A Rapid Response System for Elderly Safety Monitoring Using Progressive Hierarchical Action Recognition

Han Sun, *Student Member, IEEE*, and Yu Chen, *Senior Member, IEEE*

***Abstract*— The global trend of population aging presents an urgent challenge in ensuring the safety and well-being of elderly individuals, especially those living alone due to various circumstances. A promising approach to this challenge involves leveraging Human Action Recognition (HAR) by integrating data from multiple sensors. However, the field of HAR has struggled to strike a balance between accuracy and response time. While technological advancements have improved recognition accuracy, complex algorithms often come at the expense of response time. To address this issue, we introduce an innovative asynchronous detection method called Rapid Response Elderly Safety Monitoring (RESAM), which relies on progressive hierarchical action recognition and multi-sensor data fusion. Through initial analysis of inertial sensor data using Kernel Principal Component Analysis (KPCA) and multi-class classifiers, we efficiently reduce processing time and lower the false-negative rate (FNR). The inertial sensor identification serves as a pre-filter, enabling the identification of filtered abnormal signals. Decision-level data fusion is then executed, incorporating skeleton image analysis based on ResNet and the inertial sensor data from the initial step. This integration enables the accurate differentiation between normal and abnormal behaviors. The RESAM method achieves an impressive 97.4% accuracy on the UTD-MHAD database with a minimal delay of 1.22 seconds. On our internally collected database, the RESAM system attains an accuracy of 99%, ranking among the most accurate state-of-the-art methods available. These results underscore the practicality and effectiveness of our approach in meeting the critical demand for swift and precise responses in healthcare scenarios.**

***Index Terms*— Information fusion, action recognition, neural networks.**

I. Introduction

WE ARE witnessing a significant global demographic shift. In 2019, 9% of the world's population was aged 65 or older, projected to reach 16% by 2050, notably impacting Europe and North America. The number of individuals aged 80 and above is expected to triple from 143 million in 2019 to 426 million in 2050 [1]. This aging population trend transcends borders, affecting advanced economies on a global scale [2]. The consequence is a burgeoning demand for healthcare services, particularly concerning the health and well-being of elderly individuals who often reside independently within residential communities or extensive nursing facilities [3], [4]. The multifaceted challenges they confront include limited access to healthcare services, aggravated by physical limitations that curtail their mobility [5]. These challenges render seniors more susceptible to accidents or medical emergencies. Managing their health, medications, and chronic conditions, especially for those with multiple ailments, poses significant hurdles. Human Activity Recognition (HAR)-based surveillance algorithms gained prominence in response.

The proliferation of Internet of Things (IoT) technology has introduced fresh avenues for tackling sensor-based HAR challenges, notably utilizing time-series data from wearable devices [6]. Accelerometers and gyroscopes, compact and widespread in low-cost devices, play a key role in HAR. However, inertial sensor-based action recognition, while fast [7], [8], falls short of video-based systems that benefit from richer contextual cues. For example, deep learning-based fall detection achieves just 86% accuracy when relying solely on accelerometer data from wrist-worn devices [9] due to the limitations of single-context accelerometer data lacking 3D information for discerning wrist movements during falls [10].

HAR based on RGB images shows advantages in accuracy [11]. RGB images contain rich information, including color, texture, and spatial relationships, allowing for a comprehensive understanding of human behavior in their surroundings. Deep learning methods such as convolutional neural networks (CNNs) have successfully extracted meaningful features from RGB images and achieved high accuracy in HAR tasks [12], [13]. The ability to capture scene context enables RGB-based HAR systems to recognize complex activities and interactions beyond the limitations of inertial sensor-based approaches. However, HAR systems based on RGB images face challenges in computational complexity and data storage requirements, especially with high-resolution videos. In addition, deploying cameras for human monitoring

Manuscript received 21 November 2023; revised 11 April 2024 and 5 May 2024; accepted 31 May 2024. Date of publication 4 June 2024; date of current version 7 June 2024. *(Corresponding author: Yu Chen.)*

The authors are with the Department of Electrical and Computer Engineering, Binghamton University, Binghamton, NY 13902 USA (e-mail: hsun28@binghamton.edu; ychen@binghamton.edu).

Digital Object Identifier 10.1109/TNSRE.2024.3409197

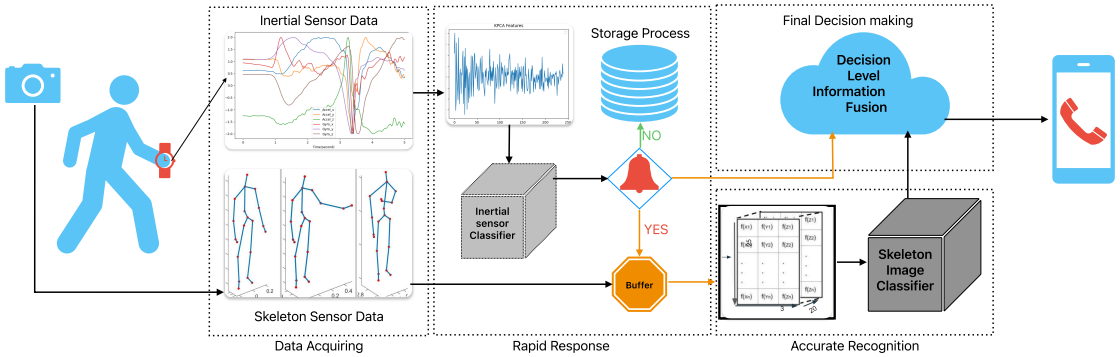

Fig. 1. Rapid response elderly safety monitoring (RESAM) system architecture.

may raise privacy concerns, limiting the locations where the cameras can be installed.

Therefore, the skeleton image is introduced to avoid the issues. By representing only key joint positions and motions, skeletal data ensures privacy while capturing human behavior's underlying spatial relationships and temporal dynamics [14]. Skeleton data has a small footprint, improving computational efficiency and reducing storage requirements. Deep learning models, such as graph convolutional networks (GCNs) [15] or recurrent neural networks (RNNs) [16], efficiently extract features from skeleton data for accurate, real-time action recognition. Fusing skeletal data with other sensor inputs, such as inertial data from wearable devices, enables multimodal HAR systems to provide comprehensive insights into human activity while preserving privacy and ethical considerations [17].

Information fusion [18] is a pivotal aspect of HAR, enhancing system accuracy by integrating data from various sources or modalities. Fusion occurs at three levels: data, feature, and decision [13]. Data level fusion combines raw sensor data, like accelerometer and gyroscope inputs, to create comprehensive datasets. Feature level fusion merges relevant features from heterogeneous data sources, such as the combination of the eigenvectors of the RGB and depth image into a single one. Decision-level fusion integrates information from multiple sources to reach final decisions. This fusion strategy enables HAR systems to recognize diverse activities while enhancing system robustness accurately.

In healthcare for elders, a timely response is critical to ensuring their safety and well-being. This paper proposes a Rapid Response Elderly Safety Monitoring (RESAM) system. This progressive HAR method combines the advantages of wearable inertial data and Skeleton Images while mitigating their respective disadvantages. By integrating information from both modalities, the RESAM system can gain a comprehensive picture of older people's activities, enabling faster and more accurate identification of their behavior. The RESAM system performs action recognition based on the motion details of the wearable device and quickly responds to potentially dangerous behaviors. In addition, skeletal data allows more precise identification results while maintaining privacy. Furthermore, the data fusion of the two recognition methods enhances the overall accuracy and robustness of the system. This holistic approach maximizes the potential of healthcare technology, ensuring prompt and effective responses to support and safeguard the well-being of the aging population. The key features are summarized below.

- **Real-time response:** The RESAM system prioritizes real-time responses to potentially dangerous behaviors of seniors. The system can quickly detect and identify critical activities by utilizing wearable device data and efficient motion recognition algorithms, enabling rapid intervention and assistance when needed.
- **Privacy-preserving human action recognition:** Skeleton data integration for human activity recognition ensures privacy preservation while maintaining accurate action recognition. Skeletal data represents key joint positions and motions without processing or storing detailed visual information, addressing privacy concerns and ethical considerations in healthcare settings.
- **Enhanced accuracy and robustness:** The RESAM system improves the accuracy and robustness of human activity recognition by fusing decisions from multiple modalities, including wearable device input and skeletal data. Combining information from the two sources allows for a more complete understanding of older adults' activities, leading to improved performance and reliability in recognizing and responding to various behaviors.

The rest of the paper is organized as follows. Section II illustrates the rationale and detailed design of the proposed RESAM system. In Section III, we introduce the datasets, evaluation metrics, experimental setup, and results cooperation. Finally, Section IV concludes the paper, summarizing the findings and suggesting future research directions.

II. RESAM: Rationale and Methods

Figure 1 illustrates the architecture of the RESAM system. It begins with the crucial step of data extraction, where skeletal and inertial sensor data are collected and labeled with corresponding tags. After that, the system performs feature extraction on the inertial sensor data, utilizing it for initial classification. If the detection results signify an emergency, the subsequent action involves obtaining more precise outcomes through the fusion of skeleton-based detection, which aligns with the inertial sensor data labels. Finally, whether it is an emergency or not, all data is stored in the cloud for future long-term health state analysis.

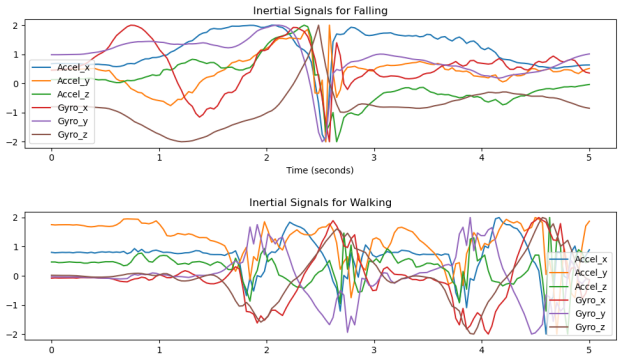

Fig. 2. Examples of inertial signal for normal and abnormal activities.

A. Inertial Data Pre-Processing

Thanks to the advancement of IoT technology, modern wearable devices are equipped with multiple sensors, such as accelerometers and gyroscopes, providing rich multi-dimensional data. Therefore, in motion recognition systems based on inertial sensors, the collected data is usually represented as a d-dimensional vector, where d is the number of sensor channels in the wearable device.

Figure 2 shows an example of both normal and abnormal inertial signals, highlighting the difference between them. The data from a wearable device containing accelerometers and gyroscopes will be represented as multiple three-dimensional vectors because each sensor measures acceleration or angular velocity along three orthogonal axes (x, y, z). Therefore, the resulting data consists of d-dimensional vectors, where $d = 6$, counting three channels from the accelerometer and three from the gyroscope. A d-dimensional vector encapsulates the measurements of all sensor channels at a particular time, forming raw sensor data collected over time. However, the information contained in the original data is not all useful. We need the features to distinguish different actions and remove the part polluted by noise.

The dataset of inertial data includes N samples in d dimensions. Each sample can be represented as a d-dimensional vector $X_i \in R^d$, where $i = 1, 2, \ldots, n$.

B. KPCA-Based Inertial Signal Analysis

Kernelized principal component analysis (KPCA) is a nonlinear dimensionality reduction method [19]. The basic idea is to map the original linearly inseparable data to a high-dimensional space through the kernel method, making it linearly separable in the high-dimensional space; The space is still linearly separable. The first step in KPCA is to compute the kernel matrix K, which measures the similarity between pairs of data points in the original space. The most commonly used kernel is the Radial Basis Function (RBF) kernel, defined as:

$$K(X_i, X_j) = exp(-\gamma \|X_i - Xj\|^2) \tag{1}$$

where γ is a parameter known as the kernel bandwidth, and $\|.\|$ the Euclidean distance between X_i and X_j.

Once the kernel matrix K is computed, the next step is to center it by subtracting the mean of each row and each column and then double-centering the matrix. The centered kernel matrix is:

$$\tilde{K} = (I_n - \frac{1}{N} 1_n 1_n^T) K (I_n - \frac{1}{N} 1_n 1_n^T)^T \tag{2}$$

where I is the identity matrix, 1 is a column vector of ones, and 1^T is its transpose.

After that, the $\tilde{K}$ eigenvalues λ and eigenvectors v are calculated. The eigenvectors represent the principal components of the data in the higher-dimensional feature space, and the eigenvalues indicate the variance captured by each principal component.

Finally, we select the top k eigenvectors corresponding to the k largest eigenvalues to form the projection matrix $W \in R^{nk}$, where k is the desired dimensionality of the feature space. The transformed data in the feature space is obtained as:

$$\phi(x_i) = W^T K(x_i, X) \tag{3}$$

where X is the original dataset matrix, and $\phi(x_i)$ is the feature representation of the data point x_i in the higher-dimensional space.

Using KPCA on inertial sensor data yields a potent feature representation, capturing nonlinear relationships and discriminative information and enhancing action recognition performance. KPCA transforms the data into a higher-dimensional space, unveiling intricate patterns not apparent in the raw data and preparing data for the classifier.

C. Rapid Response for Inertial Signal

The transformed data is prepared for classification after the feature extraction process using KPCA. In this study, multiple classifiers are individually employed to achieve the best performance. The selected classifiers include Support Vector Machine (SVM), Random Forest, and XGBoost. Each classifier is chosen based on its specific strengths and performance characteristics.

1) Support Vector Machine: The core principle of Support Vector Machine (SVM) involves identifying the optimal dividing boundary within a hyperplane to differentiate various categories [20]. Consequently, when dealing with KPCA-processed data, enhanced outcomes are frequently achieved due to the alignment within the system. SVM was initially designed for binary problems, so when faced with multi-classification issues, the natural idea is to construct multiple binary devices and combine them, called the One-against-all (OVA) algorithm. The m-th SVM is trained with all of the examples in the m-th class with positive labels and all other examples with negative labels. Thus given l training data $(x_1, y_1), \ldots, (x_l, y_l)$, where $x_i \in R^n, i = 1, 2 \ldots\ldots l$ and $y_i \in 1, \ldots, k$ is the class of x_i, the m-th SVM solves the following problem:

$$\min_{w^w, b^w, \ \xi^w} \frac{1}{2} \left\| w^k \right\|^2 + C \sum_{i=1}^{l} \xi_i^m \tag{4}$$

$$subject\ to \quad w_k^T \phi(x_i) + b_k \geq 1 - \xi_i^m, if\ \ y_i = m \tag{5}$$

$$w_k^T \phi(x_i) + b_k \leq -1 + \xi_i^m, if\ \ y_i \neq m \tag{6}$$

$$\xi_i^m \geq 0, i = 1, \ldots..l \tag{7}$$

where the training data x_i are mapped to a higher dimensional space by the function ϕ and C is the penalty parameter. When data are not linearly separable, there is a penalty term $C\sum_{i=1}^{l}\xi_i^m$, which can reduce the number of training errors. After solving the equation, the k-th decision function is obtained as:

$$f_{(x)} = w_k^T \phi(x) + b_k \tag{8}$$

When there are multiple decision outcomes with positive outputs, the input x will be classified as the one with the largest decision function value: $\hat{y} = argmax(f_i)$.

2) Random Forest: The Random Forest classifier operates through an ensemble of decision trees, where each tree independently predicts the class label of an input instance [21]. The final prediction is determined by aggregating the individual predictions through majority voting. The Random Forest construction process involves the following steps:

1) **Bootstrap Aggregating (Bagging)**: A sample of size N is drawn N times with replacement. This process generates N samples, forming the foundation for decision tree training. Each of these N samples is employed to train a decision tree, serving as the samples at the root node.
2) **Feature Subsetting:** With each sample containing M attributes, during the formation of the decision tree, at every node's split, n attributes are randomly selected from the M attributes, where $m \ll M$.
3) **Node Splitting:** For each decision tree node, a split attribute is chosen based on a criterion such as information gain. Each node must be split according to this attribute-selection strategy until further splitting is infeasible. The attribute used for splitting in the parent node is avoided for selection in the child nodes.
4) **Ensemble Formation:** This process is repeated numerous times, generating many decision trees. The ensemble of these trees constitutes the Random Forest.
5) **Prediction and Aggregation:** During prediction, an input instance is passed through each tree to obtain individual class predictions. The final prediction is determined through majority voting among the trees.

Random Forests create an ensemble of different decision trees, each trained on a different subset of data and with a different choice of attributes. Collective predictions of these trees yield robust and accurate classification models.

3) XGBoost: XGBoost (eXtreme Gradient Boosting) is an integrated learning algorithm in the integrated learning category [22]. It excels in handling sequential data, making capturing temporal dynamics in human actions practical. It manages high-dimensional inertial signal data and automatically selects the most informative features, simplifying feature engineering. As an ensemble learning method, XGBoost combines multiple models, typically decision trees, which is advantageous in recognizing complex actions. This approach mitigates the potential biases of individual sensors and sensor noise. XGBoost also offers model interpretability through feature importance scores, aiding in identifying the most relevant sensor measurements for action recognition. Its gradient-boosting mechanism allows iterative error correction, adapting well to the sequential nature of inertial data. Furthermore, XGBoost includes techniques to handle imbalanced datasets, a common challenge in human action recognition, ensuring accurate recognition across all action classes. In summary, XGBoost's capabilities in sequential data analysis, feature selection, ensemble learning, interpretability, and adaptation to inertial signal characteristics make it a robust choice for inertial signals-based HAR.

The tree model used is the CART regression tree model, which assumes a binary tree structure, repeatedly dividing based on features. For example, a node splits using the jth feature's eigenvalue: samples below the threshold s go left, others right, shown as

$$R_1(j,s) = x|x^{(j)} \le s \ and \ R_2(j,s) = x|x^{(j)} > s \tag{9}$$

Following this core concept, the process involves successive feature segmentations to expand the tree. By iteratively adding trees, we are learning new functions to match previously predicted residuals. After training with k trees, predicting a sample's score involves directing it to a leaf node in each tree, with each leaf node representing a score. Ultimately, the expected value of a sample is the sum of the scores corresponding to each tree. The comprehensive model for generating a decision tree is

$$\hat{y} = \phi(x_i) = \sum_{K}^{k=1} f_k(X_i) \tag{10}$$

where $F = f(x) = w_{q_x}(q : R^m \to T, w \in R^T)$ is a collection of all classification and regression trees, x_i is a feature vector, q is the structural information contained in the leaf nodes of the corresponding classification and regression trees, T is the number of leaf nodes on the corresponding classification regression tree, and each classification regression tree corresponds to its structural information q and leaf nodes weight w. The objective function definition of XGBoost is:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_j) + \sum_{k=1}^{K} \Omega(f_k) \tag{11}$$

where l indicates the selected loss function, which calculates the error between the predicted value $\hat{y}$ and the real value y, and the part after the plus sign is the regular term, which reduces the complexity of the model and alleviates the over-fitting of the model.

As mentioned above, the newly generated tree needs to fit the residual of the last prediction, so after iteration, the model for generating the t-th tree is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{12}$$

Therefore, the target formula can be expressed as the sum of multiple iterations:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)) + \Omega(f_t) \tag{13}$$

The next step is to find a f_t that can minimize the objective function. The idea of XGBoost is to approximate it with

its Taylor second-order expansion at $f_t = 0$. Therefore, the objective function is approximated as:

$$L^t \simeq \sum_{i=1}^{n}[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(X_i) + \frac{1}{2}h_i f_t^2(X_i)] + \Omega(f_t) \tag{14}$$

where $g_i = \partial_{\hat{y}(t-1)} l(y_i, \hat{y}^{(t-1)})$ is the first derivative and the $h_i = \partial^2_{\hat{y}(t-1)} l(y_i, \hat{y}^{(t-1)})$is the second derivative. Since the prediction scores of the first $(t-1)$ trees and the residual of y do not affect the optimization of the objective function, they can be removed directly. At the same time, each sample will eventually fall into a leaf node, allowing us to utilize identical leaf nodes. Consequently, point samples are regrouped. The objective function can be rewritten as a quadratic function of the leaf node score w, and the final calculation formula for the optimal w and objective function value using the vertex formula is as follows:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{15}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{16}$$

where T is the leaf node number, w is the L2 regularization of the leaf node score. When an internal node splits, if the loss function value is less than γ, then the split stops. λ is a similar penalty coefficient. G and H are summations of g_i and h_i, respectively.

After obtaining the final objective function, it is only necessary to continuously generate the optimal classification regression tree and integrate it into the existing model to form the final XGBoost model for classification. The fast response these algorithms produce provides initial judgment on the data obtained from the inertial sensors. However, the captured target actions need further evaluation due to the algorithm's potential inaccuracy. Therefore, the next step is an accurate classification based on the skeleton image captured simultaneously with the inertial sensors.

D. Skeleton Image Pre-Processing

The Kinect sensor can construct a simplified human skeleton model using 20 key points, as shown in Fig. 3, without needing all 206 bones. Each joint point's spatial coordinates are denoted as well $P(x, y, z)$, where x and y represent the abscissa and ordinate, respectively, and z corresponds to the distance from the camera to the human body. During movement, the relative positions of these joints change. To better represent the offsets of limb joint points with the hip joint and remove the camera distance effect, the central node of the hip is used as the central origin. The formula to calculate the initial spatial position feature is given by:

$$f = p_n - p_{hip}(n = 2, 3, \ldots . N). \tag{17}$$

where $p_n$represents the other nodes excluding the hip joint, and p_{hip} is the hip-center joint.

Therefore, the differences in X, Y, and Z coordinates are obtained using the 3D matrix vector calculation. These differences form the feature vectors for the m-th frame: $f_m = [f_x^m, f_y^m, f_z^m]$ with the size of 19×3. An entire action can be represented as a set of these feature vectors for all frames:

$$F = [f_1, f_2, \ldots . . f_M] \tag{18}$$

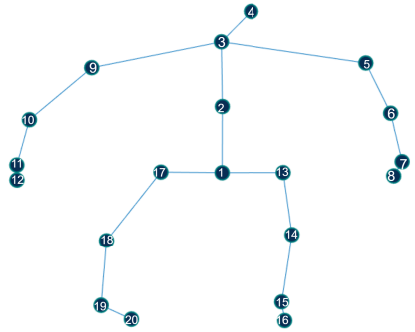

Fig. 3. Skeleton Image. The points are: 1. hip center, 2. middle-spine, 3. shoulder center, 4. Head, 5. Left shoulder, 6. Left elbow, 7. Left wrist, 8. Left-hand, 9. Right-shoulder, 10. Right elbow, 11. Right-wrist, 12. Right-hand, 13.Left-hip, 14. Left knee, 15. Left-ankle, 16. Left-foot, 17. Right-hip, 18. Right knee, 19. Right-ankle, 20. Right-foot.

Due to the varying heights of individuals, the coordinate values of skeletons can differ. Before training the database, a normalization process is conducted. Point 1 represents the hip center, while point 2 denotes the middle of the spine. The middle spine length is defined as the Euclidean distance between these points. This process involves determining the maximum length of the spine across the entire dataset and establishing it as a fixed parameter Max_Middle_Spine. This normalization aims to ensure that every sample has the same length of the middle spine, thereby canceling out the effect of different heights of individuals. Therefore, the final normalized action space feature vector is:

$$\bar{f} = \frac{f}{distance(P_1, P_2)/(Max\_Middle\_Spine)} \tag{19}$$

$$\bar{F} = [\bar{f}_1, \bar{f}_2, \ldots . . \bar{f}_M] \tag{20}$$

E. ResNet-Based Accurate Classification

Residual Networks (ResNet) represent a significant advancement in deep learning architectures, specifically designed to address the vanishing gradient problem that can hinder the training of very deep neural networks. ResNet introduces the concept of residual blocks, which allows gradients to flow more effectively during backpropagation, enabling the training of much deeper networks.

Each residual block consists of skip connections, also known as shortcut connections, which bypass one or more layers. This helps mitigate the vanishing gradient problem and enables the training of deeper networks. The output of a residual block is a combination of its input and the output of the internal layers, creating a *residual*; the output of a residual block can be represented as:

$$Output = Input + F(Input) \tag{21}$$

where F represents the transformation performed by the internal layers of the block. This formulation enables the network

TABLE I
RESNET-20 ARCHITECTURE AND PARAMETER SIZES

Layer	Input Size	Filter Size/Parameters	Output Size
$Conv1$	[19, time, 03]	16 filters (3x3)	[19, time, 16]
$Conv2_x$	[19, time, 16]	2x Convolution (16 filters)	[19, time, 16]
$Conv3_x$	[19, time, 16]	2x Convolution (16 filters)	[19, time, 16]
GAP	[19, time, 16]	No parameters	[1, 1, 16]
FC	[1, 1, 16]	Variable	[1, 1, Classes]
Softmax	[1, 1, Classes]	No parameters	[1, 1, Classes]

to learn the residual transformation, making optimizing and learning more complex features easier. In terms of layers, a typical ResNet architecture includes several convolutional layers and residual blocks. The specific architecture can vary, but the basic structure involves stacking multiple blocks together. The architecture often starts with initial convolutional and pooling layers, then a series of residual blocks, and concludes with fully connected layers for classification.

ResNet-20 is a specialized architecture designed to work efficiently with small-scale images, which is particularly relevant for scenarios like our skeleton images after processing [23]. In our case, the action represented in the skeleton matrix is $19 \times 3 \times 20$, effectively making it a small image with three channels after transposition. The ResNet-20 architecture is well-suited for such compact images and is optimized for tasks like human action recognition using skeleton data. It allows for effective feature extraction and classification, which is crucial for accurately recognizing actions based on skeletal information.

Table I shows the parameters and architecture of ResNet20. "Time" is the number of frames used to represent the whole action. "Input Size" represents the dimensions of the input Skeleton Image. The "Filter Size/Parameters" column describes the filter sizes and the number of parameters for each layer. "Output Size" shows the dimensions of the output feature maps or layers, while GAP means Global Average Pooling. The parameter count for the Fully Connected (FC) layer depends on the specific number of classes in the classification task. ResNet-20 allows the residual blocks to reach the desired depth while performing well, particularly on small-scale image classification tasks.

F. Decision-Level Information Fusion

In the last step of our RESAM scheme, the Dempster-SHAFER evidence theory (D-S theory) is adopted to integrate decision-level data [24]. The D-S theory is a robust framework for reasoning and combining data from various sources when processing uncertainty. In RESAM, we use the function of basic probability distribution (BPA).

Two distinct sources of evidence, I and S, are considered to apply the DS-Theory on the decision-level fusion. Each provides degrees of belief regarding various hypotheses within a frame of discernment, denoted as Θ, where $\Theta = H_1, H_2 \ldots\ldots H_N$, including $2^{\Theta} = A/A \in \Theta = \emptyset, H_i, H_2 \ldots\ldots H_N$ subsets. The following equations calculate the Belief (Bel):

$$m(A) = \frac{1}{1-k} \sum_{B \cap C = A} m_I(B) m_S(C) \tag{22}$$

$$k = \sum_{B \cap C = \varnothing} m_I(B) m_S(C) \tag{23}$$

where the $m_I(B)$ and $m_S(C)$ are the mass assigned to hypotheses B and C by Inertial-based HAR(I) and Skeleton-based HAR(S).

To combine evidence from both sources I and S, we employ Dempster's rule of combination to compute the Belief (Bel) and Plausibility (Pl) functions for hypothesis A:

$$Bel(A) = \sum_{B \in A} m(B) \tag{24}$$

$$Pl(A) = \sum_{A \cap B \neq \varnothing} m(B) \tag{25}$$

Here, Bel(A) represents the overall degree of belief in hypothesis A, considering both sources I (Inertial Sensor-based HAR) and S (Skeleton-based HAR). On the other hand, Pl(A) indicates the total degree of support for hypothesis A across both sources. The final decision-making process involves comparing the maximum Pl(A) and the maximum Bel(A), allowing us to choose the most plausible and believable hypothesis.

These functions are crucial for fusing Inertial Sensor-based HAR and Skeleton-based HAR evidence. Calculating belief and plausibility values enables comprehensive decision-making in human action recognition scenarios, ensuring that the system makes informed and accurate judgments based on multiple sources of evidence.

III. EXPERIMENTAL RESULTS

A. Database and Evaluation Method

1) UTD-MHAD Database: Due to the time consistency requirements of the inertial sensor and skeleton data in the system design, the experimental data set must contain two kinds of data collected concurrently. Therefore, the UTD-MHAD data set is selected because it is deliberately designed to include a single Kinect camera and wearable inertial sensor [25]. The Kinect camera captures color images at 640 × 480 pixels and 16-bit depth images at 320 × 240 pixels, running at approximately 30 frames per second. The wearable inertial sensors used in the dataset were developed at the ESSP lab at the University of Texas at Dallas. The sensor consists of a 9-axis MEMS sensor that captures 3-axis acceleration, 3-axis angular velocity, and 3-axis magnetic field strength. It integrates a 16-bit low-power microcontroller, a dual-mode Bluetooth low-energy unit for wireless data transfer to a laptop/PC and a serial link between the MEMS sensor and the microcontroller for control commands and data transfer.

The UTD-MHAD dataset covers 27 actions, including arm swings, hand waves, clapping, and motion-related movements. Motions were performed using wearable inertial sensors on the subjects' right wrist or right thigh, depending on whether the action was primarily arm- or leg-based. Exercises 1 through 21 have the sensor on the right wrist, while drills 22 through 27 use the right thigh position. This database provides a valuable resource for studying and analyzing human motion in various scenarios and activities.

TABLE II
COUNT OF ACTIONS FOR SELF-COLLECT DATABASE

Action	Walking	Retrieving	Writing	Squatting
Count	32	31	33	32
Action	Standing Up	Getting Up	Sitting Down	Falling
Count	31	30	32	35
Action	Coughing	Headache	Stomach Pains	Vomiting
Count	33	34	30	36

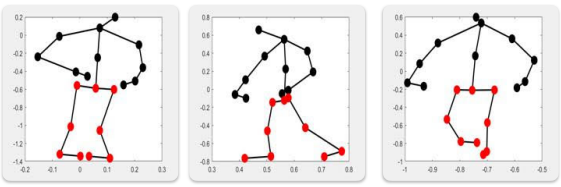

Fig. 4. Skeleton images for the same action from 3 cameras.

2) Self-Collected Database: While public databases can be used for synchronization purposes, they are not well suited to the specific requirements of elderly safety monitoring. As such, they lack action designs to address behaviors associated with potential risks. To fully validate the fast response mechanism and the precision of our RESAM scheme, we conducted laboratory experiments in a simulated home environment, collecting proprietary data.

In this experimental setup, three Kinect V2 cameras were strategically placed in front, to the right, and behind the subject, giving the skeleton image for the actions from different directions simultaneously, as shown in Fig. 4. At the same time, the TicWatch Pro3 is used to collect inertial sensor data. Fifteen volunteers actively participated in the experiment, resulting in 389 datasets. In total, the experiment consisted of 12 different actions. Table II presents the details of the action and number of each class.

This customized experimental setup allowed us to evaluate the system's efficacy in scenarios that mirror real-life situations, fine-tuning its rapid response mechanism and assessing its accuracy.

3) Evaluate Method: In healthcare, it is crucial to recognize that the accuracy rate alone may not adequately gauge diagnostic efficiency. The False Negative Rate (FNR), often called the missed diagnosis rate, assumes pivotal importance. To illustrate, when a routine action is erroneously classified as an emergency, this error can be rectified in a subsequent retest with minimal harm. However, in contrast, if a genuinely hazardous action is mistakenly perceived as usual and subsequently disregarded, the repercussions can be severe, including delayed disease diagnosis and the loss of an optimal treatment window.

$$FNR = \frac{FN}{FN + TP} \tag{26}$$

Therefore, clinical practice emphasizes minimizing the FNR, calculated as Eq.26, where FN and TP are the False Negative and True Positive, to ensure that critical conditions are not overlooked.

B. Experimental Results

1) Inertial Sensor Recognition: Table III compares our proposed approach with existing state-of-the-art methods in the inertial sensor-based action recognition field. When applying the inertia data collected by Ticwatch to accurate inertia data, our method obtained the highest accuracy. This indicates the effectiveness of our approach and suggests that it may affect the potential of various applications.

The accuracy of the type of action may decrease slightly compared with other existing studies, but it is still within the acceptability range. This adaptability and robustness make our method a multi-functional tool that recognizes extensive human behavior, even if those were previously considered more complicated or subtle.

Achieving an FNR as low as 1.2% is a significant milestone, signifying the system's ability to maintain a high level of sensitivity and reduce the risk of missed diagnoses.

2) Skeleton Image Based HAR: Recently, skeleton-based action recognition based on deep learning and neural networks has been widely used, as shown in Table IV. In this rapid development landscape, our method offers simplicity and effectiveness. Despite the complexity of challenges brought by bone motion recognition, we have reached the accuracy level of the most complicated algorithms that can be used today.

While complex algorithms can achieve impressive results, it is essential to remember that the balance between simplicity and efficiency is often the top priority in real-world applications. Our method not only meets the strict requirements of modern action recognition but also accomplishes this in a simplified and effective way. The simple algorithm allows our system to deploy in hardware with less powerful computing power, such as home computing centers, and maintain accuracy while responding quickly.

3) Decision-Level Fused Result Analysis: Table V provides a comprehensive comparison with state-of-the-art approaches, offering a detailed analysis encompassing both accuracy and power and time consumption metrics. We have achieved superior accuracy while maintaining efficient power usage and rapid processing times.

Table VI represents the running time of the experiment under different hardware configurations. On high-performance GPUs, such as GeForce 1080, GeForce 4080, and RTX A5000, the system's rapid response time is 2.04 seconds, 1.40 seconds, and 1.22 seconds, respectively, on average. In contrast, the response time extends to 36.5 seconds when using the Raspberry Pi 4 to run the system.

Our RESAM system was initially designed for the safety of older people living alone. What needs to be considered is that our target group may not be able to use the high-performance equipment used in the laboratory environment in actual application scenarios. A report in 2017 showed that the average response time for emergency services after 911 in the urban community was about seven minutes [45]. This response time can be extended to 15 minutes or longer in the rural environment. The result shows that even if the system runs on a low-computing device like a Raspberry Pi 4, it only takes half a minute to respond. Our experimental results

TABLE III
PERFORMANCE COMPARISON OF SKELETON-BASED ACTION RECOGNITION IN TOP-1 ACCURACY (%). THE BEST ONE IS IN BOLD, AND THE SECOND ONE IS UNDERLINED

Reference	Year	# of Act	Method	Accuracy(%)	Sensors
[26]	2009	20	Decision Tree	93	
[27]	2018	5	CNN & LSTM Fused Model	95.35	iPhone
[28]	2019	18	RF	86.8	
[29]	2019	5	XGBoost	84.41	SmartPhone
[30]	2021	18	LSTM	97.158	SmartWatch
[31]	2021	6	RF	90.8	AppleWatch
[32]	2022	8	Markov Model	95	SmartWatch
RESAM	2023	27	SVM RF XGBoost	80.69 91.89 95.27	Wearable Sensors
RESAM	2023	12	XGBoost	97.73	TicWatch

TABLE IV
SKELETON BASED ACTION RECOGNITION

Reference	Year	Activities	Method	Accuracy(%)	Dataset
[33]	2013	10	RF	92	UTKinect
[34]	2019	60	Actional-graph-based	94.2	NTU-RGBD(CV)
[35]	2017	20	Multi-Stream CNN	96.62	MSRC-12(CS)
[36]	2017	27	CNN	88.10	UTD-MHAD
[37]	2022		GCN	93	NTU-RGBD
[38]	2023	120	Learning Discriminative Representations	96.8	NTU-RGBD
RESAM	2023	27	ResNet	95.98	UTD-MHAD
RESAM	2023	12	ResNet	96.22	Self-Collected

TABLE V
COMPARISON OF HAR SYSTEM

Reference	Year	Power consumption	Performance metrics	Time Consumption	Notes
[39]	2014	268mW	85.9-89.7	2.5s	RGB image based HAR applied on FPGA
[40]	2011		> 90	5.7s	3D Axis-based HAR
[41]	2021	0.0209-0.0385mW*h/sample	92.11-99.41	8.653-14.626 ms/sample	Multiply Inertial sensors involved HAR
[42]	2021		93.45		Skeleton and Depth Image Fused HAR
[43]	2018		95.38		Skeleton,Depth and RGB image Fused HAR
[44]	2020		97.71	4.63s	Skeleton and Inertial Fused HAR
RESAM	2023	60.2 mW/sample	97.41 (UTD-MHAD)	1.22s	Skeleton and Inertial Fused HAR
RESAM	2023	58.6 mW/sample	99.02(Self-Collected)	1.24s	Skeleton and Inertial Fused HAR

TABLE VI
RESPONSE TIME ON DIFFERENT HARDWARE

Device	RPi 4	GeForce 1080	GeForce 4080	RTX A5000
Time (S)	36.5	2.04	1.40	1.22

demonstrate its adaptability to low-cost edge devices with limited computational and storage capacity, which aligns with our primary goal.

IV. FUTURE DIRECTION AND CONCLUSION

Action recognition, particularly when involving the fusion of inertial and skeletal data, presents notable challenges, with precise synchronization during data collection being paramount. Maintaining impeccable temporal alignment between these data sources demands meticulous attention and often necessitates specialized hardware setups. Creating a comprehensive database housing synchronized inertial and skeletal data proves essential for achieving accurate fusion-based action recognition results. One promising avenue lies in online medical care within the Metaverse, augmented by digital twins (DT). Harnessing the DT's capabilities, including real-time data integration, historical data storage, and AI-driven enhancements, enables the acquisition of vast information volumes, ultimately yielding more precise recognition outcomes. Moreover, the intricate human-machine interaction within DT models enhances the provision of insightful and actionable recommendations for decision-makers within the context of online healthcare scenarios [46].

In summary, this paper presents a fast and accurate method for human action recognition, harnessing the fusion of inertial sensor data and skeletal information. Our RESAM scheme leverages the unique strengths of both modalities, blending the spatial richness of skeletal data with the temporal dynamics captured by inertial sensors. Experimental results attest to its exceptional performance in precisely identifying diverse human actions. RESAM possesses the ability to deliver high-accuracy results while maintaining a low complexity. Moreover, its adaptability to various hardware configurations underscores its practicality and versatility. We employ the Dempster-Shafer evidence theory to provide a robust framework for decision-level data fusion, capable of integrating information from multiple sources under uncertain conditions, thereby achieving comprehensive and dependable action recognition. RESAM minimizes the risk of missed diagnoses in critical medical applications, where timely and accurate

diagnosis is paramount. This study underscores the significant potential of multimodal fusion in enhancing the accuracy and responsiveness of human action recognition systems, with anticipated advancements in healthcare, security monitoring, and beyond.

REFERENCES

[1] Department of Economic United Nations and Social Affairs, *World Population Prospects 2022: Summary of Results*, document UN DESA/POP/2022/TR/NO. 3, 2022.

[2] S. Juan and P. A. Adlard, "Ageing and cognition," in *Biochemistry and Cell Biology of Ageing: Part II Clinical Science*. Singapore: Springer, 2019, pp. 107–122.

[3] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes—Present state and future challenges," *Comput. Methods Programs Biomed.*, vol. 91, no. 1, pp. 55–81, Jul. 2008.

[4] *Factors That Affect Health-Care Utilization in Health-Care Utilization as a Proxy in Disability Determination*, Nat. Academies Press, Washington, DC, USA, 2018.

[5] H. Sun and Y. Chen, "Real-time elderly monitoring for senior safety by lightweight human action recognition," in *Proc. IEEE 16th Int. Symp. Med. Inf. Commun. Technol. (ISMICT)*, May 2022, pp. 1–6.

[6] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," *Inf. Fusion*, vol. 51, pp. 10–18, Nov. 2019.

[7] A. Raj, A. Subramanya, D. Fox, and J. Bilmes, "Rao-blackwellized particle filters for recognizing activities and spatial context from wearable sensors," in *Proc. 10th Int. Symp. Experim. Robot.* Berlin, Germany: Springer, 2008, pp. 211–221.

[8] Y.-L. Kuo, K. M. Culhane, P. Thomason, O. Tirosh, and R. Baker, "Measuring distance walked and step count in children with cerebral palsy: An evaluation of two portable activity monitors," *Gait Posture*, vol. 29, no. 2, pp. 304–310, Feb. 2009.

[9] T. Mauldin, M. Canby, V. Metsis, A. Ngu, and C. Rivera, "SmartFall: A smartwatch-based fall detection system using deep learning," *Sensors*, vol. 18, no. 10, p. 3363, Oct. 2018.

[10] H. Guo, L. Chen, L. Peng, and G. Chen, "Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2016, pp. 1112–1123.

[11] H. H. Wu, E. D. Lemaire, and N. Baddour, "Change-of-state determination to recognize mobility activities using a BlackBerry smartphone," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 5252–5255.

[12] N. Sikder, Md. S. Chowdhury, A. S. M. Arif, and A.-A. Nahid, "Human activity recognition using multichannel convolutional neural network," in *Proc. 5th Int. Conf. Adv. Electr. Eng. (ICAEE)*, Sep. 2019, pp. 560–565.

[13] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Proc. Int. Workshop Ambient Assist. Living*, Vitoria-Gasteiz, Spain. Cham, Switzerland: Springer, Dec. 2012, pp. 216–223.

[14] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2712–2716.

[15] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 7444–7452.

[16] S. Liao, T. Lyons, W. Yang, K. Schlegel, and H. Ni, "Logsig-RNN: A novel network for robust and efficient skeleton-based action recognition," 2021, *arXiv:2110.13008*.

[17] S. Qiu et al., "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Inf. Fusion*, vol. 80, pp. 241–265, Apr. 2022.

[18] E. Blasch, Y. Chen, G. Chen, D. Shen, and R. Kohler, "Information fusion in a cloud-enabled environment," in *High Performance Cloud Auditing and Applications*. New York, NY, USA: Springer, 2014, pp. 91–115.

[19] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 1997, pp. 583–588.

[20] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, Dec. 2001.

[21] T. Kam Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[23] L. Heim, A. Biri, Z. Qu, and L. Thiele, "Measuring what really matters: Optimizing neural networks for TinyML," 2021, *arXiv:2104.10645*.

[24] K. Sentz and S. Ferson, *Combination of Evidence in Dempster–Shafer Theory*. Livermore, CA, USA: Sandia National Laboratories, 2002.

[25] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 168–172.

[26] A. G. Bonomi, A. H. C. Goris, B. Yin, and K. R. Westerterp, "Detection of type, duration, and intensity of physical activity using an accelerometer," *Med. Sci. Sports Exercise*, vol. 41, no. 9, pp. 1770–1777, 2009.

[27] J. He, Q. Zhang, L. Wang, and L. Pei, "Weakly supervised human activity recognition from wearable sensors by recurrent attention learning," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2287–2297, Mar. 2019.

[28] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and smartwatch-based biometrics using activities of daily living," *IEEE Access*, vol. 7, pp. 133190–133202, 2019.

[29] W. Zhang, X. Zhao, and Z. Li, "A comprehensive study of smartphone-based indoor activity recognition via XGBoost," *IEEE Access*, vol. 7, pp. 80027–80042, 2019.

[30] H. Kim, H.-J. Kim, J. Park, J.-K. Ryu, and S.-C. Kim, "Recognition of fine-grained walking patterns using a smartwatch with deep attentive neural networks," *Sensors*, vol. 21, no. 19, p. 6393, Sep. 2021.

[31] D. Fuller et al., "Predicting lying, sitting, walking and running using apple watch and fitbit data," *BMJ Open Sport Exercise Med.*, vol. 7, no. 1, Apr. 2021, Art. no. e001004.

[32] V. R. Chifu et al., "Identifying and monitoring the daily routine of seniors living at home," *Sensors*, vol. 22, no. 3, p. 992, Jan. 2022.

[33] L. Gan and F. Chen, "Human action recognition using APJ3D and random forests," *J. Softw.*, vol. 8, no. 9, pp. 2238–2245, Sep. 2013.

[34] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3595–3603.

[35] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[36] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017.

[37] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1474–1488, Feb. 2023.

[38] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 10608–10617.

[39] K. Basterretxea, J. Echanobe, and I. del Campo, "A wearable human activity recognition system on a chip," in *Proc. Conf. Design Architectures Signal Image Process.*, Oct. 2014, pp. 1–8.

[40] A. Czabke, S. Marsch, and T. C. Lueth, "Accelerometer based real-time activity analysis on a microcontroller," in *Proc. 5th Int. Conf. Pervasive Comput. Technol. Healthcare (PervasiveHealth) Workshops*, May 2011, pp. 40–46.

[41] R. Chen, H. Luo, F. Zhao, X. Meng, Z. Xie, and Y. Zhu, "A light-weight deep human activity recognition algorithm using multi-knowledge distillation," 2021, *arXiv:2107.07331*.

[42] S. S. Rani, G. A. Naidu, and V. U. Shree, "Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition," *Mater. Today, Proc.*, vol. 37, pp. 3164–3173, 2021.

[43] P. Khaire, J. Imran, and P. Kumar, "Human activity recognition by fusion of RGB, depth, and skeletal data," in *Proc. 2nd Int. Conf. Comput. Vis. Image Process. (CVIP)*, vol. 1. Singapore: Springer, 2017, pp. 409–421.

[44] J. Imran and B. Raman, "Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 1, pp. 189–208, Jan. 2020.

[45] H. K. Mell, S. N. Mumma, B. Hiestand, B. G. Carr, T. Holland, and J. Stopyra, "Emergency medical services response times in rural, suburban, and urban areas," *JAMA Surg.*, vol. 152, no. 10, pp. 983–984, Oct. 2017.

[46] R. Minerva, G. M. Lee, and N. Crespi, "Digital twin in the IoT context: A survey on technical features, scenarios, and architectural models," *Proc. IEEE*, vol. 108, no. 10, pp. 1785–1824, Oct. 2020.