *13*, 158585–158602. https://doi.org/10.1109/access.2025.3606865

Zaoui, C., Benabbou, F., & Ettaoufik, A. (2023). Edge-Fog-Cloud Data Analysis for eHealth-IoT. *International Journal of Online and Biomedical Engineering (IJOE)*, *19*(07), 184–199. https://doi.org/10.3991/ijoe.v19i07.38903

Zhao, H. (2025). Research on privacy protection and performance optimization of 6G communication network based on the fusion of federated learning and edge computing. *Discover Artificial Intelligence*, *6*(1), 41. https://doi.org/10.1007/s44163-025-00687-x

Zhou, H., Zhang, X., Feng, Y., Zhang, T., & Xiong, L. (2025). Efficient human activity recognition on edge devices using DeepConv LSTM architectures. *Scientific Reports*, *15*(1), 13830. https://doi.org/10.1038/s41598-025-98571-2