JOIV : Int. J. Inform. Visualization, 9(2) - March 2025 779-788

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Ensemble Approach for Enhanced Classification of Timed Up and Go Test Movements

Yudhi Ardiyanto [a,b,*], Kusworo Adi [c], Kurnianingsih [d]

[a] Doctoral Program of Information System, School of Postgraduate Studies, Diponegoro University, Jawa Tengah, Indonesia
[b] Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta, Bantul, Indonesia
[c] Department of Physics, Faculty of Science and Mathematics, Diponegoro University, Jawa Tengah, Indonesia
[d] Department of Electrical Engineering, Politeknik Negeri Semarang, Jawa Tengah, Indonesia
Corresponding author: *yudhi.ardiyanto@umy.ac.id

***Abstract*—This study aims to evaluate the classification accuracy of a video-based system for Timed Up and Go (TUG) subtasks using human pose estimation through MediaPipe. Six participants were included in the validity study, all participating in the reliability study, performing various TUG subtasks. The research methodology involved acquiring video data that captured the participants' movements during the TUG activity. This video data was processed using the MediaPipe package to extract key points from each frame, resulting in a 2D skeletal representation. The dataset was imported in CSV format to train multiple machine learning algorithms. The dataset was partitioned into training data (70%) and test data (30%), and several machine learning models, including Stacking Ensemble, Hist Gradient Boosting, XGBoost, CATBoost, Random Forest, and Gradient Boosting, were evaluated for their effectiveness in classifying TUG subtasks. The evaluation was conducted by comparing the classification accuracy of each model with the posture detection outcomes and overall performance metrics. The results indicated that the Stacking Ensemble method achieved the highest overall accuracy (96.90%), outperforming models such as Hist Gradient Boosting (96.48%), XGBoost (95.63%), CATBoost (96.06%), Random Forest (95.92%), and Gradient Boosting (95.21%). Each classifier was evaluated across sub-activities, and the results consistently demonstrated the superior performance of the Stacking Ensemble. These findings suggest that the video-based system, when combined with advanced machine learning techniques and human pose estimation, is a reliable and accurate tool for measuring and classifying subtask movements in TUG among older adults.**

***Keywords*—Ensemble learning; human pose; fall risk; TUG test; elderly people.**

Manuscript received 8 Jul. 2024; revised 19 Aug. 2024; accepted 14 Oct. 2024. Date of publication 31 Mar. 2025.

I. INTRODUCTION

According to data from the World Health Organization, falls are the second most common cause of unintentional fatalities worldwide. Individuals over the age of 60 are particularly vulnerable, with falls often leading to fatal outcomes. Each year, serious falls requiring medical attention affect approximately 37.3 million people, underscoring the need for a comprehensive prevention strategy. Such a strategy should prioritize education, training, the creation of safer environments, and developing effective policies to reduce the risk of falls. Additionally, research focused on fall prevention should be prioritized [1]. Extensive research has been conducted to develop technologies to improve the quality of life for older adults. One notable advancement is the development of fall detection technologies. Mubashir et al. classified fall detection methods into three categories: wearable sensors, ambient sensors, and camera or vision systems. The field of fall detection technology continues to advance, with machine learning algorithms playing a key role in fall prevention [2]. Usmani et al. categorize systems into two distinct groups: non-wearable systems and wearable systems [3].

U-Fast technology utilizes a tri-axis accelerometer and gyroscope sensor integrated into a smartphone. In the event of fall, the system is capable of notifying registered family members via telephone and Short Message Service (SMS). The smartphone is placed in the left shirt pocket, and the location of the elderly individual can be determined using Global Positioning System (GPS) coordinates. In addition to detecting different types of falls, the system can classify various activities, such as walking and running [4]. Another innovative approach for detecting falls and daily activities in older adults involves

the use of a Shimmer™ wireless sensor attached to the chest. This device is equipped with a triaxial accelerometer sensor, and the extracted data from both the spatial and frequency domains were used to train a machine learning model. The goal was to distinguish falling events from non-falling events and identify falls from other daily activities. The system successfully classified six distinct daily activities and detected nine different fall patterns, resulting in the development of the 'ShimFall&ADL' dataset [5]. Recently, researchers have created fall detection and ADL datasets by utilising wearable sensors, in addition to the existing datasets. The sensors encompass accelerometers, gyroscopes, and magnetometers, among other types [6], [7], [8], [9]. The purpose is to create a model that can identify irregularities in the care of older individuals by analyzing their vital signs, the environment in which they live, and their mobility patterns [10].

Falls in the elderly are caused by two primary factors: intrinsic and extrinsic. Intrinsic factors refer to conditions within the individual, such as demographic characteristics, comorbid diseases, and impaired vision. Extrinsic factors are external conditions that increase the risk of falling, such as the use of multiple medications, inadequate lighting, or slippery floors [11].

Accurate fall risk assessment involves compiling and analyzing multiple risk factors, which can be challenging to identify and evaluate. Intrinsic factors necessitate intensive medical examination, while extrinsic factors can vary with environmental conditions and time. Fall risk assessment is technically complex because not all gait abnormalities are directly associated with a high risk of falls, making gait analysis alone insufficient for predicting falls. Additionally, some risk factors may occur intermittently, requiring continuous and real-time gait monitoring. A brief outpatient visit may not provide clinicians with sufficient time to detect and objectively evaluate these factors, emphasizing the need for remote monitoring outside hospital settings. The Inertial Measurement Unit is one sensor that can be used for gait analysis [12].

Screening for fall risk in hospitals can help identify patients at risk of injury and prevent falls. A systematic approach is needed to ensure timely and effective screening of patients using risk assessment tools. However, certain considerations should be taken into account before implementing these tools in every inpatient setting. Screening tools should be easy and quick to administer. The introduction of assessment tools necessitates the training of clinical staff, and simpler tools can facilitate the learning process and ensure consistent and accurate application. This is particularly important in hospital management, where high workloads prevail, especially since periodic reassessment is required [13]. Fall risk assessment encompasses a wide range of evaluations to determine fall risk. Various methods are employed in this process, one of which involves administering a series of questions. Based on the responses, the physiotherapist evaluates the patient's fall risk level according to established standards [14].

Fall risk assessment tools can be broadly categorized into two types: Multifactorial Assessment Tools (MAT) and Functional Mobility Assessments (FMA). MAT covers a wide range of fall risk factors, while FMA focuses more on physiological conditions such as balance, gait, and related factors. In this process, the assessor, typically a physiotherapist or physician, instructs the subject to perform specific physical activities. The assessor monitors these activities and compares them against established standards [15]. Several fall risk assessments use a series of functional tests, such as the Berg Balance Scale (BBS), Mini BBS, 5 Times Sit to Stand (5TSTS) test, Timed Up and Go TEST (TUGT), and others [16]. The TUG test is an adaptation of the Get-Up and Go test, modified to include time as a factor for test completion. The equipment required includes an armchair with a height of approximately 46 cm, a 3-meter track area, and a stopwatch. In the TUG test, the participant begins seated in the chair with their back against the backrest, arms resting on the armrests, and, if necessary, a walking aid in hand. Upon the physiotherapist's instruction to "go," the participant must rise from the chair and walk at a comfortable, safe speed along the 3-meter track, then turn around, return to the chair, and sit down. [17].

The Timed Up and Go Test (TUGT) is a rapid, straightforward, and highly efficient tool for evaluating mobility and fall risk. Its minimal equipment and time requirements make it suitable for widespread use in both clinical and community settings. With a 15-second threshold, the TUGT demonstrates optimal sensitivity and specificity, making it a robust predictor of fall risk, particularly when combined with cognitive evaluations. Its user-friendliness and adaptability across diverse populations highlight its importance as an effective screening tool for fall prevention programs [18]. The TUGT is one of the tests recommended by the World Guidelines for the Prevention and Management of Falls in Older Adults [19].

There are several categories of fall risk assessments based on the time required to complete a series of tests. The first is the Timed Up and Go Test (TUGT), one of the most widely used fall risk assessment tools. In this test, participants are asked to stand up from a chair, walk 3 meters, turn around, walk back 3 meters, and sit down again. The Berg Balance Scale (BBS) is another fall risk assessment tool, but it takes longer to administer compared to the TUG test, as it involves 14 different activities. The Tinetti test, which has several variations, is also used for fall risk assessment. One version of the test, the Performance Oriented Mobility Assessment (POMA), takes approximately 20 minutes to complete [20].

Eichler et al. applied a Microsoft Kinect camera to capture characteristics from each phase of the Berg Balance Scale test. The categorization process was performed using machine learning techniques. According to their fall risk prediction model, the 14 activities of the Berg Balance Scale test can be reduced to 4 to 6 activities. The experimental results, referred to as the Efficient-Berg Balance Scale (E-BBS), demonstrate that the number of tasks can be reduced by approximately 50%, while still maintaining an accuracy level of 97%. The assessment results are classified into three categories: low, medium, and high fall risk. This study utilized two cameras in total [21]. Kampel et al. [22] presented an automated TUG method using an RGB-D camera. It employed an automated subtask approach to assess functional decline in 11 elderly individuals with Kinect Version 2. Rule-based strategies utilized features such as shoulder z-axis velocity in conjunction with other parameters. Researchers have since developed alternative methods for various purposes, including an innovative deep learning-based approach for

segmenting subtasks in the TUG test. This system uses a single RGB-D camera and a dilated temporal convolutional network [22].

Previous research on the automated segmentation of fall risk assessment subtasks can be categorized into four types based on the technology used: wearable devices, video-based systems, ambient technologies, and smartphone-based solutions. Each technology has its own advantages and limitations. Video-based technology offers several advantages, including being non-intrusive, as the device does not need to be attached to the body, and the ability to synchronize with other technologies. Additionally, video recordings can be replayed for later assessment, providing a valuable tool for detailed analysis. However, this approach also has limitations. Privacy concerns are significant, as individuals may be uncomfortable being recorded. In crowded environments, multiple people within the camera's field of view can lead to confusion or misidentification. The camera's viewing area can be obstructed, and it must be positioned correctly to capture the necessary footage. Furthermore, effective use of video-based technology requires adequate lighting, which may not always be available [23].

The use of video-based systems has gained increasing attention in movement analysis. The markerless video-based approach is a highly adaptable method for data collection, allowing participants to move naturally in various ambient settings. However, few studies have examined TUG subtasks using traditional video-based methods. One study employed the Microsoft Kinect environmental sensor to automate this process, reducing the subjectivity of outcome measurements and providing additional data on patient performance. The Kinect's depth imaging automatically detects each stage of the TUG test [24]. A new system was developed to automate the TUG test using the Kinect camera, version 2. This system was specifically designed to directly compare the performance of RGB and RGB-D based techniques. The methodology uses advanced machine learning and refinement techniques to generate 3D skeletal structures from a single RGB video. The effectiveness of both the proposed deep learning-based and Kinect-based RGB-D skeletons is then evaluated in segmenting the TUG test, using manually labeled ground truth data for comparison [25].

Other researchers developed a video-based system that allows for the assessment of individual movement characteristics. The objective of this study was to investigate the accuracy and consistency of a video-based system for measuring the speed of several tasks within the Timed Up and Go (TUG) test among older adults. The validity study involved twenty older participants, while the reliability study included ten older adults. We measured the speed at which participants completed each subtask of the TUG test under both comfortable and fast speed conditions across two sessions. The Pearson correlation coefficient was used to evaluate the validity of the video-based system compared to the motion analysis method [26].

There remains a need for further development of technologies capable of accurately measuring TUG and 5TSTS repeatedly and without continuous supervision in community settings or therapeutic rehabilitation environments. Dependable, closely monitored measurements conducted by older adults in such settings are crucial. These systems utilize a range of sensors, including RGB-D cameras, RFID, accelerometers, gyroscopes, magnetometers, and barometers [27], [28]. Another system was developed using a Raspberry Pi embedded system equipped with three cameras and additional sensors. This system serves multiple functions, including the assessment of the TUG test, as well as the monitoring and evaluation of walking speed and standing balance. The work introduces an automated camera-based device for monitoring and assessing walking speed, standing balance, and the 5-Times Sit-to-Stand (5TSTS) test. The data collected can be used to evaluate the physical performance of elderly individuals undergoing cancer treatment [29]. This paper makes two primary contributions:

a. A novel approach to the TUG test action recognition using the MediaPipe Pose architecture and ensemble learning model.
b. A new dataset was generated by utilizing videos from six participants, each of whom performed six distinct types of actions, including the stand-to-sit, walking in, turning, walking out, turning-around, and sit-to-stand phases. The videos were tagged and processed under the standards of benchmark datasets.

II. Materials and Method

A. General Context

The Health Research Ethics Committee of the Health Polytechnic, Ministry of Health, Semarang, Indonesia approved this study. The present work developed an ensemble machine learning approach that employed Hist Gradient Boosting, XGBoost, CATBoost, Random Forest, Gradient Boosting, and Stacking Ensemble models to estimate the subtasks of TUG test activities. This approach is illustrated in Figure 1, which presents a systematic method for assessing fall risk through the TUG test by integrating computer vision and machine learning techniques. The data collection phase involved high-resolution 1080p video recordings documenting participants' movements during the TUG exam. These recordings captured key movements, including standing, walking, turning, and sitting, which are critical for evaluating a subject's mobility and potential fall risk.

In the next phase, MediaPipe Pose Estimation, a component of the MediaPipe library, was used to analyze the recorded videos by identifying key human body points in two-dimensional space for each frame. These key points correspond to various joints and anatomical landmarks, and their movement patterns are crucial for assessing the subject's physical performance. The identified key points from each frame were aggregated into a 2D Keypoints Dataset and stored in CSV format for further data manipulation and machine learning model training. After generating the dataset, it was divided into training and testing subsets, with 70% designated for training and 30% for testing. Labels representing various activities were encoded to organize the dataset for machine learning applications. This balanced partitioning ensures that the model can generalize effectively to new data while minimizing the risk of overfitting.

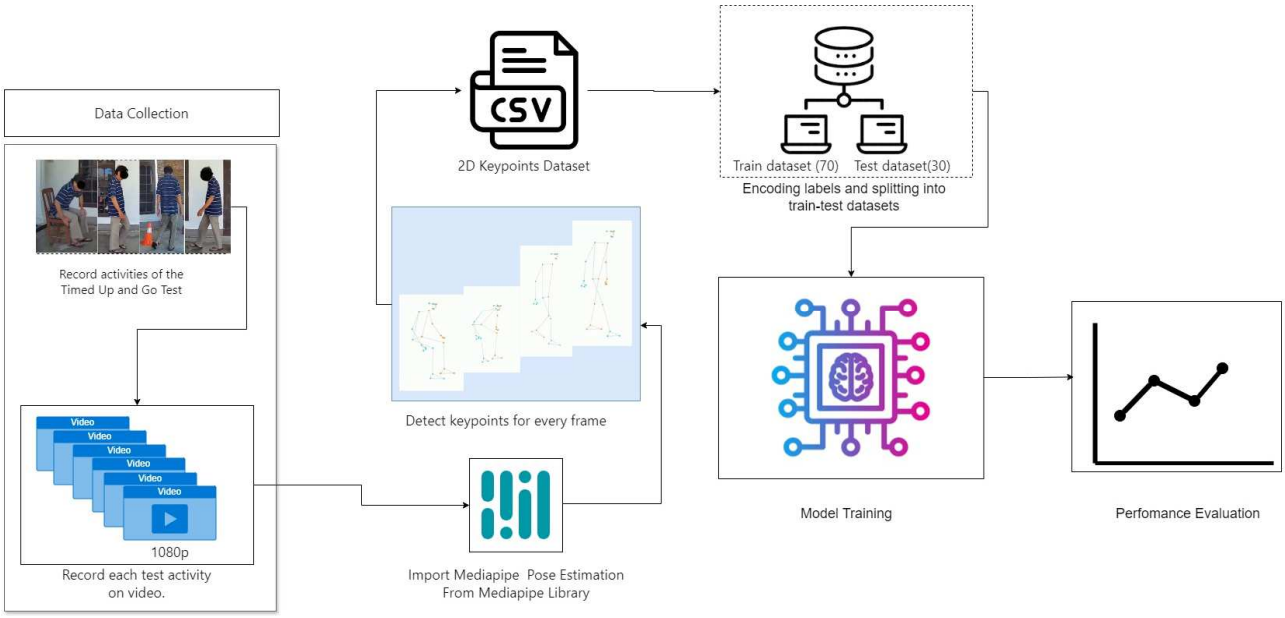

Fig. 1 The method that is being proposed for the model.

The model training phase involves inputting the training data into a machine learning algorithm, aiming to identify patterns in the subject's actions that may signify an elevated risk of falling. With time, the model acquires the ability to categorize various activities and evaluate fall risk based on the trajectory and configuration of keypoints. The performance evaluation phase assesses the model's efficacy. This phase entails utilizing the trained model on the test dataset and evaluating its accuracy, precision, and recall, among other metrics, to verify the TUG test's classification.

The experiments in this study were conducted in a room measuring 6 meters in length and 6 meters in width. The trial had six able-bodied participants, two males and four females, who had no documented mobility limitations. The participants' ages ranged from 17 to 75. Figure 2 illustrates the setup for capturing TUGT video footage.

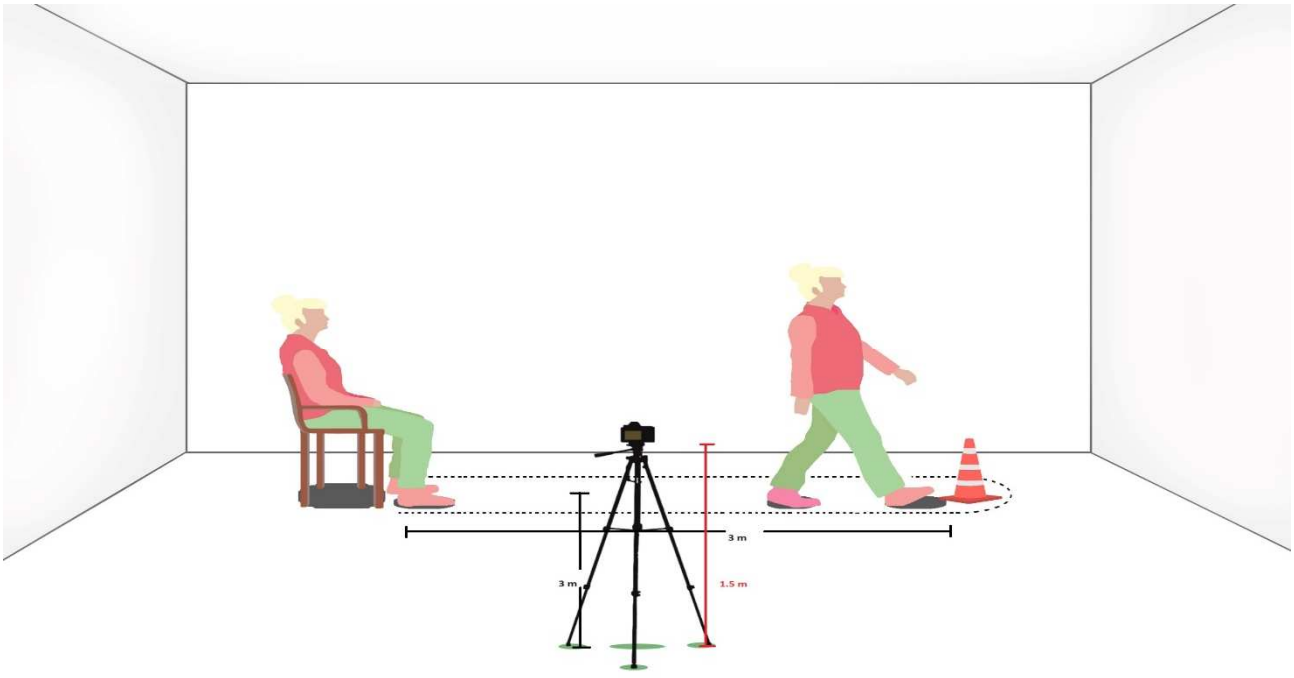

Fig. 2 Illustration of the space utilized for the TUG Test

The chair and cone were positioned 3 meters apart, following the specifications of the 3-meter TUG test, a standard balance assessment. The camera was mounted on a tripod at a height of 1.5 meters above the ground. It was placed laterally to the participant, with a distance of 3 meters between the camera and the track. It was assumed that any object moving along the track would remain within the camera's field of view. The TUG test comprises six activities, categorized according to Hsieh et al. [30]. The subject begins seated in a chair and, upon receiving the "go" signal, performs the SIT_TO_STAND activity, transitioning from seated to standing. The next activity, WALKING_OUT, involves the

participant advancing towards the cone. The TURNING activity requires the subject to navigate around the cone, while the subsequent WALKING_IN activity involves walking back towards the chair. Upon reaching the chair, the participant performs the TURNING_AROUND maneuver. The final action, termed STAND_TO_SIT, involves transitioning from a standing posture back to a seated position in the chair. The video recordings of the TUG test activities varied in duration, starting with the initiation of the SIT_TO_STAND phase and ending with the completion of the STAND_TO_SIT phase. Each video was recorded at a resolution of 1080p and a frame rate of 30 frames per second.

B. TUGT Activity Feature Extraction

The Camera application on Windows 11 is compatible with the JETE 1080P Webcam, which was utilized to take video of the TUG test activities. This webcam delivers 1080p HD video resolution at a frame rate of 30 fps, rendering it appropriate for detailed motion capture. The standards include centered and wide-angle coverage, ensuring a clear and comprehensive view of the subject during balance assessments, which is essential for accurately capturing the subtle movements required for precise evaluation of the TUG test. The JETE 1080P camera is equipped with low-light capabilities, allowing for consistent video quality in varying lighting conditions. This feature is critical for maintaining the integrity of video data across multiple sessions.

Video recordings of each test activity were extracted using the Mediapipe framework. Mediapipe is an adaptable framework that combines open-source technology to create pipelines for processing perceptual data, such as audio, video, and images. Mediapipe offers machine-learning-powered solutions such as hand gestures, face detection, hand tracking, iris tracking, body pose tracking, and other functionalities [28]. We applied the MediaPipe posture estimation method to each frame of the video to segment the TUG test activities and assign labels to the initial locations. The parameters used were min_detection_confidence = 0.5 and model_complexity = 2. This study involves the extraction of two-dimensional (x and y) data from each video frame. The objective is to generate 33 skeleton points, each corresponding to 33 coordinates (x and y). Each skeleton point is assigned two unique identifiers when stored in the CSV file used for model training. This study employed the same six classes of TUG test sub-tasks as those proposed by Hsieh *et al.* [30]. The Mediapipe framework is used to estimate poses in a TUG test video, as shown in Figure 3.

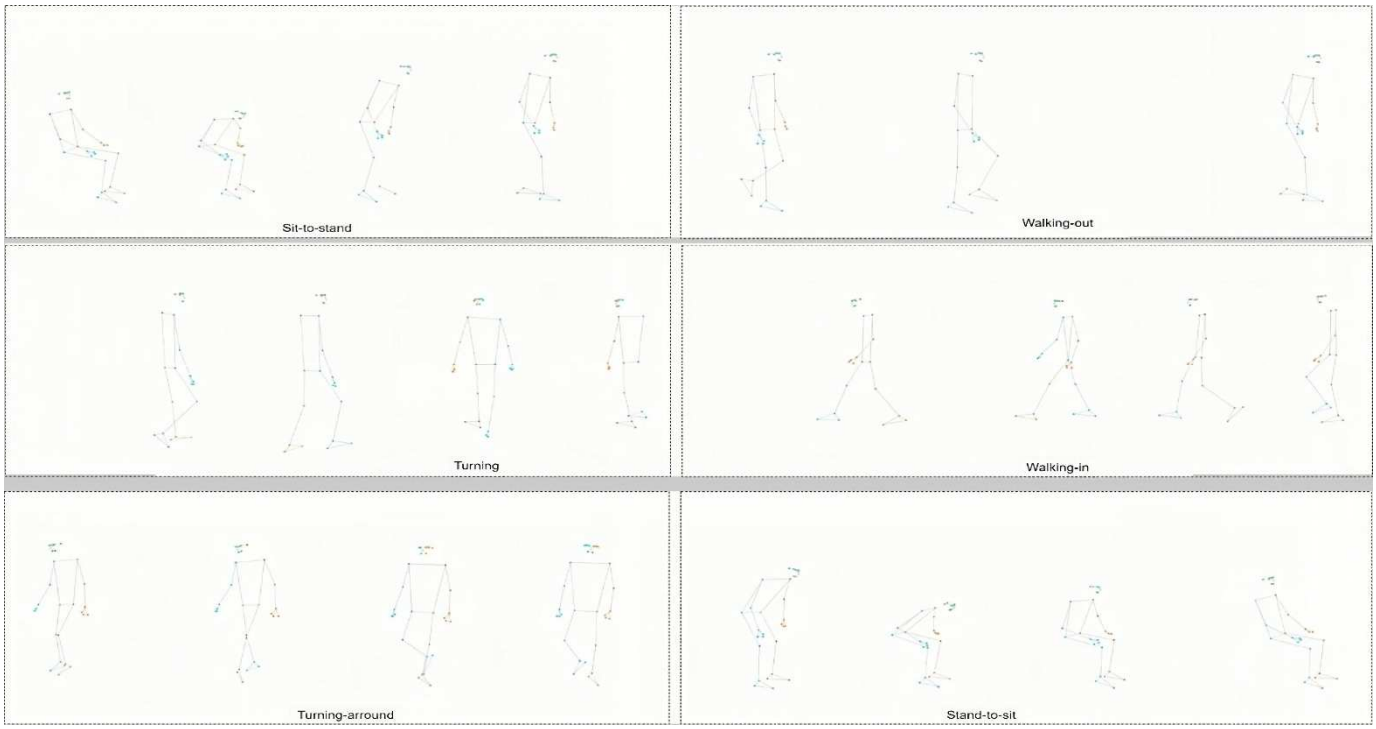

Fig. 3 Examples of pose estimation subtask TUG tests using Mediapipe

As a result, we collected a total of 2365 frames from the dataset of six activity classes and six participants. This research employed an intra-person methodology utilizing the 5-fold cross-validation technique. The data from each participant was partitioned into many folds, with each fold sequentially serving as test data while the remaining folds were utilized for training. This guarantees that the model is both trained and evaluated using data from the same individual, to enhance the prediction of fall risk for each participant based on their prior data. To train and evaluate the samples, the dataset is separated into training (70% of samples) and testing (30% of samples). After selecting the frame videos with well-matched key points, they were input into machine learning to train the TUG test activity key point detection model. Finally, the key point label results of each accurate TUG test activity are used for further processing. All experiments were implemented on a workstation with an Intel® Core™ i7-12700H central processing unit, 16 GB of RAM, and an NVIDIA GeForce RTX 3050 GPU on a Windows 11 64-bit operating system. The experiment was conducted using Python as the programming language and Anaconda 3.0 as the software development environment.

Figure 4 illustrates the distribution of different activities performed by six subjects, designated as Subject_A to Subject_F. The activities include SIT_TO_STAND, STAND_TO_SIT, TURNING, TURNING_AROUND, WALKING_IN, and WALKING_OUT, with the y-axis representing the frequency of each activity. Each individual demonstrates unique patterns in activity frequency, highlighting inter-subject heterogeneity. For Subject_A, the predominant activities are TURNING and STAND_TO_SIT, each performed approximately 90-100 times, while the least frequent activity is SIT_TO_STAND, with fewer than 30 occurrences. This pattern suggests that Subject_A frequently engages in dynamic activities, such as turning or transitioning between postures, rather than rising from a seated position. Similarly, Subject_B exhibits a comparable pattern, with TURNING being the most frequent activity and SIT_TO_STAND the least. The consistency observed in both subjects indicates that turning and postural adjustments may play a significant role in their daily routines.

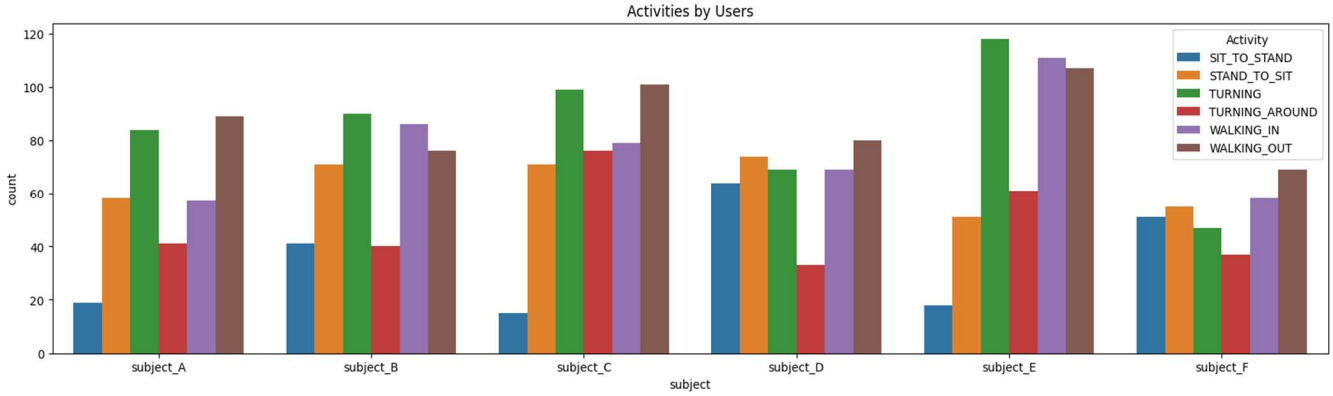

Fig. 4 The quantity of frames allocated for each participant's sub-task activity.

Subjects C and D display somewhat different distributions. For Subject_C, TURNING remains the most frequent activity, while WALKING_IN and WALKING_OUT are also notably represented. Subject_D shows a more balanced distribution of activities, with TURNING_AROUND and WALKING_OUT occurring more frequently than SIT_TO_STAND, which remains below 20 occurrences. These variations highlight the distinct movement patterns of each individual, potentially influenced by their daily routines or physical habits. For Subjects E and F, TURNING and WALKING_OUT are the predominant activities, with Subject_E demonstrating the highest frequency of TURNING among all subjects. Subject_F also regularly engages in these activities, albeit at slightly lower frequencies. In both cases, SIT_TO_STAND remains consistently low, indicating that transitions from sitting to standing are less frequent for these individuals compared to more active behaviors such as walking and turning.

C. Performance Metric

To evaluate the findings of the study, we employed four widely accepted performance metrics: accuracy, F1-score, precision, and recall. These evaluation metrics are computed using the following definitions: TP represents the number of true positive samples correctly identified in the testing set, TN represents the number of true negative samples correctly identified in the testing set, FP represents the number of false positive samples incorrectly identified in the testing set, and FN represents the number of false negative samples mistakenly identified in the testing set.

The accuracy metric measures the proportion of correctly identified samples in the testing set out of the total number of data samples. Precision measures the ratio of correctly identified positive samples in the testing set to the total number of both false positives (FP) and true positives (TP). Recall is calculated by dividing the number of true positive (TP) instances in the testing dataset by the sum of TP and false negative (FN) instances. The F1-score, which provides a balanced measure of precision and recall, can be calculated using Equation 4, based on the precision and recall values. [31].

$$Accuracy = \frac{(TP + TN)}{(TP+FP+TN+FN)} \quad (1)$$

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2)$$

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

$$F1-score = 2\,x\,\frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (4)$$

III. Results and Discussion

The variation in the number of frames for each participant's TUG test sub-task is attributed to individual differences in the time taken to complete each task. Figure 4 presents the number of frames associated with each participant's sub-task activity. The bar chart illustrates the frequency of six distinct actions performed by six individuals, identified as Subjects A to F. The activities are color-coded as follows: Sit-to-stand (blue), Stand-to-sit (orange), Turning (green), Turning-around (red), Walking-in (purple), and Walking-out (brown). For Subject A, the predominant activities are Walking-in and Turning-around, each occurring approximately 80 to 90 times, followed by Walking-out and Turning, which occur around 70 times each. The Stand-to-sit action occurs 60 times, while Sit-to-stand happens around 20 times.

Subject B shows Walking-in and Walking-out as the most frequent behaviors, occurring more than 80 times. Turning and Turning-around occur between 70 and 75 times. The Stand-to-sit action occurs 50 times, whereas Sit-to-stand occurs fewer than 20 times. Subject C primarily engages in

Walking-out, which occurs 90 times. Turning and Walking-in are the next most frequent activities, each occurring 80 times. Turning-around occurs 70 times, and Stand-to-sit occurs 60 times. The Sit-to-stand activity is recorded fewer than 20 times. Subject D's activity pattern is generally consistent, with the exception of the Sit-to-stand action, which occurs 20 times. The frequency of other activities ranges from 60 to 80, with Turning and Walking-out being the most predominant.

Subject E exhibits a high frequency of walking-in, with over 100 counts, and walking-out, with over 90 counts, as the most common activities. The activities of turning-around and turning have roughly 80 counts each. Stand-to-sit has an approximate count of 60, while sit-to-stand has about 20 counts. Subject F exhibits the highest number of occurrences in the Walking-out category, with approximately 90 instances, and in the Walking-in category, with around 70 instances. This is followed by Turning-around with 60 instances, Turning with about 50 instances, Stand-to-sit with around 40 instances, and Sit-to-stand with about 20 instances. The chart indicates that Walking-in and Walking-out are the most frequently performed activities across all subjects, whereas Sit-to-stand is the least frequent. This provides a clear understanding of the distribution and frequency of various activities undertaken by each individual.

Figures 5–10 display confusion matrices used to evaluate the TUG test sub-task activity classification model based on an ensemble learning method. The matrices consist of six activities: Sit-to-stand, Walking-out, Turning, Walking-in, Turning-around, and Stand-to-sit. The matrices are color-coded in shades ranging from deep blue to pale blue, corresponding to different levels of predictions, with the color scale indicated on the right side of the diagrams. Each element in the matrix represents the number of predictions relative to the true labels.

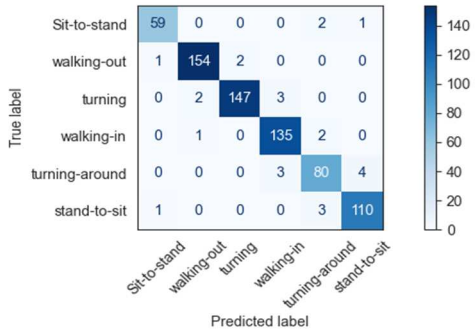

Fig. 5 Confusion Matrix of Hist Gradient Boosting.

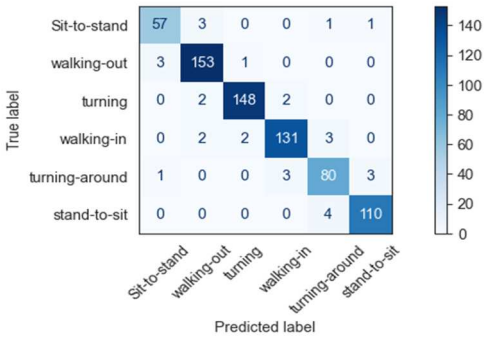

Fig. 6 Confusion Matrix of Extreme Gradient Boosting.

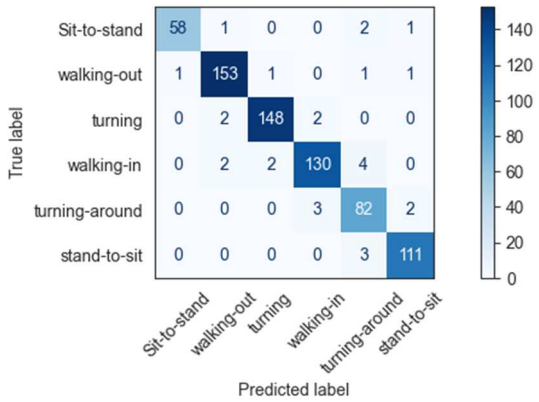

Fig. 7 Confusion Matrix of CATBoost.

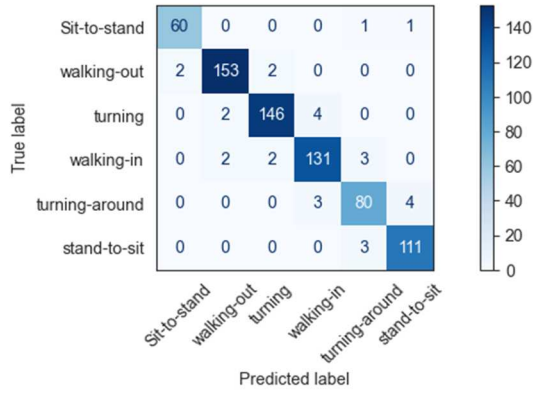

Fig. 8 Confusion Matrix of Random Forest

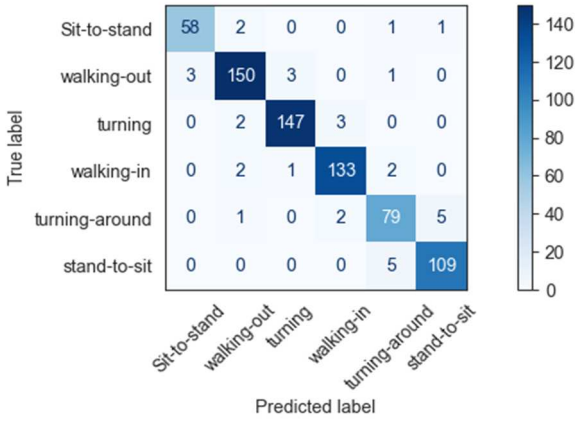

Fig. 9 Confusion Matrix of Gradient Boosting.

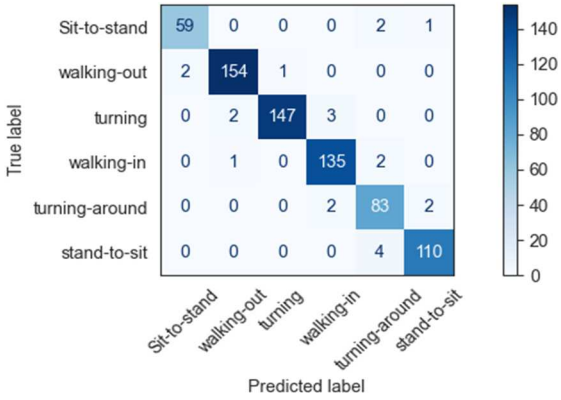

Fig. 10 Confusion Matrix of Stacking Ensemble

Figure 5 presents the confusion matrix of the Hist Gradient Boosting model. The classifiers demonstrate varying degrees of accuracy for different sub-activities. For instance, the 'Walking-out' activity is classified with 100% accuracy, achieving 154 correct predictions out of 154 cases. This indicates the classifier's high proficiency in recognizing this particular activity. Similarly, the 'Turning' and 'Walking-in' activities also exhibit high accuracy, with 147 and 135 correct classifications, respectively. Conversely, the activities 'Sit-to-stand' and 'Turning-around' show lower classification accuracy, with 59 and 80 correct predictions, respectively, suggesting that the classifier has more difficulty accurately identifying these activities. This challenge may be attributed to the similarity of motion patterns in certain activities, leading to misclassifications.

The non-diagonal elements provide insight into specific instances of misclassification. For example, the action known as 'Sit-to-Stand' is occasionally misclassified as 'Turning-Around' in two cases and as 'Stand-to-Sit' in one case. Similarly, the action of 'Turning' is frequently misclassified as 'Walking-In' in three instances, while 'Turning-Around' is misclassified as 'Stand-to-Sit' in four instances. These misclassifications suggest potential avenues for improving the model, possibly through implementing more advanced feature extraction techniques or fine-tuning the model.

Figure 10 illustrates that the stacking ensemble model accurately predicted the walking-out behavior in 154 instances; however, it did make a few errors, including misclassifying two instances of walking-out as sit-to-stand.

Table 1 compares the performance of six ensemble machine learning models: Hist Gradient Boosting, XGBoost, CATBoost, Random Forest, Gradient Boosting, and Stacking Ensemble. These models were evaluated based on their ability to classify different sub-activities, namely sit-to-stand, walking-out, turning, walking-in, turning-around, and stand-to-sit. Each sub-activity was evaluated using metrics such as Precision, Recall, and F1-score. The highest F1-score for Hist Gradient Boosting was found in the turning activity, with a value of 97.67%. The walking-out and sit-to-stand activities exhibited F1-scores of 98.09% and 95.93%, respectively, resulting in a total accuracy of 96.48% for the model. The XGBoost model demonstrated superior performance in turning, achieving a Precision of 98.01% and an F1-score of 97.69%, leading to an overall accuracy of 95.63%. CATBoost achieved the best overall accuracy of 96.06%. Among the various activities, turning had the highest Precision of 98.01%, while walking-out received an F1-score of 97.14%.

TABLE I
PERFORMANCE ANALYSIS OF DIFFERENT ALGORITHM IN CLASSIFICATION

	Hist gradient boosting			**XGBoost**			**CATBoost**		
Sub-Activity	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
					(%)				
Sit_to_Stand	96.72	95.16	95.93	93.44	91.94	92.68	98.31	93.55	95.87
Walking_out	98.09	98.09	98.09	95.62	97.45	96.53	96.84	97.45	97.14
Turning	98.66	96.71	97.67	98.01	97.37	97.69	98.01	97.37	97.69
Walking_in	95.74	97.83	96.77	96.32	94.93	95.62	96.30	94.20	95.24
Turning_arround	91.95	91.95	91.95	90.91	91.95	91.43	89.13	94.25	91.62
Stand_to_sit	95.65	96.49	96.07	96.49	96.49	96.49	96.52	97.37	96.94
Overall Accuracy			96.48			95.63			96.06
	Random Forest			**Gradient Boosting**			**Stacking Ensemble**		
Sub-Activity	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
					(%)				
Sit_to_Stand	96.77	96.77	96.77	95.08	93.55	94.31	96.72	95.16	95.93
Walking_out	97.45	97.45	97.45	95.54	95.54	95.54	98.09	98.09	98.09
Turning	97.33	96.05	96.69	97.35	96.71	97.03	99.32	96.71	98.00
Walking_in	94.93	94.93	94.93	96.38	96.38	96.38	96.43	97.83	97.12
Turning_arround	91.95	91.95	91.95	89.77	90.80	90.29	91.21	95.40	93.26
Stand_to_sit	95.69	97.37	96.52	94.78	95.61	95.20	97.35	96.49	96.92
Overall Accuracy			95.92			95.21			96.90

The Random Forest model, positioned in the lower section of the table, exhibited an overall accuracy of 95.92%, with the highest F1-score of 97.45% for the walking-out category. The Gradient Boosting and Stacking Ensemble models demonstrated overall accuracies of 95.21% and 96.90%, respectively. Among these models, the Stacking Ensemble exhibited the highest overall accuracy, particularly excelling in the walking-out activity with an F1-score of 98.09% and in the turning activity with an F1-score of 98.00%. This thorough comparison examines the advantages and disadvantages of each model, revealing that CATBoost and Hist Gradient Boosting generally achieve a good balance between accuracy and performance. However, the Stacking Ensemble model outperforms the others in certain TUG test activities, demonstrating higher overall performance.

Each study presents a distinct method for evaluating physical mobility through the Timed Up and Go (TUG) test. Your research achieves high accuracy using video-based pose estimation, suitable for non-invasive environments. In contrast, the IMU-based research is particularly relevant in clinical settings. The camera-based system enables real-time monitoring for cancer patients, while the Kinect-based research integrates machine learning with fall risk assessment, providing an economical home-use solution. Collectively, these studies highlight the adaptability of the TUG test across various demographics and technological contexts. A comparison was conducted between various state-of-the-art approaches for segmenting the subtasks of TUG tests and the proposed method, with findings reported in Table II.

TABLE III
COMPARISON OF METHODS FOR SEGMENTING TUG SUBTASKS AND THE PROPOSED APPROACH

Parameter	Reference			
	This work	Hsieh et al. [30]	Duncan et al. [32]	Dubois et al. [24]
Technology	Video-based (MediaPipe)	IMUs (accelerometers)	Multi-camera (Raspberry Pi)	Depth sensor (Kinect)
Data Modality	Video pose data	Motion data (IMU)	Camera-based video	Depth data (Kinect)
Participants	6 subjects	26 subjects	8 subjects	43 subjects
Machine Learning Models	Stacking, XGBoost, Random Forest	AdaBoost, Support Vector Machine	CSRT (Channel Spatial Reliability Tracking)	SVM, RF, Neural Netwotk , Naive Bayes
Tested Population	General subtasks, low risk	TKA patients	Older adults with cancer	Elderly individuals (fall risk)
Accuracy	96.90% (Stacking Ensemble)	92% (AdaBoost)	>95% (gait speed), >97% (timing)	100% (SVM, RF with two parameters)

IV. CONCLUSIONS

We present a fully automated segmentation technique for the subtasks of the Timed Up and Go (TUG) test in video recordings. Our method employs a human learning-based ensemble machine learning methodology for pose estimation, making it significantly more practical to adopt than previous systems. Among the models studied, the Stacking Ensemble approach achieved the highest overall accuracy of 96.90%, surpassing the performance of other algorithms such as Hist Gradient Boosting and CATBoost, both of which also demonstrated commendable precision and F1-scores. Although XGBoost is robust, it exhibited marginally inferior precision and recall in the majority of subtasks compared to the leading methodologies. Despite Random Forest and Gradient Boosting displaying competitive efficacy, they failed to surpass the performance of the Stacking Ensemble. While the efficiency enhancements of the Stacking Ensemble method are significant, particularly in practical applications, the increase in accuracy relative to simpler techniques like Hist Gradient Boosting may appear minimal. We argue that this modest enhancement justifies the added complexity when considering the broader context of fall-risk screening, where even small improvements in accuracy can yield substantial clinical benefits. However, the computational complexity of ensemble methods remains a potential limitation that requires careful consideration. In real-world applications, evaluating the trade-offs between model complexity and performance improvements is crucial, particularly in resource-limited settings.

In the future, researchers intend to explore techniques to reduce computational costs while maintaining accuracy, thereby enhancing the method's accessibility in clinical environments. Furthermore, utilizing multimodal sensor data could improve the method's efficacy, providing a more comprehensive solution for early fall-risk assessment by healthcare practitioners, including physicians and physiotherapists.

ACKNOWLEDGMENT

The research was funded by Universitas Muhammadiyah Yogyakarta, Indonesia, through the Center for Research and Innovation for the year 2023.

REFERENCES

[1] World Health Organization, "Falls." Accessed: Jul. 30, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/falls

[2] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, pp. 144–152, 2013, doi: 10.1016/j.neucom.2011.09.037.

[3] S. Usmani, A. Saboor, M. Haris, M. A. Khan, and H. Park, "Latest research trends in fall detection and prevention using machine learning: A systematic review," *Sensors*, vol. 21, no. 15, pp. 1–23, 2021, doi:10.3390/s21155134.

[4] A. Z. Rakhman, Kurnianingsih, L. E. Nugroho, and Widyawan, "U-FASt: Ubiquitous fall detection and alert system for elderly people in smart home environment," *Proceeding - 2014 Makassar Int. Conf. Electr. Eng. Informatics, MICEEI 2014*, no. November, pp. 136–140, 2014, doi: 10.1109/miceei.2014.7067326.

[5] T. Althobaiti, S. Katsigiannis, and N. Ramzan, "Triaxial accelerometer-based falls and activities of daily life detection using machine learning," *Sensors (Switzerland)*, vol. 20, no. 13, pp. 1–19, 2020, doi: 10.3390/s20133777.

[6] M. J. Al Nahian *et al.*, "Towards an Accelerometer-Based Elderly Fall Detection System Using Cross-Disciplinary Time Series Features," *IEEE Access*, vol. 9, pp. 39413–39431, 2021, doi:10.1109/access.2021.3056441.

[7] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "Up-fall detection dataset: A multimodal approach," *Sensors (Switzerland)*, vol. 19, no. 9, 2019, doi:10.3390/s19091988.

[8] E. Casilari, J. A. Santoyo-Ramón, and J. M. Cano-García, "UMAFall: A Multisensor Dataset for the Research on Automatic Fall Detection," *Procedia Comput. Sci.*, vol. 110, pp. 32–39, 2017, doi:10.1016/j.procs.2017.06.110.

[9] D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, vol. 7, no. 10, 2017, doi:10.3390/app7101101.

[10] Kurnianingsih, L. E. Nugroho, Widyawan, L. Lazuardi, A. S. Prabuwono, and M. Pratama, "Anomaly detection for elderly home care," *Int. J. Bus. Intell. Data Min.*, vol. 16, no. 4, pp. 418–444, 2020, doi: 10.1504/IJBIDM.2020.107545.

[11] N. Lundebjerg, "Guideline for the prevention of falls in older persons," *J. Am. Geriatr. Soc.*, vol. 49, no. 5, pp. 664–672, 2001, doi:10.1046/j.1532-5415.2001.49115.x.

[12] C. Tunca, G. Salur, and C. Ersoy, "Deep Learning for Fall Risk Assessment with Inertial Sensors: Utilizing Domain Knowledge in Spatiooral Gait Parameters," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 7, pp. 1994–2005, 2020, doi: 10.1109/JBHI.2019.2958879.

[13] S. P. Teo, "Fall risk assessment tools-validity considerations and a recommended approach," *Ital. J. Med.*, vol. 13, no. 4, pp. 202–204, 2019, doi: 10.4081/itjm.2019.1196.

[14] S. L. S. Koh, N. Hafizah, J. Y. Lee, Y. L. Loo, and R. Muthu, "Impact of a fall prevention programme in acute hospital settings in Singapore," *Singapore Med. J.*, vol. 50, no. 4, pp. 425–432, 2009.

[15] V. Scott, K. Votova, A. Scanlan, and J. Close, "Multifactorial and functional mobility assessment tools for fall risk among older adults in community, home-support, long-term and acute care settings," *Age Ageing*, vol. 36, no. 2, pp. 130–139, 2007, doi: 10.1093/ageing/afl165.

[16] V. Strini, R. Schiavolin, and A. Prendin, "Fall Risk Assessment Scales: A Systematic Literature Review," *Nurs. Reports*, vol. 11, no. 2, pp. 430–443, 2021, doi: 10.3390/nursrep11020041.

[17] P. Richardson, "the Timed & Go," *Jags*, vol. 39, no. 2, pp. 142–148, 1991.

[18] J. C. Whitney, S. R. Lord, and J. C. T. Close, "Streamlining assessment and intervention in a falls clinic using the Timed Up and Go Test and Physiological Profile Assessments," *Age Ageing*, vol. 34, no. 6, pp. 567–571, 2005, doi: 10.1093/ageing/afi178.

[19] E. Eckstrom *et al.*, "American Geriatrics Society response to the World

Falls Guidelines," *J. Am. Geriatr. Soc.*, vol. 72, no. 6, pp. 1669–1686, Jun. 2024, doi: https://doi.org/10.1111/jgs.18734.

[20] W. M. A. Meekes, J. C. Korevaar, C. J. Leemrijse, and I. A. M. Van De Goor, "Practical and validated tool to assess falls risk in the primary care setting: A systematic review," *BMJ Open*, vol. 11, no. 9, pp. 1–10, 2021, doi: 10.1136/bmjopen-2020-045431.

[21] N. Eichler, S. Raz, A. Toledano-Shubi, D. Livne, I. Shimshoni, and H. Hel-Or, "Automatic and Efficient Fall Risk Assessment Based on Machine Learning," *Sensors*, vol. 22, no. 4, pp. 40–48, 2022, doi:10.3390/s22041557.

[22] M. Kampel, S. Doppelbauer, and R. Planinc, "Automated timed up & go test for functional decline assessment of older adults," *PervasiveHealth Pervasive Comput. Technol. Healthc.*, no. July, 2018, doi: 10.1145/3240925.3240960.

[23] G. Sprint, D. J. Cook, and D. L. Weeks, "Toward Automating Clinical Assessments: A Survey of the Timed Up and Go," *IEEE Rev. Biomed. Eng.*, vol. 8, pp. 64–77, 2015, doi: 10.1109/RBME.2015.2390646.

[24] A. Dubois, T. Bihl, and J. P. Bresciani, "Automating the timed up and go test using a depth camera," *Sensors (Switzerland)*, vol. 18, no. 1, pp. 1–13, 2018, doi: 10.3390/s18010014.

[25] P. Savoie, J. A. D. Cameron, M. E. Kaye, and E. J. Scheme, "Automation of the Timed-Up-and-Go Test Using a Conventional Video Camera," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 4, pp. 1196–1205, 2020, doi: 10.1109/JBHI.2019.2934342.

[26] T. Kamnardsiri *et al.*, "Conventional video-based system for measuring the subtask speed of the Timed Up and Go Test in older adults: Validity and reliability study," *PLoS One*, vol. 18, no. 6 June, pp. 1–20, 2023, doi: 10.1371/journal.pone.0286574.

[27] S. Fudickar, S. Hellmers, S. Lau, R. Diekmann, J. M. Bauer, and A. Hein, "Measurement system for unsupervised standardized assessment of timed 'up & go' and five times sit to stand test in the community—a validity study," *Sensors (Switzerland)*, vol. 20, no. 10, 2020, doi:10.3390/s20102824.

[28] S. Fudickar *et al.*, "Measurement System for Unsupervised Standardized Assessments of Timed Up and Go Test and 5 Times Chair Rise Test in Community Settings—A Usability Study," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22030731.

[29] L. Duncan *et al.*, "Camera-Based Short Physical Performance Battery and Timed Up and Go Assessment for Older Adults with Cancer," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 9, pp. 2529–2539, 2023, doi:10.1109/tbme.2023.3253061.

[30] C. Y. Hsieh, H. Y. Huang, K. C. Liu, K. H. Chen, S. J. P. Hsu, and C. T. Chan, "Subtask segmentation of timed up and go test for mobility assessment of perioperative total knee arthroplasty†," *Sensors (Switzerland)*, vol. 20, no. 21, pp. 1–17, 2020, doi: 10.3390/s20216302.

[31] A. Gumaei *et al.*, "A deep learning-based driver distraction identification framework over edge cloud," *Neural Comput. Appl.*, vol. 1, 2020, doi: 10.1007/s00521-020-05328-1.

[32] L. Duncan *et al.*, "Camera-Based Short Physical Performance Battery and Timed Up and Go Assessment for Older Adults with Cancer," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 9, pp. 2529–2539, 2023, doi:10.1109/tbme.2023.3253061.