Biophysics Reviews ARTICLE pubs.aip.org/aip/bpr

vioHAR: A visual-inertial odometry-based human activity recognition system

Cite as: Biophysics Rev. **7**, 021408 (2026); doi: 10.1063/5.0304928
Submitted: 8 October 2025 · Accepted: 11 May 2026 ·
Published Online: 25 June 2026

View Online Export Citation CrossMark

Kazi Md. Shahiduzzaman,[1,2] Md Salah Uddin Yusuf,[1,a)] Md Sajjad Hossen,[2] and Md Nazmul Alam Munna[2]

AFFILIATIONS

[1]Department of Electrical and Electronics Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh
[2]Department of Electrical and Electronic Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh

[a)]Author to whom correspondence should be addressed: suyusuf@eee.kuet.ac.bd

ABSTRACT

Human activity recognition (HAR) remains a pivotal research domain with extensive applications in healthcare, elderly care, and assistive technology. However, traditional systems relying solely on inertial measurement units (IMUs) face significant limitations, particularly regarding precision and vulnerability to varying device orientations. To address this urgent need, this study introduces visual inertial odometry based human activity recognition (vioHAR), a novel methodology and public dataset that seamlessly fuses visual and inertial data. By leveraging the visual-inertial odometry system for monocular cameras (VINS-Mono) state estimator, the proposed system generates precise, orientation-independent trajectories, facilitating robust classification across eight distinct daily behaviors and fall scenarios. Extensive evaluation using leave-one-subject-out (LOSO) cross-validation demonstrates that the vioHAR method achieves a robust average accuracy of 73.96%. This represents a substantial improvement over the standard IMU-only baseline (43.35%), yielding specific accuracy gains of +27.60% for the random forest (RF) model and +41.11% for the long short-term memory (LSTM) model. Furthermore, comparing computational efficiency, the LSTM network trained nearly ten times faster than the RF model (44.79 vs 415.81 s) while maintaining a minimal false alarm rate of 0.038, highlighting its superior potential for edge computing applications. Despite hardware challenges such as asynchronous sensor clocks, vioHAR establishes a rigorous and efficient foundation for real-world HAR deployment on emerging head-worn devices.

I. INTRODUCTION

Recognition of different kinds of human activity has gained significant prominence in research because of its significant contributions to human-centered fields that aim to improve people's quality of life. These include healthcare, sports, elderly care, investigation activities, automated public place surveillance for crime prevention, and assistive technologies. In the medical sector, precise human movement detection can significantly advance the development of autonomous diagnostic systems. Additionally, the detection of activities such as running and walking can contribute to advances in video surveillance, smart home technologies, and related applications.[1] Recognition of human activity lies at the heart of numerous automation projects that make our lives easier.[2] The integration of motion sensor-based lighting, which activates upon positively detecting a user's presence, represents just one application of advanced technology. This not only enhances convenience but also contributes to improved safety and significant time savings in our daily lives. This common practice involves collecting signals from both ambient and wearable sensors and subsequently processing them through machine-learning algorithms to classify various human activities.

The prevalent use of smartphones, smartwatches, fitness trackers, and *ad hoc* wearable devices exemplifies the growing trend in monitoring human activities. Smartphones, equipped with a range of motion sensors like accelerometers and gyroscopes, have become widely utilized inertial sensors capable of discerning various physical conditions in humans. Among these, accelerometers are the most commonly used sensors for interpreting body motion signals.[3] However, a noticeable drawback lies in the limitations associated with the number and location of body sensors, making the system highly obtrusive. Gyroscopes have also been employed for HAR and have been demonstrated to improve recognition performance when used in combination with accelerometers.[4] The diversity of sensors supported by smartphones positions them as an alternative to wearable sensing technologies. The combination of these sensors with the processing power and wireless communication capabilities of smartphones makes them robust tools for activity recognition. Moreover, smartphones offer advantages over other ambient sensing methods, such as multi-modal sensors in home environments or surveillance cameras, as they are ubiquitous and require minimal or no static infrastructure for operation.[2]

The effectiveness of the algorithms created for HAR is greatly affected by the accessibility of training (labeled) data. Making such data publicly accessible would enable researchers to objectively compare different techniques.[5–9] Despite the advancements in HAR, existing studies often rely on data collected in controlled environments, with devices positioned in specific ways on different body parts. Consequently, the efficacy of those models may be influenced by the data collected in a controlled environment, which includes specific device orientations and a limited number of activities. In the context of these orientations, the current situation is far from ideal, given that individuals may use these devices, particularly smartphones, in various ways. Even for the same person, different clothing choices can lead to variations in the orientation and placement of the device. In the same way, for different individuals, their body shape, as well as their behavior, can make an enormous difference too.[10] In our previous work,[11] the uActivity framework introduces a user-centric human activity recognition (uHAR) model that fuses advanced sensor technology with machine learning algorithms to analyze and categorize physical movements. By integrating multiple sensor data streams, the uActivity algorithm ensures high reliability in daily activity classification and has demonstrated an exceptional accuracy rate of over 99% in detecting falls.[11] This architecture provides a robust solution for improving the accuracy and dependability of monitoring systems specifically tailored for elderly care. However, trajectory datasets relying solely on IMU sensors lack robust mathematical modeling capabilities.

Although numerous HAR datasets exist, our work introduces a novel approach to this field. Our work introduces a unique dataset that combines acceleration and gyroscope data from an Inertial Measurement Unit (IMU) along with video visuals, aiming to improve the accuracy and precision of HAR and fall detection. The calibration of the IMU emerges as a crucial factor influencing system accuracy and precision. Notably, the combination of IMU and video data resulted in improved accuracy during testing compared to using IMU alone. This improved accuracy in activity recognition significantly benefits the research community.[12] In our approach, a calibrated IMU captures acceleration and gyroscope data at a frequency of 50 Hz. Utilizing an Android device equipped with inbuilt sensors, we seamlessly collected video and IMU data. An essential goal is to overcome biases introduced by specific device orientations in controlled environments, fostering a more comprehensive dataset independent of such orientation limitations. The overarching ambition is to generate accurate trajectories, enabling the effective classification of activities using machine learning algorithms. Subsequent processing resulted in a final dataset comprising images and IMU measurements, forming the foundation of our system. As technology continues to advance rapidly, tools like the Apple Vision Pro may soon become widely accessible to the public. The dataset we have compiled will serve as a valuable resource for researchers looking to explore and work on such emerging technologies. Although privacy concerns pose a drawback to vision-based technologies—often restricting their application in various locations,[13] we have mitigated these issues during data collection by opting for offline approaches. Our methodology involves accurate positioning and activity assessment, employing a smartphone mounted on a helmet to appropriately capture video data. The gathered data are subsequently processed and classified offline utilizing diverse machine learning methods. The model was evaluated across eight distinct activities, yielding a noteworthy accuracy level. This work provides a dataset that is more realistic and is not influenced by the orientation of the device, positioning, and user variability to solve the identified problems. These differentiations distinguish our work from others in the domain. Additionally, we demonstrate the feasibility of our approach by implementing the uActivity[11] algorithm on the collected data. This model, trained on eight different activities, achieved high accuracy, proving the effectiveness of our approach for real-world HAR applications.

A. Main contributions of this work

In summary, the primary contributions of this research are as follows:

- The introduction of vioHAR, a novel public dataset and methodology that fuses visual and inertial data to enhance recognition precision while specifically addressing the orientation-dependency limitations of traditional IMU-only systems.
- The implementation of the VINS-Mono state estimator to generate accurate, three-dimensional trajectories for eight distinct daily behaviors and fall scenarios, providing a robust foundation for activity classification regardless of device placement.
- A rigorous cross-subject validation (LOSO) demonstrating that the proposed VIO-based approach achieves a 73.96% accuracy, effectively doubling the 43.35% baseline accuracy achieved by standard accelerometer and gyroscope data alone.
- Evidence of computational efficiency for real-world deployment, showing that LSTM networks can train approximately ten times faster than traditional machine learning models in LOSO validation, facilitating deployment on resource-constrained head-worn edge devices.

II. LITERATURE REVIEW

Machine learning algorithms have been successfully employed in recognizing various activities in healthcare and other areas as demonstrated by the fall detection cum emergency response system presented by Tao *et al.*[14] These approaches, focused on improving fall detection accuracy and reliability, highlight the transformative potential of machine learning in healthcare. Guerra *et al.* have developed a methodology applied in human activity recognition (HAR) for ambient assisted living, covering basic threshold-based techniques to advanced deep learning algorithms[15] incorporating both wearable and non-wearable sensors for data collection and classification.

In the realm of computer vision, video surveillance commonly relies on support vector machines and piece-wise linear support vector machine (PL-SVM).[16] While effective for interpreting two-dimensional views, this approach may lack precision in identifying activities, posing a challenge in scenarios requiring more precise activity identification. Smartphone sensors provide omnipresent and easily accessible data for detecting human activity. Nonetheless, Chen and Shen's performance analysis underscored the critical importance of technically processing raw data to achieve optimal output.[17] Yacchirema *et al.* introduced the IoTE-Fall system, utilizing IoT and ensemble machine learning algorithms to detect falls among elderly individuals in indoor environments.[18] Huynh *et al.* developed a wearable fall detection system that relies on an accelerometer and gyroscope for activity recognition and fall detection.[19] Additionally, off-the-shelf machine learning toolkits like WEKA have been proposed as practical solutions for addressing classification or detection-related issues in data mining, as demonstrated by Ratra *et al.*[20]

Recognition techniques for daily living activities and fall detection typically perform their tasks by analyzing samples from sensors. These sensors can be physically deployed in the ambient surroundings and include ambient sensors such as cameras, vibration sensors, and microphones or worn by people (wearable sensors, e.g., accelerometers, gyroscopes, and barometers).[21] In order to train and evaluate HAR researcher make their own dataset or use datasets publicly available.[3,22–24] The diverse methodologies employed to enhance the accuracy and reliability of fall detection systems underscore the potential of machine learning to revolutionize healthcare. Various machine-learning methods and techniques have been proposed to enhance the accuracy and reliability of fall detection systems. The integration of machine learning algorithms in fall detection systems has shown promising results in improving accuracy and reducing false positives.

Cameras (categorized as ambient sensors), and inertial measurement sensors (utilized as wearable devices), are the primary types of sensors employed to capture publicly accessible datasets containing kinematic data of human subjects. HumanEva,[25] CMU Mocap,[26] and datasets mentioned in the survey by Chaquet *et al.*[27] are instances of datasets that comprise samples gathered from cameras, while the works by Wang *et al.*[28] and Urtasum *et al.*[29] are some examples of techniques for obtaining dynamic models of human motion from camera data.[3] Janidarmian *et al.* amalgamated 14 publicly accessible datasets concentrating on acceleration patterns to scrutinize feature representations and classification techniques for human activity recognition.[30] However, they did not make the resulting dataset publicly available for download. Publicly available datasets can be primarily divided into three main sets: those acquired by ambient sensors, those acquired by wearable devices, and datasets that are a combination of the two. Recently, there has been considerable emphasis on wearable sensors due to their non-intrusive nature, outdoor functionality, and typically lower cost compared to ambient sensors. This trend is supported by the growing number of techniques relying on wearable sensors, as highlighted in surveys such as the one conducted by Luque *et al.* on fall detection techniques using smartphone data.[31]

Exploring publicly available datasets related to human body movement mainly with a focus on human activity recognition and others with a pure biomechanical focus. Notably, datasets in the first category often lack very fine movement tracking, while those in the second category utilize high-end laboratory equipment, making it impractical to collect data in natural environments. Consequently, there exists a demand for datasets that cater to telemedicine settings, aiming to recognize patients' activities and comprehensively track their movements in their natural environment.[32–35] Table I shows few publicly available data using the smartphone and our hypothesis on the proposed vioHAR dataset.

III. DATASET DEVELOPMENT

This section marks the transition from the theoretical and methodological design of our system, as outlined in Sec. III, to the practical realization of the vioHAR dataset. Dataset development is the cornerstone of this research, transforming raw visual and inertial sensor streams into a fully calibrated and labeled resource suitable for human activity recognition (HAR) training and evaluation. We detail the entire lifecycle of the dataset, beginning with the rationale for

TABLE I. Publicly available dataset collected using smartphone. ADL: Activities of Daily Living, F: Falls, A: Accelerometer, G: Gyroscope, CO: Compass, M: Magnetometer, CA: Camera LA: Linear Acceleration, AT: Atmospheric Pressure, OR: Orientation, L: Light, S: Sound, GPS: Global Positioning System, SP: Smartphone, SW: Smartwatch, IMUs: Inertial Measurement Units.

Dataset	Year	No. of activity	No. of subject	Sensors	Sampling rate (Hz)	Device
UCIHAR	2012	6 ADL	30	A,G	50	SP
Gravity	2016	7 ADL, 12 F	3	A, G, CO	25	SP, SW
MobiFall	2014	9 ADL, 4F	24	A,G	H	SP
HARUS	2013	6 ADL	30	A,G	50	SP
HHAR	2015	6 ADL	9	A,G	H	SP, SW
Shoaib PA	2013	6 ADL	4	A,G, M	50	SP
Shoaib SA	2014	7 ADL	10	A,G,M,LA	50	SP
CHAD	2016	13 ADL	10	A,G,LA	50	SP
Motion Sense	2018	6 ADL	24	A,G,AT	50	SP
MobiAct	2016	11 ADL, 4 F	67	A,G,OR	87	SP
UniMiB SHAR	2016	8 ADL, 8 F	30	A	50	SP
UMA Fall	2017	12 ADL, 3 F	19	A, G, M	200, 20	SP, IMUs
RealWorld (HAR)	2016	8 ADL	15	A, G, GPS, L, M, S	50	SP, SW
WISDM	2012	6 ADL	29	A	20	SP
vioHAR	**2024**	**6 ADL, 2F**	**14**	**CA, A, G**	**30, 50**	**SP**

selecting the eight specific daily living and fall activities. Following this, we meticulously describe the data collection setup and protocol, emphasizing the head-mounted configuration and the safety measures employed during subject participation. Finally, this chapter presents the critical results of the visual-inertial odometry (VIO) trajectory extraction, which forms the unique and precise foundation of our new dataset.

A. Selection of the activities

The choice of specific activities in our Human Activity Recognition (HAR) study was influenced by a thorough evaluation of everyday human movements, including diverse physical activities. Each selected activity exhibits distinctive motion patterns and has specific recognition issues, necessitating a robust HAR system. Gait patterns differentiate Walking and Running as basic motor activities and present typical daily behaviors. Jumping is a dynamic and noisy activity that enhances the variation of human movement. Sitting and Standing are static postures adopted during long periods of rest or engagement. “Climbing stairs” and “descending stairs” were chosen to represent vertical movement, which includes variations in acceleration and orientation that are relevant for detecting transitions between levels. These activities simulate real-world scenarios where participants navigate multi-level environments, enhancing the complexity of recognition tasks. The integration of both “falling forward” and “falling backward” scenarios is important in improving the safety component for our research. Identifying such actions is very fundamental in developing systems that have the ability to quickly act on the identification of possible accidents or emergencies and improve healthcare-assisted living applications. he actions were selected with care to provide a wide and comprehensive dataset for our system of HAR to generalize well to different human movements. This inclusiveness allows the model to comprehend various activities thoroughly, laying a foundation for applications in healthcare, fitness tracking, and intelligent environments where accurate activity recognition is indispensable.

B. Data collection setup and safety measures

We developed a comprehensive testbed for data acquisition, emphasizing efficiency and safety measures. The method entailed affixing a mobile device to a helmet, which was thereafter provided to the subject for data acquisition.

The setup illustrated in Fig. 1 allowed us to record various activities comfortably. We integrated the smartphone into a wearable helmet equipped with safety gear, ensuring efficient data collection while prioritizing subject comfort and safety. The experimental design sought to equilibrate extensive data acquisition with participant safety to improve the reliability and ethical integrity of the study.

C. Subjects and ethical considerations

Data were collected from a diverse group of healthy adult subjects recruited from various age groups. Before recording the data, each participant received comprehensive instructions, both in written and oral form, regarding the experiment. Brief criteria of participants are listed in Table II. Additionally, they were informed about the creation of a public dataset for HAR, wherein their video and IMU data records would be utilized.

FIG. 1. Data collection setup for our approach.

TABLE II. The characteristics of the subjects.

No. of subjects (male and female)	Age (year)	Height (cm)	Weight (kg)
14	18–28	148–186	46–92

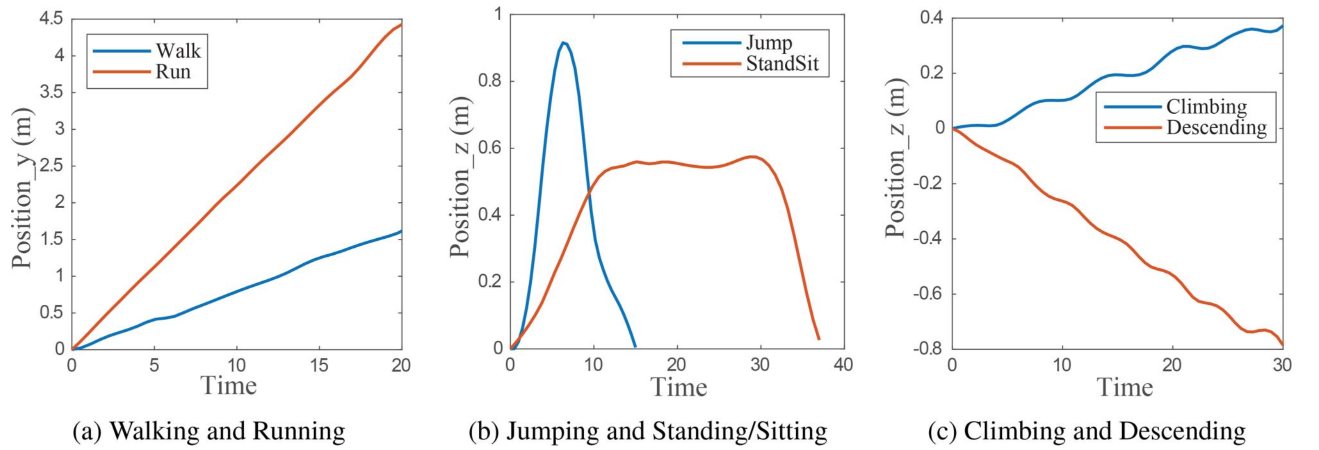
D. Data acquisition protocol

To acquire the raw data from subjects, we used a head-mounted system. We asked each volunteer to wear a helmet as shown in Fig. 1, and we placed our smartphone in the same position as a smart glass, or VR goggles would be used. This also helped to reduce the background noise captured by the IMU sensor and Camera. We have followed the protocols mentioned in Table III.

Using the MARS Logger for every activity session generates a folder and creates four files. For the data annotation process, we recorded the activity in the same sequence; then we renamed every folder with the activity annotation keyword. Finally, for every subject, we created a folder with that subject keyword annotation. We manually start the data recording process by tapping the record button of

TABLE III. The protocols for activity data acquisition.

Protocol	Action	Iteration
Protocol 1	Walking for 20s Running for 10s Climbing Stairs for 15s Descending Stairs for 15s	1 time
Protocol 2	Jumping Sitting and Standing	5 times
Protocol 3	Falling Forward Falling Backward	1 time



(a) Walking and Running (b) Jumping and Standing/Sitting (c) Climbing and Descending

FIG. 2. Positional changes across different dimensions over time.

the MARS Logger. For data processing flexibility, we asked the subject to wait at least 5 s before starting the activity sequence.

For protocols, we tend to follow moderate walking and running conditions as it will give a better result for elderly people. It also helps to acquire a better dataset because of the low noise it records by conducting moderate activities. The other protocols were done in the most natural way possible. Volunteers were asked to sit up and down in a chair, starting from the sit-down position in a natural motion. A mattress was utilized to collect fall activity data for the safety of the volunteers.

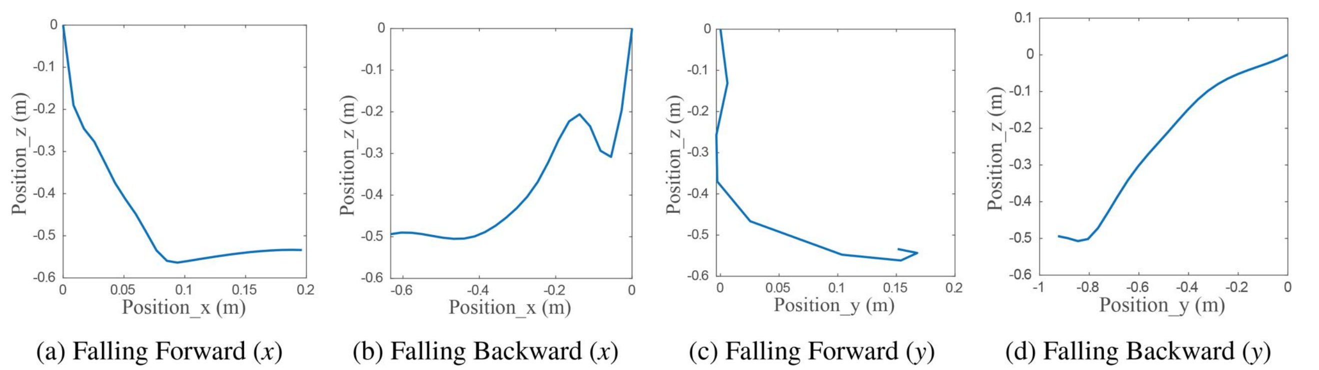
E. Analysis of extracted trajectory data

The retrieved trajectory data has several parameters: Position_x, Position_y, Position_z, Orientation_x, Orientation_y, Orientation_z, and Orientation_w. Each column encapsulates a distinct aspect of position and orientation data, enhancing the overall depiction of motion. To improve the dataset's clarity, each created trajectory is carefully annotated for various activities.

The trajectories represent eight distinct activities: Walking, Running, Jumping, Sitting, Standing on a chair, Climbing Stairs, Descending Stairs, Falling Forward, and Falling Backwards. Fourteen volunteers have engaged in the activities, each distinctly recognized by a Subject_ID. The activities have been designated with the Activity_ID.

This trajectory dataset is formatted as CSV files and Matlab tables for enhanced usage in research and practical applications. Its structured organization enables efficient data management and analysis, rendering it adaptable for diverse research fields. This extensive trajectory dataset, equipped with appropriate labels and an organized structure, serves as a valuable resource for study in motion analysis, activity recognition, and related fields.

We visualized the positional changes across different dimensions over time for activities such as "Walking," "Running," "Jumping," "Sitting and Standing," "Climbing Stairs," and "Descending Stairs" (designated as A01–A08 respectively) in Fig. 2. Additionally, Fig. 3 illustrates the positional changes across diverse dimensions for activities involving Falling Forward and Falling Backward. Through the trajectory plotting, we have identified many curve properties. Upon meticulous examination of the graphs, unique trends and patterns emerge for each activity area. The peaks and troughs in the curves correspond to particular points during the activities, reflecting the variations in vertical position. Analyzing these trajectories provides a qualitative understanding of how different activities manifest in their



(a) Falling Forward (x) (b) Falling Backward (x) (c) Falling Forward (y) (d) Falling Backward (y)

FIG. 3. Navigating positional changes across diverse dimensions.

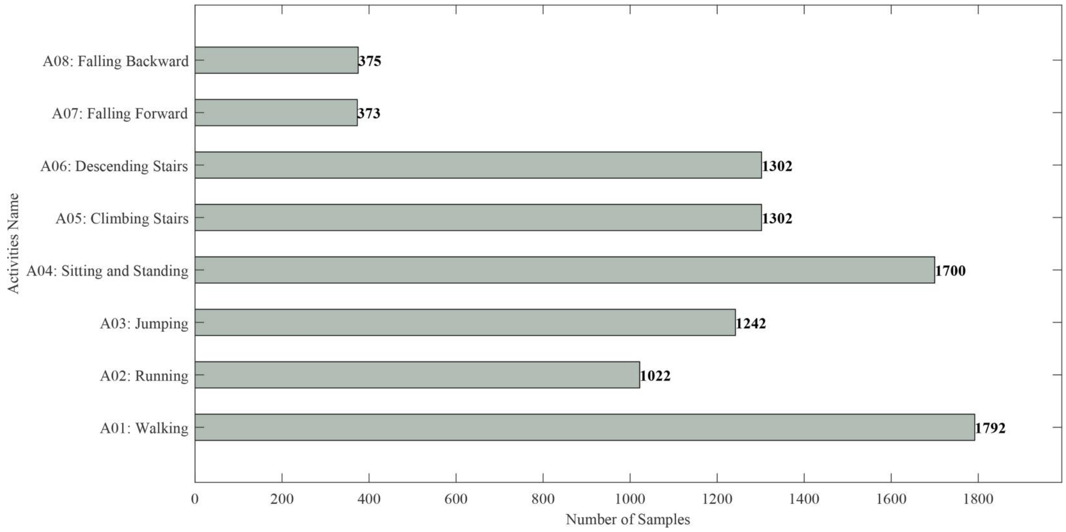


FIG. 4. Activity-wise sample counts for machine learning.

movement. Our method using Visual-Inertial Odometry (VIO) shows that this visual representation helps recognize and understand the features of other activities.

IV. DATASET EVALUATION

The assessment of datasets constitutes a significant aspect of our work. We graphically illustrate the general distribution of data samples in Fig. 4. Notably, the sample sizes for activities A07 and A08 are smaller relative to the other activities. The diminished samples of A07 and A08 can be ascribed to the brief length of the forward and backward falling actions. The volunteers' descent transpires within a brief interval, lasting under 2 s. This elucidates the reason for the reduced number of samples for A07 and A08 in our dataset compared to other activities.

A. Performance evaluation metrices

We assessed our dataset using traditional machine learning methods and deep learning models, noting exceptional results. To ensure a rigorous evaluation, we employed two distinct validation strategies depending on the model's architecture and purpose: a 50:50 train-test random split and a leave-one-subject-out (LOSO) cross-validation. User-specific and activity-specific models: For models trained to recognize activities for a single user, or to predict a user's identity based on a specific activity, we utilized datasets specific to those individual constraints. These models were evaluated strictly using a 50:50 train-test split. This method guarantees the model is trained on a portion of the dataset and evaluated on its capacity to generalize to unseen data from that same specific context. Global models: For the global model, we consolidated the data into a single dataset containing all activities and users. We evaluated these global models using two methods: first, a standard 50:50 train-test split to establish baseline performance metrics against the user-specific models, and second, a rigorous leave-one-subject-out (LOSO) cross-validation to assess the model's ability to generalize to completely unseen subjects. Our process of dataset evaluation is illustrated in Fig. 5.

For the user-specific evaluation, we utilized datasets specific to individual users. We conducted training and testing using the 50:50 train-test split, concentrating on activity recognition for a single user. Similarly, we selected a single activity to predict user identity for the activity-specific user recognition model, again applying the 50:50 train-test split for evaluation. We report the results using the arithmetic mean accuracy for both the user-specific and activity-specific datasets. All outcomes and comparative analyses are depicted in Figs. 6 and 7. The two figures offer a clear visual depiction of each model's performance metrics, emphasizing significant accuracy discrepancies between the user-specific and activity-specific methodologies.

We have measured the following performance evaluating metrics:

- Accuracy: This is a basic measurement that tells us the percentage of times the system correctly identified an activity compared to the total number of attempts. It is calculated as the number of correct classifications divided by the total number of instances.
- Precision: This metric measures how precise the positive predictions are. It is calculated as the ratio of true positives (correctly predicted activities) to the sum of true positives and false positives (activities incorrectly predicted as that specific action).
- Recall: This measures the ability of the system to find all the relevant cases within a dataset. It is calculated as the number of true positives divided by the sum of true positives and false negatives (instances where the system failed to identify the activity).
- F1-Score: This is a single score that combines both precision and recall to give a better overall picture of the system's performance. It is useful because it balances the two, especially when one metric might be high and the other low.
- False Alarm Rate (FAR): Also known as the false positive rate, is a metric that analyzes the tendency of the model to make an incorrect prediction for the positive class. To better understand

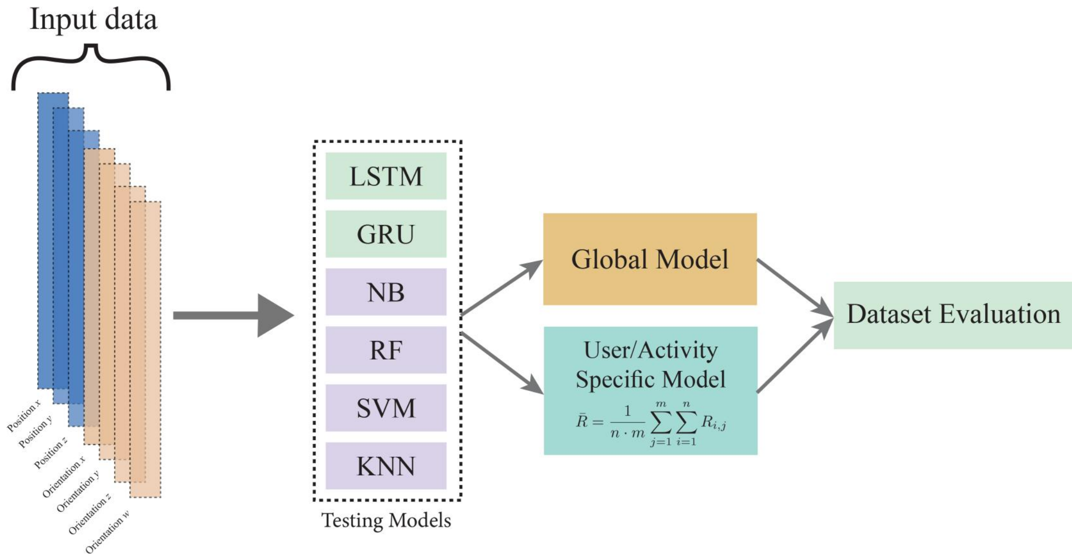


FIG. 5. Evaluation process of the vioHAR dataset.

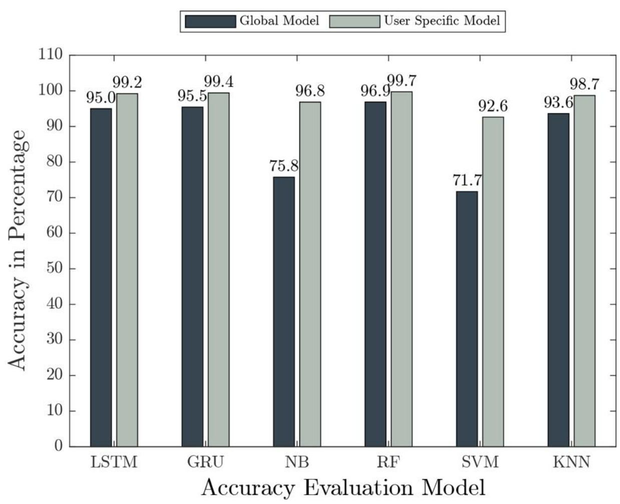


FIG. 6. Comparison of average accuracy for activity recognition between global and user-specific models across different ML and DL algorithms.

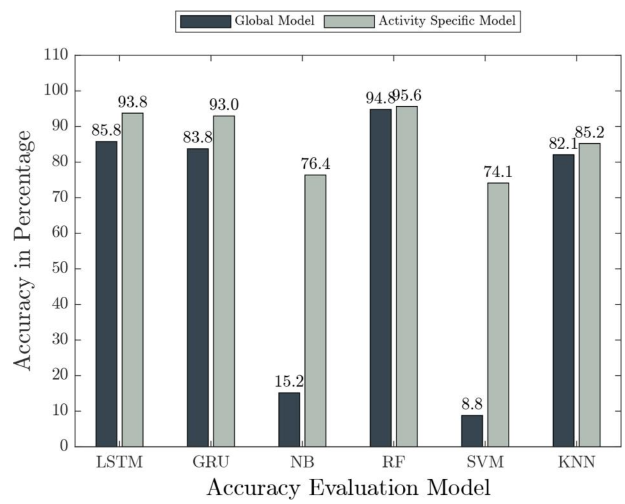


FIG. 7. Comparison of average accuracy for subject recognition between global and user-specific models across different ML and DL algorithms.

the model's specificity, the false positive rate is calculated as the ratio of false positives to the sum of false positives and true negatives.

- Overall Balanced Accuracy: This metric is the average of sensitivity (recall) and specificity, which helps when some activities happen much less often than others. While not explicitly defined in the text, it is calculated as the arithmetic mean of sensitivity and specificity to handle imbalanced data where some classes have fewer samples (like falls) compared to others (like walking).
- Macro F1 Score: This is calculated by finding the F1-score for each individual activity first and then taking the average of all those scores. This ensures that the performance on less frequent activities (like falling) counts just as much as the performance on frequent ones (like walking), providing a fair view of how the system works across all different types of actions.

The aforementioned metrics collectively furnish a thorough quantitative evaluation of a model's performance across multiple dimensions. They yield nuanced perspectives on the model's predictive accuracy, its capacity to reconcile precision and recall, as well as its efficacy in minimizing false positives. In real-world applications, these assessments inform the iterative refinement of the model, thereby guaranteeing its reliability and accuracy in forecasting outcomes on previously unseen data. Application of such performance evaluation metrics aligns with the purpose of the enhancement of the

TABLE IV. Activity-specific performance evaluation matrices for different users using the RF algorithm.

ID	Activities	Accuracy	Recall	Precision	F1 Score	FAR	Overall balanced accuracy	Micro F1-Score
A01	Walking	0.9899	0.9898	0.9903	0.9900	0.000 77	0.9982	0.9967
A02	Running	0.9804	0.9810	0.9822	0.9816	0.0015	0.9864	0.9742
A03	Jumping	0.9484	0.9515	0.9501	0.9508	0.0041	0.9652	0.9332
A04	Sitting and Standing	0.9070	0.9098	0.9118	0.9108	0.0077	0.9556	0.9159
A05	Climbing Stairs	0.9815	0.9832	0.9812	0.9822	0.0014	0.9899	0.9812
A06	Descending Stairs	0.9769	0.9777	0.9796	0.9787	0.0018	0.9866	0.9750
A07	Falling Forward	0.9572	0.9281	0.9565	0.9421	0.0034	0.9330	0.8689
A08	Falling Backward	0.9095	0.8990	0.9134	0.9061	0.0074	0.9186	0.8428
Average		0.9564	0.9525	0.9581	0.9553	0.0035	0.9667	0.9360

accuracy, reliability, and pertinence of machine learning models for a whole bunch of real-world scenarios.

B. Assessment of vioHAR using multiple machine learning algorithms

We used two deep learning algorithms, long short-term memory (LSTM) and gated recurrent units (GRUs). We also used four different machine learning algorithms to analyze our dataset: naive bayes (NB), random forest (RF), support vector machine (SVM), and k-nearest neighbors (k-NN).

We used each algorithm to evaluate its effectiveness in classifying human activities and users. Traditional machine learning techniques have been crucial in addressing the complexities inherent in human activity recognition (HAR) tasks. We found that the user-specific model works better in all situations, as shown in Figs. 6 and 7 for the activity recognition and user recognition using different machine learning and deep learning algorithms. Figures 6 and 7 also illustrates that the Random Forest algorithm outperformed all other algorithms when using the traditional machine learning algorithm. Furthermore, it demonstrates that deep learning algorithms, specifically long short-term memory (LSTM) and gated recurrent units (GRUs), performed similarly to the random forest algorithm in terms of accuracy and predictive power.

It shows recognizing a user from both the ADL and falls is more challenging than user-specific activity recognition because the variability in individual behaviors and contexts can lead to significant differences in the recognition patterns. Furthermore, integrating data from both activities enhances the model's accuracy but requires sophisticated algorithms to effectively manage the increased complexity.

As the RF (random forest) exhibits best accuracy, therefore it is worthy to go deeper analysis with Tables IV and V. It can be seen that FAR is very less close to 0.04% in both activity recognition and subject detection. The comparison between Accuracy and Overall Balanced Accuracy in Table IV is crucial because the dataset is not perfectly distributed across all activities. Since there are fewer samples for falling actions compared to walking or running, a simple Accuracy metric

TABLE V. User-specific performance evaluation matrices for different activities using the RF algorithm.

ID	Accuracy	Recall	Precision	F1 Score	FAR	Overall balanced accuracy	Micro F1-score
S01	0.9876	0.9643	0.990 2	0.9771	0.0018	0.9812	0.9744
S02	1.0000	1.0000	1.000 0	1.0000	0	1.0000	1.0000
S03	0.9967	0.9976	0.997 2	0.9974	0.0004	0.9965	0.9946
S04	1.0000	1.0000	1.000 0	1.0000	0	1.0000	1.0000
S05	1.0000	1.0000	1.000 0	1.0000	0	1.0000	1.0000
S06	0.9968	0.9896	0.996 2	0.9929	0.0004	0.9946	0.9927
S07	0.9755	0.9252	0.980 1	0.9519	0.0036	0.9608	0.9461
S08	0.9819	0.9510	0.988 9	0.9696	0.0028	0.9689	0.9556
S09	0.9773	0.9357	0.981 1	0.9577	0.0035	0.9661	0.9531
S10	0.9969	0.9972	0.998 0	0.9976	0.0005	0.9984	0.9976
S11	0.9843	0.9725	0.990 23	0.9813	0.0024	0.9850	0.9804
S12	1.0000	1.0000	1.000 0	1.0000	0	1.0000	1.0000
S13	0.9970	0.9911	0.997 8	0.9944	0.0005	0.9953	0.9943
S14	0.9939	0.9799	0.996 2	0.9879	0.0009	0.9963	0.9936
Average	0.9916	0.9789	0.993 2	0.9859	0.0012	0.9888	0.9844

could be misleading if the model ignored those rare events. The standard Accuracy average of 0.9564 is slightly lower than the Overall Balanced Accuracy of 0.9667, which is a positive sign. It suggests that the system maintains high specificity and sensitivity even for the minority classes, ensuring that falls are not missed despite their lower frequency. In the user-specific results of Table V, these two values are almost identical, with an Accuracy of 0.9916 and a Balanced Accuracy of 0.9888. This convergence indicates that the personalized models are robust enough that class imbalance does not significantly skew the performance metrics.

The distinction between the F1-Score and the Micro F1-Score highlights how the model performs on a per-class basis vs the dataset as a whole. In Table IV, the average F1-Score is 0.9553, which is the simple average of the scores for all eight activities. This high value shows that the model is generally effective at identifying each distinct activity type. However, the Micro F1-Score is lower at 0.9360. This metric aggregates all the predictions globally, meaning that misclassifications in any category affect the total score directly based on their frequency. The drop in the Micro F1-Score suggests that while the per-class performance is strong, there is a cumulative error rate across the entire dataset that slightly reduces global precision and recall. In contrast, Table V shows a Micro F1-Score of 0.9844 that is nearly equal to the average F1-Score of 0.9859, confirming that personalization minimizes these cumulative errors.

The average confusion matrix of the activity recognition and subject identification supports the claims in the Tables VI–IX. We can observe that the user centric model outperforms the global model.

Based on the comprehensive performance analysis illustrated in the bar charts (Figs. 8–13) and tables (Tables IV and V), random forest (RF) and long short-term memory (LSTM) networks emerged as the standout performers, with user-specific models consistently outperforming their global counterparts. Random forest demonstrated exceptional robustness, achieving an average overall balanced accuracy of 98.88% in the user-specific activity analysis. This aligns with the bar charts, where RF reached a user-specific activity overall balance of approximately 97.3% and a Subject Macro F1-Score of 99.1%. LSTM also proved highly effective, securing a user-specific activity overall balance of 96.9% (Figs. 14–17). However, a direct comparison in the Global models reveals that Random Forest offers superior stability; for instance, RF achieved a Global Activity Micro F1-Score of 94.0%, notably higher than the 88.9% observed for LSTM. These data indicates that while both algorithms benefit significantly from personalization, Random Forest retains a distinct advantage in generalizing across subjects in the Global model settings. This superiority is detailed further in Tables IV and V. Specifically, Table IV shows that the random forest algorithm maintains an impressive average overall balanced accuracy of 96.67% across all activities, proving its effectiveness even on imbalanced minority classes like falling backward, which still achieved a robust 91.86%. Table V confirms that when trained on individual user data, the model achieves near-perfect generalization for several subjects, with an average Micro F1-Score of 0.9844, significantly minimizing the false alarm rates compared to the global approach. This balance is critical because a high F1-Score cannot exist if either precision or recall is compromised. Furthermore, the "Overall Balanced Accuracy" charts reinforce this finding by accounting for class imbalances, implicitly confirming that the false alarm rate (fAR) is kept to a minimum (Figs. 18 and 19); a high balanced accuracy requires the model to correctly identify nonevents just as well as events. The visual gap between the global (dark bars) and user-specific (light bars) data further highlights that personalizing the model significantly tightens these metrics, effectively reducing false positives and enhancing the system's reliability in real-world scenarios.

Establishing a standard IMU-only baseline is necessary to scientifically isolate and quantify the specific performance gains contributed by our visual-inertial odometry (VIO) method. Since most existing human activity recognition systems rely solely on accelerometer and gyroscope data, we must demonstrate that the additional complexity of processing visual trajectory data yields a statistically significant improvement in accuracy. We selected random forest (RF) and long short-term memory (LSTM) as the representative algorithms for this baseline comparison because they serve as the "gold standards" for their respective categories. Random forest was chosen as it demonstrated the highest accuracy among traditional machine learning classifiers in our initial tests, while LSTM was selected for its proven ability to handle temporal dependencies in sequential deep learning tasks. By pitting our vioHAR approach against these two rigorous baselines, we can conclusively prove whether the trajectory data provides a tangible benefit over conventional sensing methods (see Table X).

While the 50:50 random split provided strong baseline metrics for both user-specific and global models, random splitting within a

TABLE VI. Average confusion matrix for activity recognition using the user-specific model with the random forest algorithm.

	|	A01	A02	A03	A04	A05	A06	A07	A08
Actual Activities	A01	869	0	0	0	0	0	0	0
	A02	0	506	0	0	0	0	0	0
	A03	0	0	649	6	0	0	0	0
	A04	0	0	2	859	0	0	0	0
	A05	0	0	0	0	672	0	0	0
	A06	0	0	0	0	0	620	0	0
	A07	0	0	3	0	0	1	169	0
	A08	0	0	0	1	0	0	0	199
Predicted Activities									

TABLE VII. Confusion matrix for activity recognition using global model with the random forest algorithm.

	|	A01	A02	A03	A04	A05	A06	A07	A08
Actual Activities	A01	881	6	5	8	0	2	0	0
	A02	5	481	1	0	0	0	0	0
	A03	0	0	551	62	0	0	0	0
	A04	0	0	11	823	0	0	0	0
	A05	0	0	0	4	650	4	0	0
	A06	0	0	0	0	11	648	0	0
	A07	0	0	3	1	0	3	194	0
	A08	0	0	5	10	0	2	0	183
Predicted Activities									

TABLE VIII. Average confusion matrix for user recognition using the activity-specific model with the random forest algorithm.

	|	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14
Actual Subjects	S01	319	0	0	0	0	0	0	2	3	1	0	3	0	6
	S02	0	327	0	0	0	3	1	2	0	2	0	0	2	0
	S03	0	2	293	0	4	0	1	0	1	0	0	1	0	1
	S04	0	0	0	342	0	0	0	7	0	0	0	0	0	2
	S05	0	0	4	0	344	0	0	0	0	0	1	0	0	1
	S06	0	4	0	0	0	272	2	0	1	0	0	1	1	3
	S07	4	6	0	0	0	0	312	0	1	1	0	1	0	0
	S08	2	3	0	2	0	0	1	302	0	9	2	2	1	1
	S09	10	1	2	0	2	0	2	0	283	4	2	1	1	0
	S10	1	0	0	0	0	1	4	3	0	323	0	1	3	0
	S11	6	0	0	0	0	0	0	0	1	2	301	0	1	1
	S12	1	1	0	0	0	0	4	0	0	3	0	330	0	1
	S13	0	0	0	0	0	2	0	4	0	4	0	0	329	3
	S14	2	0	0	0	0	2	1	2	0	0	0	0	2	286
Predicted Subjects															

TABLE IX. Average Confusion matrix for user recognition using global model with the Random Forest (RF) algorithm.

	|	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14
Actual Subjects	S01	308	0	0	0	0	0	5	2	6	2	3	3	0	9
	S02	0	316	2	2	1	0	0	0	0	6	0	1	6	1
	S03	1	0	305	0	3	1	0	0	2	0	0	10	0	1
	S04	0	0	0	342	1	1	1	1	0	0	0	0	0	1
	S05	0	1	8	0	328	3	0	3	0	0	0	0	0	3
	S06	0	0	0	0	1	306	1	0	3	1	0	1	5	2
	S07	0	2	0	1	0	0	303	2	0	3	0	5	1	4
	S08	1	1	0	9	0	0	1	295	2	1	0	2	1	2
	S09	5	1	7	1	0	0	2	4	276	4	0	1	2	0
	S10	1	4	0	4	0	0	1	8	0	311	0	2	0	0
	S11	1	0	2	1	0	0	0	1	0	0	303	2	4	2
	S12	2	0	2	0	0	0	1	0	1	2	0	327	0	0
	S13	1	0	0	1	0	1	1	1	0	4	0	1	296	0
	S14	1	0	0	0	3	2	0	3	7	1	0	0	0	301
Predicted Subjects															

global dataset can allow specific user traits to leak into the training process. To rigorously evaluate the true generalization capability of the global model, we implemented the second phase of our validation: a leave-one-subject-out (LOSO) cross-validation protocol. This method involves training the classifier on data from all subjects except one, which is reserved exclusively for testing, repeating the process so that each of the 14 participants serves as the test subject exactly once.

The integration of VIO significantly enhances the robustness and generalization of human activity recognition (HAR) systems compared to standard IMU-only baselines, particularly in real-world scenarios involving unseen users. According to Table XI, the proposed vioHAR method more than doubles the performance of the IMU baseline during leave-one-subject-out (LOSO) cross-validation, elevating accuracy from approximately 0.43 to over 0.73 for both LSTM and random forest (RF) models. This drastic improvement is further supported by a reduction in the false alarm rate (FAR) from over 8% to under 4%, indicating that fused trajectory data provides a much clearer signal for activity differentiation than raw accelerometer and gyroscope data alone.

While Table XII illustrates a performance gap between random 50:50 splits (approximately 95.6% accuracy) and LOSO testing (approximately 73.9% accuracy), the LOSO results represent a more "honest" assessment of how the system performs on completely new subjects. Table XIII and XIV reveal that while RF and LSTM perform similarly in terms of overall balanced accuracy (approximately 85%), they offer distinct computational tradeoffs. LSTM demonstrates a

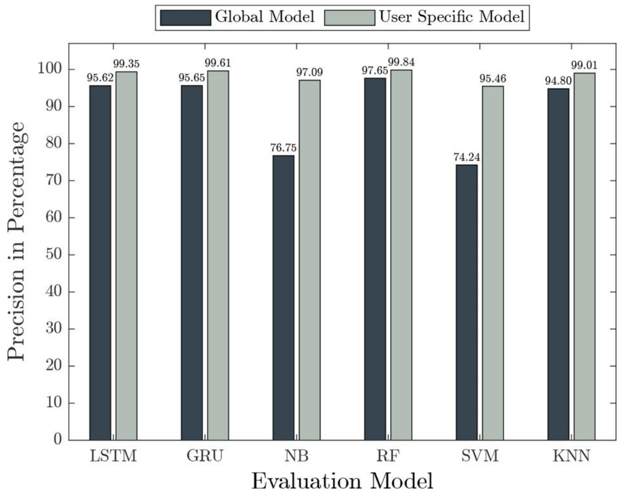


FIG. 8. Comparison of average precision for activity recognition between global and user-specific models across different ML and DL algorithms.

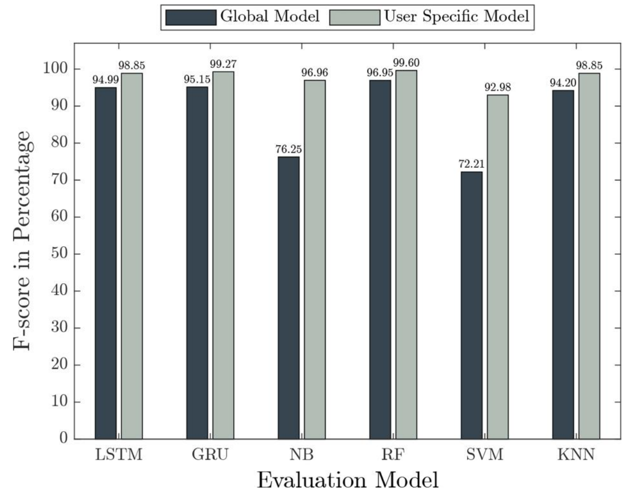


FIG. 10. Comparison of average F-score for activity recognition between global and user-specific models across different ML and DL algorithms.

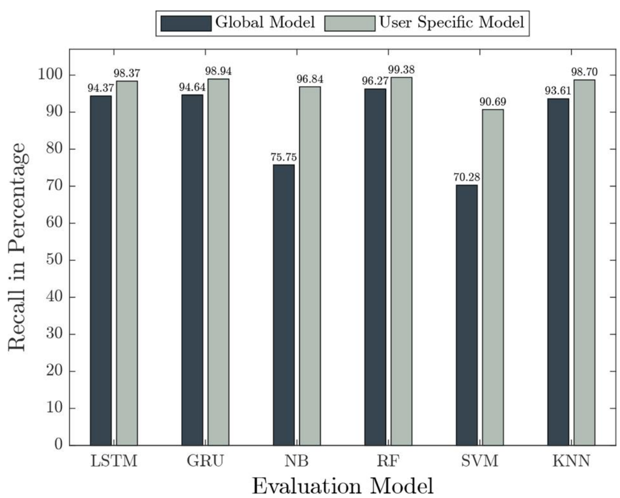


FIG. 9. Comparison of average recall for activity recognition between global and user-specific models across different ML and DL algorithms.

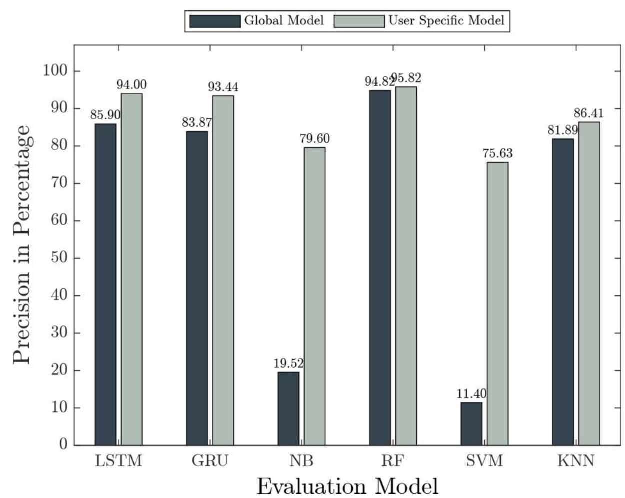


FIG. 11. Comparison of average precision for subject recognition between global and user-specific models across different ML and DL algorithms.

significant advantage in efficiency, requiring only 44.8 s for training compared to the 415.8 s required by random forest, making it a more viable candidate for resource-constrained applications despite the two models sharing nearly identical predictive power.

One of the ways to determine computational complexity is through training time. It is easily observed from Table XV that user-specific models require less time than global models. This can be interpreted as the CPU requires less computational resources for user-specific model development. We can also assume that the power consumption will be less, leads to better battery performance in terms of discharge. However, the computational resources required for deep learning models were significantly higher than the classical machine learning techniques which can be a critical consideration in practical applications.

C. Performance comparison: IMU baseline vs proposed VIO method

To scientifically isolate the performance gains contributed by our methodology, we established a standard IMU-only baseline using raw accelerometer and gyroscope data. Both the baseline and the proposed VIO method utilized identical preprocessing steps, including a 2.56 s window size (128 samples), 50% overlap, and Min-Max normalization to ensure a fair comparison.

1. Quantitative accuracy gains

The integration of fused trajectory data yielded statistically significant improvements across all tested models:

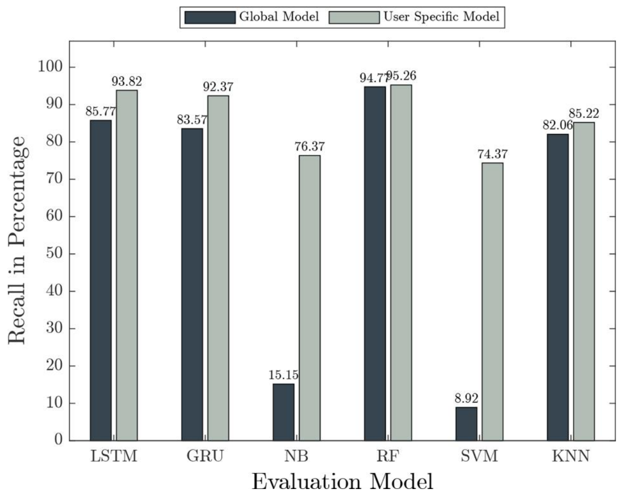


FIG. 12. Comparison of average recall for subject recognition between global and user-specific models across different ML and DL algorithms.

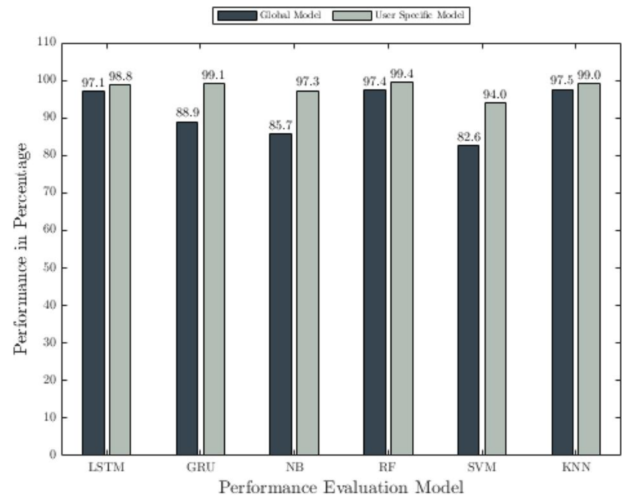


FIG. 14. Comparison of average overall balance for activity recognition between global and user-specific models across different ML and DL algorithms.

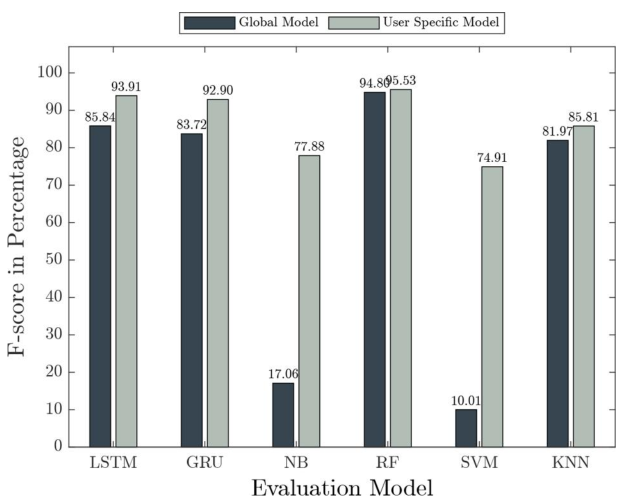


FIG. 13. Comparison of average F-score for subject recognition between global and user-specific models across different ML and DL algorithms.

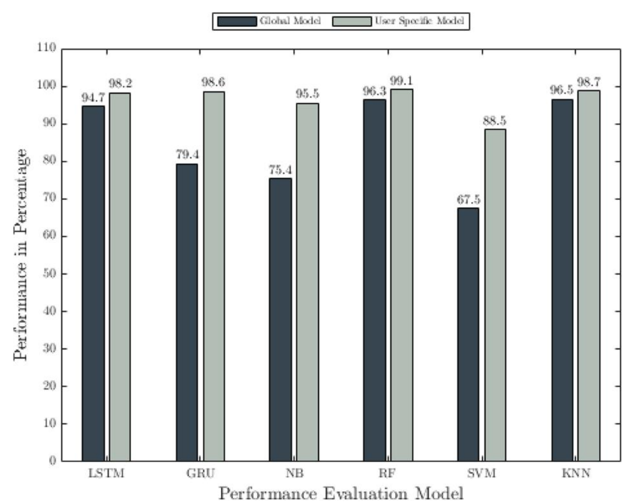


FIG. 15. Comparison of average FAR for activity recognition between global and user-specific models across different ML and DL algorithms.

- Random Forest (RF): Accuracy increased by 27.60% compared to the IMU baseline.
- Long short-term memory (LSTM): Accuracy increased by 41.11%, demonstrating the trajectory data's superior compatibility with deep learning architectures.
- Cross-subject robustness: In rigorous leave-one-subject-out (LOSO) testing, the VIO method achieved 73.96% accuracy, effectively doubling the 43.35% baseline recorded by the IMU-only system.

2. Technical justification for orientation invariance

The primary limitation of traditional IMU-only systems is their high sensitivity to device orientation; raw sensor readings change drastically if a device is tilted or placed differently on the body. Our VIO approach addresses this through:

- Global Pose Estimation: By fuzing IMU data with visual features via the VINS-Mono state estimator, the system calculates the camera's global position and orientation.
- Reference Point Calibration: Unlike IMUs, which suffer from cumulative drift and lack a starting reference, the camera provides visual landmarks to anchor the trajectory in a global frame.
- Feature Transformation: This process transforms local, orientation-dependent sensor signals into a scene-independent trajectory, allowing the model to recognize the fundamental movement pattern (e.g., the specific arc of a "Running" gait) regardless of how the helmet is positioned.

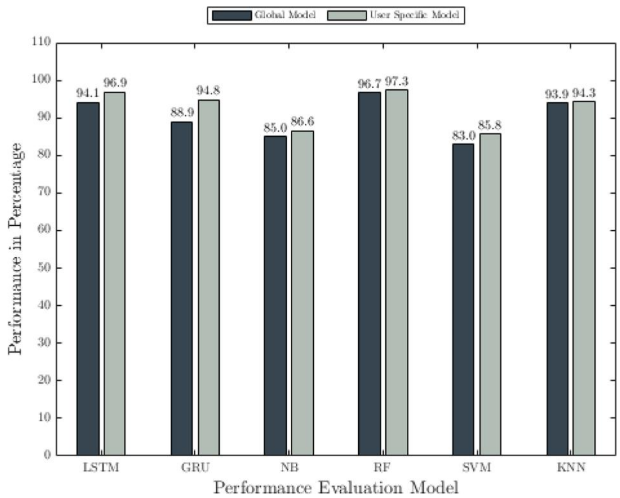


FIG. 16. Comparison of average overall balance accuracy for subject recognition between global and user-specific models across different ML and DL algorithms.

3. False alarm rate (FAR) reduction

The clarity provided by fused trajectories significantly reduced the False Alarm Rate (FAR) from over 8% in the IMU baseline to under 4% in the VIO method. This reduction is critical for real-world applications like elderly care, where minimizing false positives is essential for system trust and reliability.

V. DISCUSSION

From the experimental results and performance metrics presented, we can derive the following critical insights regarding the vioHAR system:

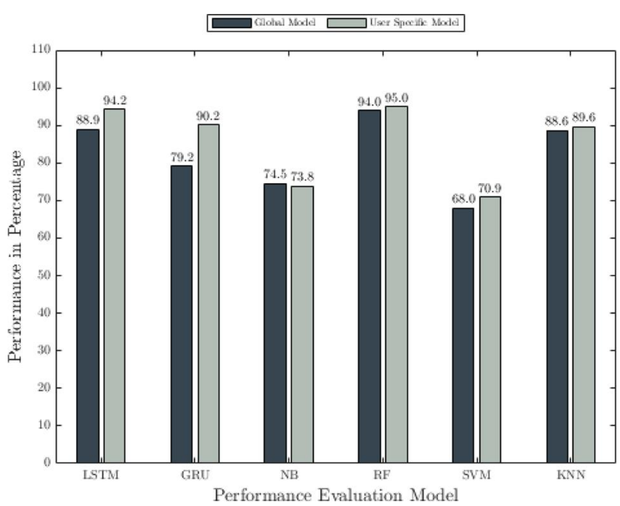


FIG. 17. Comparison of average Macro F-1 for subject recognition between global and user-specific models across different ML and DL algorithms.

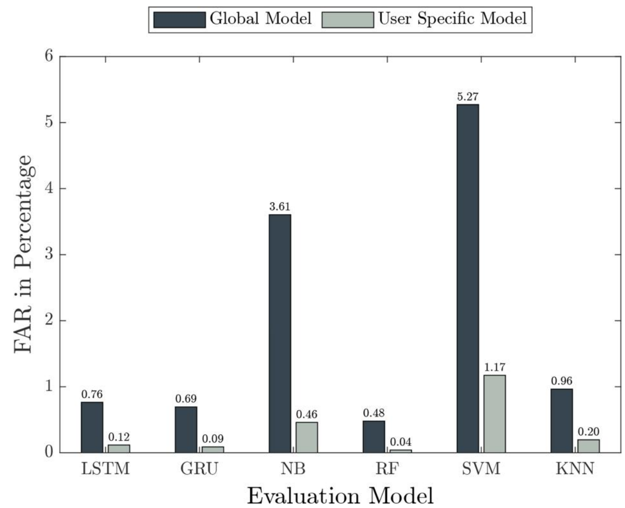


FIG. 18. Comparison of average FAR for activity recognition between global and user-specific models across different ML and DL algorithms.

A. Enhanced trajectory analysis

- Superior Efficiency: Trajectory analysis of human movement is established as an efficient and robust methodology for both human activity recognition (HAR) and patient/subject identification.
- Validation through VIO: The integration of visual-inertial odometry (VIO) data demonstrates a statistically significant improvement over raw IMU signals. As detailed in Table XIII, this approach yielded average accuracy improvements of approximately 27.6% for Random Forest and 41.1% for LSTM models.
- New Benchmarks: In rigorous LOSO testing, the VIO-based trajectory analysis achieved an accuracy of 73.96%, a substantial advancement over the meager 0.4335 achieved by the IMU

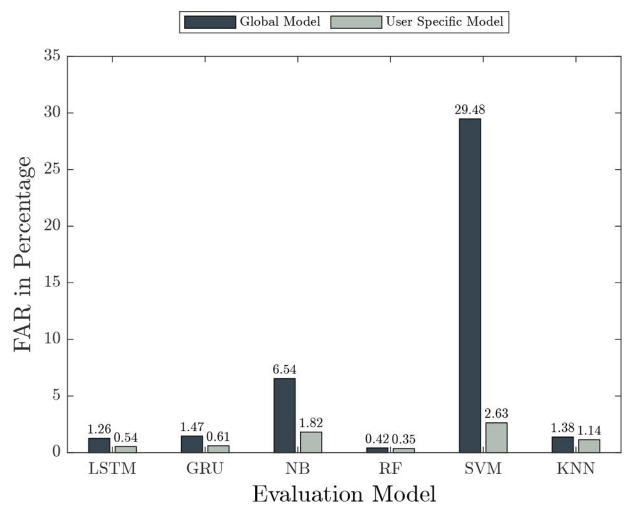


FIG. 19. Comparison of average FAR for subject recognition between global and user-specific models across different ML and DL algorithms.

TABLE X. Performance Comparison: Standard IMU Baseline vs Proposed vioHAR Method. This evaluation quantifies the specific accuracy gains achieved by integrating visual trajectory data compared to using raw accelerometer and gyroscope features alone.

	IMU Baseline (Raw Data) (Standard Accel + Gyro)		Proposed vioHAR Method (Fused Trajectory Data)		vioHAR Improvement (% Increased)	
Activity	RF Accuracy	LSTM Accuracy	RF Accuracy	LSTM Accuracy	RF	LSTM
A01 Walking	0.7958	0.7144	0.9989	0.9996	**+25.5216**%	**+39.9278**%
A02 Running	0.6938	0.6149	0.9981	1.0	**+43.8516**%	**+62.6183**%
A03 Jumping	0.65	0.5685	0.9791	0.9757	**+52.751 66**%	**+71.6255**%
A04 Sitting & Standing	0.826	0.7461	0.9956	0.9891	**+20.5332**%	**+32.57**%
A05 Climbing Stairs	0.7778	0.69	0.9982	0.9929	**+28.3062**%	**+43.8995**%
A06 Descending Stairs	0.8141	0.7523	0.9984%	0.9912	**+22.635**%	**+31.7569**%
A07 Falling Forward	0.8189	0.7870	0.9566%	0.9450	**+16.8075**%	**+20.0783**%
A08 Falling Backward	0.8161	0.7772	0.9008%	0.9825	**+10.38**%	**+26.4130**%
Average Accuracy	**0.773**	**0.7063**	**0.9782**%	**0.00**%	+27.5983%	+41.1112%

TABLE XI. LOSO Cross-Validation Performance: IMU Baseline vs vioHAR Method.

Method	Model	Accuracy	Recall	Precision	F1-Score	Bal. Acc.	FAR (%)
IMU Baseline	LSTM	0.4377	0.4053	0.5120	0.4524	0.6609	0.0835
	Random Forest	0.4335	0.4137	0.4415	0.4271	0.6652	0.0832
vioHAR Method	LSTM	0.7335	0.7315	0.7467	0.7358	0.8539	0.0388
	Random Forest	0.7396	0.7438	0.7389	0.7379	0.8530	0.0378

TABLE XII. Comparison of Global Model Performance: Random Split vs Leave-One-Subject-Out (LOSO). Values represent the aggregate average across all activities.

Metric	Random split (50:50)	LOSO (unseen user)
Accuracy	0.9564	0.7396
Precision	0.9581	0.7438
Recall	0.9525	0.7389
F1-score	0.9553	0.7413
Overall balanced accuracy	0.9667	0.8530
Micro F1-score	0.9360	0.7379
False alarm rate (FAR)	0.0035	0.0380

baseline. This confirms that fused trajectory data is a far more reliable indicator of movement in real-world scenarios.

B. Mathematical modeling and user-centricity

- Predictive Power: Human activities can be effectively modeled mathematically, allowing for user-centric frameworks that reduce the necessity for extensive volunteer involvement during data acquisition.
- Generalization Performance: While user-specific models achieve near-perfect results, reaching an average accuracy of 99.16%. The LOSO findings underscore that these models maintain a solid 85.3% balanced accuracy even when encountering completely unseen subjects. These preliminary results align with our established research.
- Time-Series Superiority: Utilizing time-series data for activity modeling continues to exhibit superior predictive performance, as substantiated by our previous methodologies.

C. Adaptive algorithms and computational efficiency

- Algorithm-Specific Strengths: Random forest (RF) demonstrates slightly better accuracy and stability across global model configurations, whereas long short-term memory (LSTM) networks excel in computational efficiency and temporal dependency management.
- Resource Management: For user-specific models, training time is significantly optimized; LSTM requires only 12.79 s compared to the 94.21 s necessary for global models. This reduction in complexity is consistent with our prior work on localized processing.
- Edge Computing Potential: The extreme efficiency of LSTM in LOSO validation (44.79 vs. 415.8 s for RF) indicates that a VIO-based, user-specific adaptive algorithm can be effectively deployed on edge devices to optimize battery performance and reduce discharge rates.

D. Analysis of misclassification patterns

A detailed examination of the average confusion matrices for activity recognition reveals specific patterns of misclassification that provide insight into the system's performance limits:

TABLE XIII. Detailed Per-Activity Performance Comparison between LSTM and Random Forest (RF) using Leave-One-Subject-Out (LOSO) Cross-Validation. This evaluation highlights the generalization capability of each model for specific activities on unseen subjects.

	Accuracy (Recall)		Precision		F1-Score		Balanced Accuracy		FAR	
Activity	LSTM	RF	LSTM	RF	LSTM	RF	LSTM	RF	LSTM	RF
A01 Walking	0.6192	0.6806	0.7782	0.7916	0.6897	0.7319	0.7880	0.8184	0.0432	0.0438
A02 Running	0.7045	0.7358	0.5333	0.5884	0.6071	0.6539	0.8133	0.8354	0.0779	0.0651
A03 Jumping	0.6015	0.6248	0.6530	0.5812	0.6262	0.6023	0.7755	0.7769	0.0504	0.0711
A04 Sitting & Standing	0.79	0.6877	0.7461	0.7576	0.7674	0.7209	0.8642	0.8186	0.0617	0.0505
A05 Climbing Stairs	0.8326	0.8993	0.8542	0.8844	0.8433	0.8919	0.9044	0.9399	0.0237	0.0196
A06 Descending Stairs	0.8295	0.8410	0.8333	0.8365	0.8314	0.8387	0.9009	0.9068	0.0277	0.0274
A07 Falling Forward	0.8123	0.7292	0.7652	0.8608	0.7880	0.7896	0.9008	0.8621	0.0106	0.0050
A08 Falling Backward	0.784	0.752	0.6885	0.6104	0.7332	0.6738	0.8844	0.8657	0.0152	0.0206
Macro Average	**0.7467**	**0.7438**	**0.7315**	**0.7389**	**0.7358**	**0.7379**	**0.8539**	**0.8530**	**0.0388**	**0.0378**

TABLE XIV. Comparative Performance of Deep Learning (LSTM) and Machine Learning (Random Forest) models using the Leave-One-Subject-Out (LOSO) cross-validation protocol. This evaluation demonstrates model robustness when tested on unseen subjects.

Performance metric	LSTM (LOSO)	Random forest (LOSO)
Accuracy	0.7335	0.7396
Recall (sensitivity)	0.7467	0.7438
Precision	0.7315	0.7389
F1-score	0.7358	0.7379
Overall balanced accuracy	0.8539	0.8530
Micro F1-score	0.7379	0.7358
False alarm rate (FAR)	0.0388	0.0378
Training time (s)	44.7936	415.814

TABLE XV. Training Time for Different Algorithms under User/Activity Specific and Global Models.

Algorithm name	Dataset type	Training time (s) User/activity specific	Training time (s) Global model
KNN	Activity Recognition	0.033 640	0.059 333
LSTM	Activity Recognition	12.792 610	94.219 247
NB	Activity Recognition	0.053 614	0.026 781
RF	Activity Recognition	0.332 476	1.544 113
SVM	Activity Recognition	0.231 232	2.123 858
GRU	Activity Recognition	4.381 260	60.944 978
KNN	Subject Recognition	0.011 334	0.022 965
LSTM	Subject Recognition	16.119 101	165.262 020
NB	Subject Recognition	0.023 057	0.029 642
RF	Subject Recognition	0.373 547	2.820 801
SVM	Subject Recognition	0.227 707	1.338 943
GRU	Subject Recognition	14.820 276	65.646 249

- Static vs Dynamic Transitions: The most frequent misclassifications occurred between Sitting and Standing (A04). This is attributed to the nearly identical vertical trajectory and orientation signatures at the beginning and end of these transitions.
- Jumping vs Vertical Transitions: Some Jumping (A03) instances were misclassified as sitting/standing. Biomechanically, both activities involve a rapid change in the z-axis (vertical) position. While jumping produces a distinct peak, the stabilization phase after a jump can mirror the postural shift seen when a subject sits down.
- Gait Similarities: Minimal confusion was observed between walking (A01) and running (A02), which reached near-perfect accuracy. This suggests that the VIO-derived velocity and trajectory features effectively differentiate between these two gait patterns, even when device orientation varies.
- Falling Scenarios: Falling Forward (A07) and Falling Backward (A08) showed slight confusion with one another. This is likely due to the extremely short duration of these events (under 2 s), which limits the number of temporal features the models can extract for differentiation.
- User-Specific Robustness: Notably, the user-specific models significantly reduced these misclassifications compared to the global models. By training on an individual's unique "movement signature," the models better distinguished between subtle transitions, such as a specific user's way of sitting vs jumping.

VI. CONCLUSION

This research introduces an advanced methodology for human activity recognition (HAR) that leverages visual-inertial odometry (VIO) to achieve superior precision and accuracy. Our findings conclusively demonstrate that the integration of fused trajectory data significantly outperforms traditional IMU-only baselines. Under the rigorous leave-one-subject-out (LOSO) cross-validation protocol, which tests the system on entirely unseen subjects, the proposed method maintained a robust accuracy of 73.96%. This effectively doubles the 43.35% accuracy recorded by the IMU-only baseline, yielding specific accuracy improvements of +27.60% for the Random Forest (RF) model and +41.11% for the long short-term memory (LSTM) model.

The system achieved 100% accuracy in fundamental activities such as walking and running, while subject detection achieved 100% accuracy for 7 of the 14 participants, validating the strength of the proposed feature extraction process. Furthermore, computational analysis revealed that the LSTM network is approximately 10 times more efficient than random forest in LOSO training (requiring only 44.79 s compared to 415.81 s for RF). This suggests that an adaptive, user-specific LSTM algorithm can be effectively deployed on low-power edge devices.

VII. LIMITATIONS

Despite the noteworthy results, this study faced specific restrictions that define its current scope. First, for the ease of our aged care-focused research, a consumer smartphone (*Xiaomi Poco F1*) was employed, positioned at eye level utilizing a motorcycle helmet.[11] We acknowledge that this head-mounted configuration, chosen to simulate emerging AR/VR goggles and smart glasses, differs significantly from typical pocket-based human activity recognition (HAR).[7]

While the current study utilizes data from 14 subjects to establish the vioHAR framework, we acknowledge that a larger and more age-diverse cohort—particularly including elderly participants—would further enhance the model's generalizability. To mitigate the risk of over-fitting to this specific group, we employed LOSO cross-validation, which yielded a robust macro-average accuracy of 73.96% across all subjects. Furthermore, the standard deviation across LOSO folds indicates consistent performance, though future iterations of the vioHAR dataset will aim to expand the subject count to further reduce statistical variability.

Furthermore, technical challenges emerged from the hardware itself. The smartphone's camera features a rolling shutter effect, and in the absence of official calibration values, a conventional rolling shutter parameter was utilized.[11] The existence of IMU noise and asynchronous clocks between the IMU and camera sensors further complicated the enhancement of sensor fusion for long-term state estimation. Finally, the methodology currently relies on offline processing.[11] Future work will focus on addressing unconstrained "in-the-wild" data, exploring diverse sensor placements, and transitioning toward a real-time processing framework to improve the system's overall adaptability.

VIII. METHODS

Our methodology, depicted in Fig. 20, culminated in the classification of activities based on the extracted features. Utilizing machine learning or statistical techniques, our model learned to discern and categorize different activities based on identified patterns. This classification process served as the cornerstone of our efforts to recognize and understand the observed activities.

Our dataset includes 8 activities collected from 14 subjects obtained by using inertial measurement units (IMUs) and camera-equipped device which is a smartphone mounted on a helmet. The acquired raw data undergo postprocessing to provide a series of 3D joint positions for each movement. Furthermore, recordings require synchronization, which involves the simultaneous acquisition of video and IMU sensor data. We conduct additional processing to transform the dataset into a usable format for analysis. Our research project commenced with the careful acquisition of raw data from a varied range of study participants. After obtaining the data, an important part of our workflow was preprocessing, during which we methodically cleaned and arranged the raw data to improve its quality and usability. The purpose of this preliminary stage was to reduce noise and inconsistencies that could have a negative impact on subsequent assessments. Afterwards, we segmented the data, dividing continuous streams into separate and meaningful segments. This segmentation facilitated a more granular analysis, enabling us to extract patterns and insights at a finer level. Feature extraction followed, where relevant and informative characteristics of the segmented data were identified and isolated.

To precisely predict human activity, we need to first figure out the trajectory of the subject. Trajectory generation typically requires information about the object's position, velocity, and orientation over time. Relying solely on an inertial measurement unit (IMU) for trajectory generation presents limitations; while IMUs capture short-term trajectory data effectively, they suffer from significant drift and fail to estimate long-term positions precisely. Moreover, IMUs can provide information about the motion but we cannot get a reference point to start so they cannot measure actual position. Using the IMUs (acceleration and angular velocity) and camera metadata together and sensor fusion technology we can calculate a very precise global position reference. However, integration introduces cumulative errors, leading to drift in the estimated position and orientation. This drift can become significant over time, especially in the absence of external position updates.

To understand the advantage of the vioHAR approach, consider the limitations of standard IMU sensors. Relying solely on an IMU is similar to navigating a room with closed eyes; while one can sense steps and turns (acceleration and rotation), small errors quickly accumulate, leading to a loss of the true position over time (drift). The vioHAR system resolves this by "opening its eyes." By fuzing the IMU

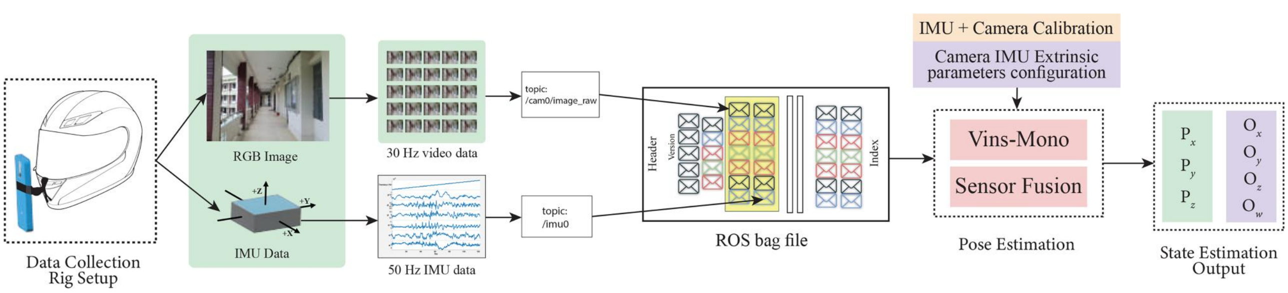


FIG. 20. Comprehensive process of VIO-based human activity recognition (HAR) dataset.

data with video frames, the system continuously identifies static visual landmarks in the environment. These visual reference points act as anchors, constantly correcting the IMU's drift and generating a highly precise, three-dimensional trajectory of the user's movement. This allows the system to recognize the fundamental pattern of an activity regardless of how the device is oriented on the user's body. In this research, we are focusing on elderly people where the user experience and convenience are more important. In the recent announcement of virtual reality (VR) headsets by Apple Vision Pro, smart glass like Ray-Ban Meta smart glasses, or even pocket wearable devices like Rabbit R1 AI device, Humane Ai Pin, which all have equipped with the camera sensor and IMU sensor built in. In our research, we are focusing on these targeted product areas where we'll be able to detect and identify human activities very accurately.

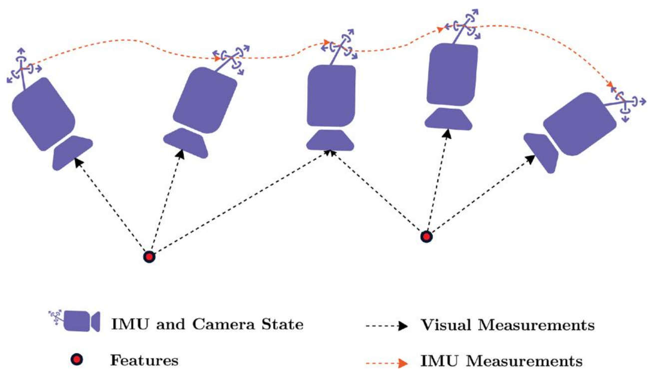


FIG. 22. An illustration of visual-inertial localization and mapping.

A. Data acquisition

To gather the raw data, we employed a Xiaomi Poco F1 smartphone running Android version 10 QKQ1.190828.002 and MIUI version MIUI Global 12.0.3 Stable (QEJMIXM). This device is equipped with a 12MP Sony IMX363 1/2.55″ Exmor RS stacked CMOS sensor for the camera and a Bosch BMI160 IMU sensor. This smartphone was chosen for its accessibility, cost-effectiveness, and the capabilities of its integrated sensors. To record the human activity data, we have used the Mobile Augmented Reality Sensor (MARS) Logger Android app developed by OSU Photogrammetric Computer Vision Lab.[36] Addressing the challenge of accessing comprehensive data from mobile AR sensors, the MARS logger is a viable solution. The purpose of the MARS logger is to collect visual and inertial data for Simultaneous Localization and Mapping (SLAM) and Augmented Reality (AR) applications using commodity mobile devices. It records camera frames at approximately 30 Hz and inertial measurement unit (IMU) measurements at about 50 Hz synced to one clock source on Android (API 21+) shown in Fig. 21.

The IMU sensor features a 3-axis, 16-bit resolution accelerometer and gyroscope, with noise parameters of $120\,\mu\text{g}/\sqrt{\text{Hz}}$ for the accelerometer and $0.017^\circ/\text{s}/\sqrt{\text{Hz}}$ for the gyroscope. Camera frames are recorded and stored as H.264/MP4 videos with a video frame resolution of 1280×720. The recording process is facilitated using the Camera2 API, OpenGL ES, and MediaCodec along with MediaMuxer. The OpenGL library is utilized for efficient rendering, while background processes handle frame encoding and I/O operations. The metadata associated with camera frames is saved to a CSV file, and timestamps for each frame are recorded in a txt file. Simultaneously, inertial data are captured on a background HandlerThread. This metadata encompasses crucial parameters such as rolling shutter skew and focus distance shown in Fig. 22, providing valuable information for SLAM (simultaneous localization and mapping) algorithms.

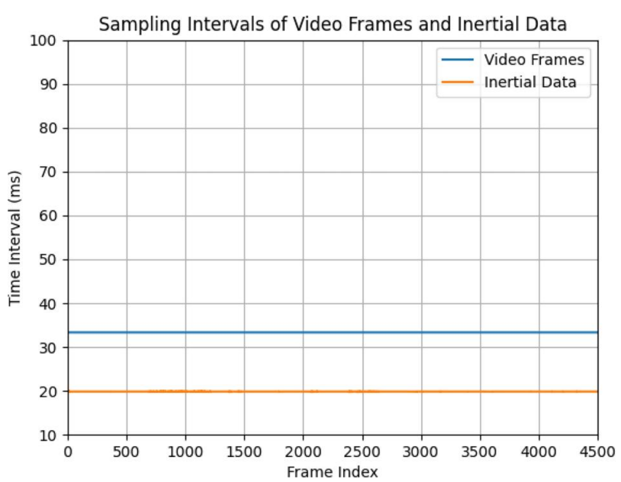


FIG. 21. Sampling intervals of captured visual inertial data with Xiaomi Poco F1.

There are multiple publicly available datasets for ADL and Falls. Most of the datasets are collected using consumer-level smartphones and smartwatches. For recording the raw data, the accelerometer and gyroscope are the top choices of all those datasets. In some of the datasets, we observe the use of magnetometer, light and sound sensors. Some datasets use both smartphones and smartwatches. What we have observed in these publicly available datasets is that there's no camera sensor used to capture the raw signal. In Table I, we can see a clear view of the different approaches to those datasets.

B. Data processing

Android devices cannot ensure the consistency of the requested sampling rate, necessitating the definition of noise parameters in offline processing. Our annotations play a crucial role in processing individual activity data. For this processing, we employed the VINS-Mono (visual-inertial odometry system for monocular cameras), an advanced open-source visual-inertial odometry system. VINS-Mono,[37] a robust and versatile monocular visual-inertial state estimator that combines pre-integrated IMU measurements and feature observations to achieve highly accurate visual-inertial odometry. The system utilizes only one camera, alongside inertial sensors like accelerometers and gyroscopes, to estimate the camera's pose and movement. Its support for rolling shutter cameras, like the smartphone camera we used with a rolling shutter parameter of approximately 0.020, is noteworthy. VINS-Mono dramatically enhances motion-tracking performance by addressing challenges such as the loss of visual tracks due to factors like illumination changes, textureless areas, or motion blur.

C. Data segmentation and model configuration

To prepare the extracted visual-inertial odometry (VIO) trajectory data for classification, several preprocessing and architectural parameters were defined. These ensure the models effectively capture the temporal dynamics of human motion while maintaining computational efficiency.

- Windowing and Overlap: The continuous streams of trajectory data (Position x, y, z and Orientation x, y, z, w) were segmented using a sliding window approach. We utilized a window size of 2.56 s, which corresponds to 128 samples at a 50 Hz sampling rate. To ensure continuity between segments and prevent the loss of transitional motion patterns, a 50% overlap was applied between successive windows.
- Input Tensor Shape: For the deep learning models, specifically the Long short-term memory (LSTM) and gated recurrent unit (GRU) networks, the input data were reshaped into a three-dimensional tensor. The input shape was defined as $[Batch\_Size, Time\_Steps, Features]$, where *Time_Steps* represents the 128 samples per window and *Features* includes the 7 parameters derived from the VIO trajectory (3 position coordinates and 4 orientation quaternions).
- Data Normalization: To ensure stable gradient descent and prevent features with larger numerical ranges (such as global position) from dominating the learning process, the trajectory data underwent normalization before being fed into the classifiers. We implemented Min-Max scaling to transform all feature values into a uniform range of $[0, 1]$, which is particularly beneficial for the activation functions used in the LSTM and GRU architectures.

D. Deep learning model architectures and hyperparameters

To ensure the reproducibility of our results, the configurations for the long short-term memory (LSTM) and gated recurrent unit (GRU) networks are detailed below:

- Input Layer: The models accept a three-dimensional input tensor with a shape of `[Batch_Size, 128, 7]`.
- Time Steps: 128 samples, representing a 2.56-s window at a 50 Hz sampling rate.
- Features: seven total parameters, consisting of three global position coordinates (x, y, z) and four orientation quaternions (x, y, z, w) derived from the VIO trajectory.
- Network Depth and Units: Both the LSTM and GRU architectures consist of two hidden layers, each containing 64 recurrent units to effectively capture the temporal dynamics of human motion.
- Activation and Dropout: A tanh activation function is utilized for the recurrent steps, while a Softmax layer is applied to the final output for class probability estimation. To prevent overfitting, a dropout rate of 0.2 was implemented between the recurrent layers.
- Optimization and Training:
- Optimizer: The Adam optimizer was employed for stable gradient descent.
- Loss Function: Categorical cross-entropy.
- Learning Rate: Set to 0.001.
- Training Epochs: Models were trained for 50 epochs with an early stopping mechanism to monitor validation loss.
- Training Efficiency: In the LOSO validation, the LSTM demonstrated superior computational efficiency, completing training in approximately 44.79 s, which is ten times faster than traditional machine learning models like Random Forest.

E. Camera-IMU calibration

Understanding the spatial transformations and temporal offsets between cameras and IMUs is essential for achieving optimal performance in visual-inertial odometry (VIO). The offline spatial calibration of cameras and IMUs has been extensively studied, with solutions employing filters and batch optimization. Newer VIO algorithms often incorporate the unknown spatial transformation between cameras and IMUs into their states, estimating it alongside the motion of the sensor suite for online self-calibration. When dealing with unsynchronized visual-inertial sensor suites, the estimation of temporal offsets between cameras and IMUs becomes necessary. Kalibr is a widely utilized tool for both spatial and temporal calibration of camera-IMU systems.[38] In real-life scenarios, sensor timestamps are affected by activation and transmission delays, leading to temporal misalignment among sensors. This misalignment considerably affects sensor fusion reliability. Our approach involves adjusting the temporal disalignment between visual and inertial data through the joint optimization of time offset, camera and IMU states, and feature positions inside a SLAM (simultaneous localization and mapping) framework. The IMU configuration parameters are enlisted in Table XVI.

The camera projection parameters were set according to the values provided. The camera time offset relative to the IMU can be set to zero if the camera and IMU timestamps do not differ significantly. In our case, we calculated every time offset between camera and IMU which we have provided in the VINS-Mono config YAML file. The IMU and camera on nearly all smartphones (Android or iOS) have the same nominal orientation as shown in Fig. 23. Because the IMU

TABLE XVI. IMU Configuration parameters.

VIO	Gyro white noise	Accelerometer white noise	Gyro random walk	Accelerometer random walk
VINS-Mono	24.0×10^{-3}	16.0×10^{-2}	2.0×10^{-4}	5.5×10^{-4}

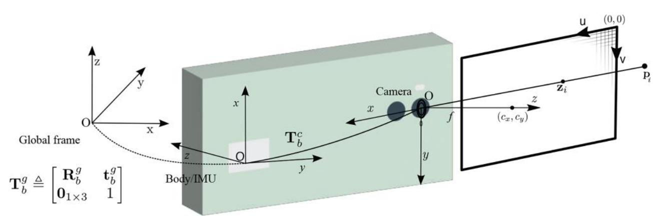


FIG. 23. Camera and IMU body frame to global frame.

and camera are not too far, within 10 cm, their relative translation can be safely ignored for practical VINS. Thus, their relative pose is

$$\text{T\_imu\_cam} = [0, -1, 0, 0; -1, 0, 0, 0; 0, -1, 0, 0; 0, 0, 0, -1]$$

such that point of homogeneous coordinates in imu frame = T_imu_cam ∗ point of homogeneous coordinates in the camera frame.

An IMU remains scene-independent, making it impervious to challenges encountered by cameras, particularly in scenarios with low texture, high speed, and High Dynamic Range (HDR) conditions. Moreover, an IMU boasts a high output rate. However, it does face challenges such as a poor signal-to-noise ratio at low accelerations and low angular velocities. The presence of sensor biases in an IMU's motion estimation often leads to quick drift accumulation. Hence, a combination of both cameras and IMUs proves effective in achieving accurate and robust state estimation across various situations. The IMU measures the angular velocity ω and the external acceleration a

$$\omega = I\omega + b_g + n_g, \tag{1}$$

$$a = R_{IW}({}_{W}a{-}_{W}g) + b_a + n_a, \tag{2}$$

where Iω is the angular velocity of the IMU expressed in the IMU frame, W_a is the acceleration of the IMU in the world frame, and W_g is the gravity in the world frame. b and n are the biases and additive noises respectively.[39]

In VIO, the environment is represented through a set of 3D landmarks W_p, projected by the camera to 2D image coordinates u,

$$u = \text{project}(T_{CW} \cdot W_p). \tag{3}$$

Cameras and IMUs are frequently combined in low-cost and self-assembled visual-inertial sensor configurations that lack strict time synchronization. Triggering and transmission delays may result in a discrepancy between the generated timestamp and the actual sampling time. Consequently, a temporal offset commonly exists between different measurements, as illustrated in Fig. 24. The time offset t_d is the amount of time by which we should shift the camera timestamps so that the camera and IMU data streams become temporally consistent. t_d may be a positive or negative value. Specifically, we define t_d as

$$t_{IMU} = t_{CAM} + t_d. \tag{4}$$

If the camera sequence has a longer latency than the IMU sequence, t_d is a negative value. Otherwise, t_d is a positive value.[39] In typical cases, the time offset between sensors is a constant but unknown value. In more challenging situations, sensors may have different clocks, leading to a time offset drift over time, rendering them unsuitable for sensor fusion. This paper considers the general case where the time offset t_d is a constant but unknown value. The timestamp generated by the camera sensor and IMU is not synchronized due to delays in triggering, transmission, and unsynchronized clocks. This results in a temporal misalignment between the camera and IMU.

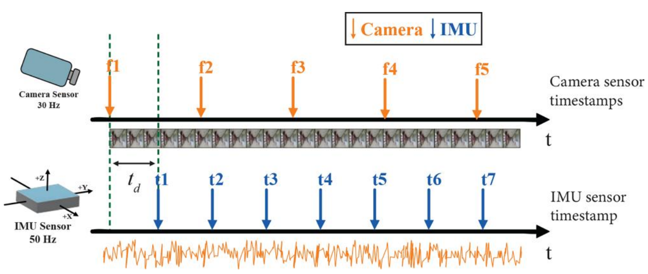


FIG. 24. An illustration of temporal misalignment (time offset) between camera and IMU streams.

F. Trajectory analysis

Using various techniques and processes, our objective was to capture eight different types of daily human activities. This procedure entailed integrating data from Visual-Inertial Odometry (VIO), enabling the combination of visual and inertial sensor information. Sensor fusion was essential in improving the accuracy and reliability of the created trajectories, offering a thorough comprehension of spatial dynamics by considering both visual characteristics and inertial data. The timestamp generated by the camera sensor and IMU is not synchronized due to delays in triggering, transmission, and unsynchronized clocks. This results in a temporal misalignment between the camera and IMU.

To acquire trajectory data from VINS-Mono, our raw data underwent conversion to a format optimized for running on VINS-Mono. This conversion involved creating a rosbag file using the `kalibr_bagcreater` tool from the Autonomous Systems Lab @ ETH Zürich.[40] Running the rosbag file in VINS-Mono requires a configuration YAML file. The rosbag file contains topics for IMU sensor data and camera image sequences. Following the preparation of VINS-Mono, we executed the rosbag file, resulting in the generation of an accurate trajectory as depicted in Figs. 25 and 26. VINS-Mono performed calculations on the trajectory points, and the resulting trajectory data were saved in CSV files. Subsequently, these data were utilized for additional processing and evaluation in the subsequent steps of our research.

The crucial part of this research is to generate a precise trajectory using the consumer-level sensor, which will be able to provide better results in the actual use case of the HAR research. The goal of this study is not only to promote the development of HAR technologies but also to provide a valuable dataset to the researchers and

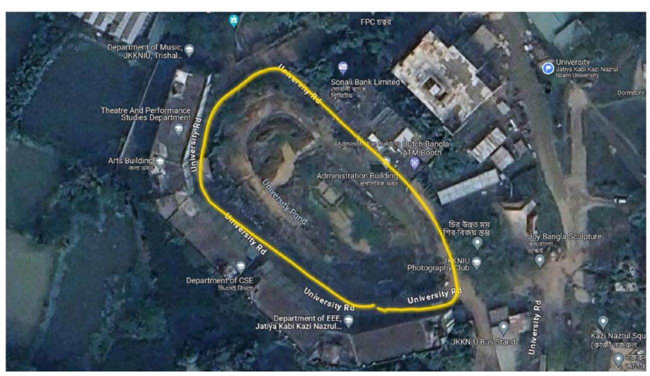


FIG. 25. Outdoor experimental results: Monocular visual-inertial state estimator. Trajectory overlaid with Google map for visual comparison.

FIG. 26. Experimental results: Climbing stairs with monocular visual-inertial state estimator. Trajectory overlaid over stairs for visual comparison.

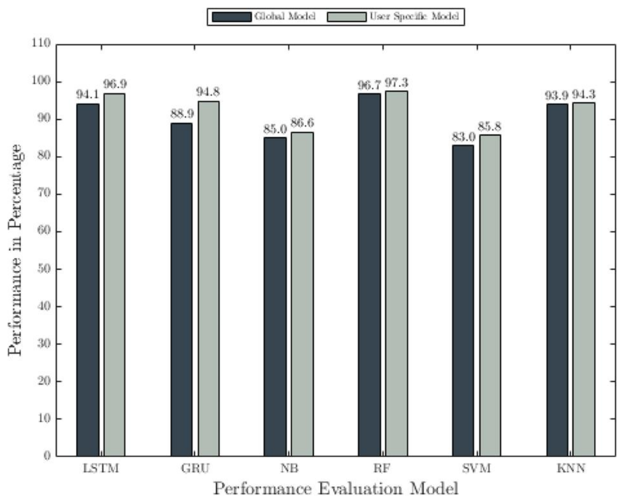


FIG. 27. Comparison of average overall balance accuracy for subject recognition between global and user-specific models across different ML and DL algorithms.

professionals who are actively working in this field. Our focus is to generate an accurate trajectory for the activities of daily living (Fig. 27).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Kazi Md. Shahiduzzaman: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Resources (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Md Salah Uddin Yusuf:** Conceptualization (equal); Investigation (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Md Sajjad Hossen:** Data curation (equal); Formal analysis (supporting); Resources (equal); Software (equal); Validation (supporting); Visualization (equal); Writing – review & editing (lead). **Md Nazmul Alam Munna:** Data curation (supporting); Resources (equal); Software (supporting); Validation (supporting); Visualization (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in Figshare at https://figshare.com/s/911dc88878c5b5593e06, Ref. 41.

REFERENCES

[1]M. H. Arshad, M. Bilal, and A. Gani, "Human activity recognition: Review, taxonomy and open challenges," Sensors **22**, 6463 (2022).
[2]N. Ahmed, J. I. Rafiq, and M. R. Islam, "Enhanced human activity recognition based on smartphone sensor data using hybrid feature selection model," Sensors **20**, 317 (2020).
[3]D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," Appl. Sci. **7**(10), 1101 (2017).
[4]E. Casilari and C. A. Silva, "An analytical comparison of datasets of real-world and simulated falls intended for the evaluation of wearable fall alerting systems," Measurement **202**, 111843 (2022).
[5]K. M. Shahiduzzman, J. Peng, Y. Gao, X. Hei, and W. Cheng, "Towards accurate and robust fall detection for the elderly in a hybrid cloud-edge architecture," in *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (IEEE, 2019), pp. 1626–1631.
[6]K. M. Shahiduzzaman, X. Hei, C. Guo, and W. Cheng, "Enhancing fall detection for elderly with smart helmet in a cloud-network-edge architecture," in *IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)* (IEEE, 2019), pp. 1–2.
[7]K. M. Shahiduzzaman and M. S. U. Yusuf, "Generating a visual-inertial odometry dataset based on a helmet prototype for recognizing human activities," in *International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)* (IEEE, 2022), pp. 1–6.
[8]K. M. Shahiduzzaman and M. S. U. Yusuf, "EdgeFall: A promising cloud-edge-end architecture for elderly fall care," Int. J. Electr. Comput. Eng. **13**, 4721 (2023).
[9]K. M. Shahiduzzaman and M. S. U. Yusuf, "Performance enhancement of human activity recognition and fall related system using multi-modal broad learning system (mBLS)," AIP Conf. Proc. **2901**, 070019 (2023).
[10]D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, and M. R. Luaces, "A public domain dataset for real-life human activity recognition using smartphone sensors," Sensors **20**, 2200 (2020).
[11]K. M. Shahiduzzaman, M. S. U. Yusuf, and M. S. Hossen, "uActivity: A user-specific human activity recognition and fall detection for elderly care," Eng. Technol. Appl. Sci. Res. **15**, 30169–30174 (2025).
[12]M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," Future Gener. Comput. Syst. **81**, 307–313 (2018).
[13]L. Kisselburgh and J. Beever, "The ethics of privacy in research and design: Principles, practices, and potential," in *Modern Socio-Technical Perspectives on Privacy* (Springer, 2022), pp. 395–426.
[14]T. Xu, H. Se, and J. Liu, "A fusion fall detection algorithm combining threshold-based method and convolutional neural network," Microprocess. Microsyst. **82**, 103828 (2021).
[15]B. M. V. Guerra, E. Torti, E. Marenzi, M. Schmid, S. Ramat, F. Leporati, and G. Danese, "Ambient assisted living for frail people through human activity recognition: State-of-the-art, challenges and future directions," Front. Neurosci. **17**, 1256682 (2023).

[16]C. M. Patil, B. Jagadeesh, and M. N. Meghana, "An approach of understanding human activity recognition and detection for video surveillance using HOG descriptor and SVM classifier," in *International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)* (IEEE, 2017), pp. 481–485.
[17]Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition," IEEE Access **5**, 3095–3110 (2017).
[18]D. Yacchirema, J. S. de Puga, C. Palau, and M. Esteve, "Fall detection system for elderly people using IoT and ensemble machine learning algorithm," Pers. Ubiquitous Comput. **23**, 801–817 (2019).
[19]Q. T. Huynh, U. D. Nguyen, L. B. Irazabal, N. Ghassemian, and B. Q. Tran, "Optimization of an accelerometer and gyroscope-based fall detection algorithm," J. Sens. **2015**, 1–8.
[20]R. Ratra, P. Gulia, and N. S. Gill, "Performance analysis of classification techniques in data mining using WEKA," in Proceedings of the International Conference on Innovative Computing & Communication (ICICC) (2021). .
[21]M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," Neurocomputing **100**, 144–152 (2013).
[22]J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy, "Wearable sensors for reliable fall detection," in IEEE Engineering in Medicine and Biology 27th Annual Conference (2005).
[23]Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou, "Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information," in Sixth International Workshop on Wearable and Implantable Body Sensor Networks (2009).
[24]O. Ojetola, E. Gaura, and J. Brusey, "Data set for fall events and daily activities from inertial sensors," in Proceedings of the 6th ACM Multimedia Systems Conference (2015).
[25]L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," Int. J. Comput. Vision **87**, 4–27 (2010).
[26]G. Bo, K. Onishi, T. Takiguchi, and Y. Ariki, "3D human pose estimation from a monocular image using model fitting in eigenspaces," J. Software Eng. Appl. **3**, 1060–1066 (2010).
[27]J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," Comput. Vision Image Understanding **117**, 633–659 (2013).
[28]J. Wang, D. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," IEEE Trans. Pattern Anal. Mach. Intell. **30**, 283–298 (2008).
[29]R. Urtasun, D. J. Fleet, A. Geiger, J. Popović, T. J. Darrell, and N. D. Lawrence, "Topologically-constrained latent variable models," in Proceedings of the 25th International Conference on Machine learning (ICML '08) (2008).
[30]M. Janidarmian, A. Roshan Fekr, K. Radecka, and Z. Zilic, "A comprehensive analysis on wearable acceleration sensors in human activity recognition," Sensors **17**, 529 (2017).
[31]E. Casilari, R. Luque, and M.-J. Morón, "Analysis of android device-based solutions for fall detection," Sensors **15**, 17827–17894 (2015).
[32]P. E. Caicedo, C. F. Rengifo, L. E. Rodriguez, W. A. Sierra, and M. C. Gómez, "Dataset for gait analysis and assessment of fall risk for older adults," Data Brief **33**, 106550 (2020).
[33]L. W. F. Sy, N. H. Lovell, and S. J. Redmond, "Estimating lower limb kinematics using a lie group constrained extended Kalman filter with a reduced wearable IMU count and distance measurements," Sensors **20**, 6829 (2020).
[34]M. Ciliberto, V. F. Rey, A. Calatroni, P. Lukowicz, and D. Roggen, "Opportunity++: A multimodal dataset for video- and wearable, object and ambient sensors-based human activity recognition," Front. Comput. Sci. **3**, 792065 (2021).
[35]S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," Knowl.-Based Syst. **223**, 106970 (2021).
[36]J. Huai, Y. Zhang, and A. Yilmaz, "The mobile AR sensor logger for android and iOS devices," in IEEE Sensors (2019). .
[37]T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," IEEE Trans. Rob. **34**, 1004–1020 (2018).
[38]D. Scaramuzza and Z. Zhang, "Visual-inertial odometry of aerial robots," in *Encyclopedia of Robotics* (Springer, 2020), pp. 1–9.
[39]T. Qin and S. Shen, "Online temporal calibration for monocular visual-inertial systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2018).
[40]Ethz-Asl, see https://github.com/ethz-asl/kalibr for "ethz-asl/kalibr: The Kalibr visual-inertial calibration toolbox."
[41]K. M. Shahiduzzaman, H. M. S. Yusuf, M. Salah Uddin, and M. N. A. Munna (2025). "vioHAR: A novel approach on developing visualinertial odometry-based human activity recognition dataset," Figshare. https://figshare.com/s/911dc88878c5b5593e06