JURNAL INFOTEL
VOL. 17, NO. 3, AUGUST 2025, PP. 456–471. DOI:10.20895/INFOTEL.V17I3.1420

RESEARCH ARTICLE

TelUP Human Fall Dataset: A Motion Forecasting Study of Human Falls

Agung Widiyanto[1], Raphon Galuh Candraningtyas[2], Andi Hisyam Helmi F.F[3], Mayesq Prameswari[4], Himam Bashiran[5], Geugeut Nyarikawanti Surahmat[6], Balqis Awaluna Rahmah[7], A.A. Istri Candra Manika Dewi[8], and Andi Prademon Yunus[9,*]

[1,2,3,4,5,6,7,8]Study Program in Data Science, Telkom University, Purwokerto, 53114, Indonesia
[9]Department of Informatics, Telkom University, Purwokerto, 53114, Indonesia

*Corresponding author: andiay@telkomuniversity.ac.id

Received: July 25, 2025; Revised: August 20, 2025; Accepted: August 25, 2025.

Abstract: This study investigates multitask learning approaches for human motion forecasting and fall classification using pose data extracted from video sequences. A custom dataset, the TelUP HumanFall Forecasting Dataset, was developed, containing annotated video frames representing fall and non-fall scenarios captured from six participants. Pose information was extracted using YOLOv11, producing 17 keypoints per frame, which were normalized and segmented into temporal sequences for training. Three deep learning architectures, Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM), were implemented and evaluated. The models were assessed in a subject-independent test set consisting of two participants to ensure generalization. Quantitative evaluation measured the forecast error using the mean per joint position error (MPJPE) and classification accuracy. The MLP achieved the lowest MPJPE of 0.2630 (131.5 pixels), while the LSTM obtained the highest classification accuracy of 92.89%. Qualitative analysis revealed limitations in the capture of complex joint dynamics. Despite fast training convergence, the results emphasize a trade-off between forecast precision and classification accuracy. Future work will explore more expressive architectures and improved pose extraction methods to enhance forecast realism.

Keywords: Deep Neural Networks, Fall Detection, Human Motion Forecasting, Multi-task Learning, Pose Estimation

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Falls are a common but often underestimated occurrence in daily life, with most people experiencing unintentional falls at least once in their lifetime. Although falls may seem trivial, they pose significant risks, particularly among older adults. According to the World Health Organization (WHO), falls are the second leading cause of unintentional injury death worldwide, where 68,400 individual deaths are estimated annually, where adults older than 60 years of age suffer the highest number of fatal falls [1]. In recent years, studies have shown that the elderly population is growing faster than any other age group, and by October 2022, 10% percents of the world's total population will be older than 65 years [2]. The prediction based on related studies suggests that by the end of 2050, there will be 1.5 billion older adults worldwide [3]. As humans get older, their physical, cognitive, and motor skills will decline with age. Hence, falls become a significant challenge for them and can significantly reduce life expectancy as older adults [4]. Approximately 35% of adults 65 years and older experience at least one fall per year [5]. In addition to age, other factors that can contribute fall varies, such as physical activity, the environment, and cardiovascular disease. It is a major cause of physical harm, which typically requires hospitalization [6], according to WHO 37.3 million fall accident require medical attention, while 650,000 falls result in death yearly [1].

Given these alarming statistics, proactive measures to mitigate falls-related injuries are urgently needed. Research has been done to detect fall so that we can differentiate fall action with other daily activity living (ADLs) such as utilizing sensor-based system like accelerometer and gyroscope with machine learning method and algorithm for deciding the states of fall or using image-based fall detection system to detect fall while doing daily activities using deep learning methods. Where both methods are highly dependent on the accuracy of human pose estimation that measures the sensitivity of the system to differenitiate fall state with other state [7]. Although existing research has focused predominantly on fall detection, merely identifying falls after they occur offers limited preventive value. A more impactful approach lies in predicting falls before they occur, enabling timely interventions to prevent injuries altogether. By predicting human motion, it makes biometric identification possible and offers a safe way to authenticate individuals. Furthermore, the ability to anticipate human motion makes it easier to create realistic gestures, which can be useful in various applications, including virtual reality (VR), animation, and human-robot interaction applications [7]. In addition, in the healthcare field, the use of human motion prediction can be used to design personalized rehabilitation programs tailored to patients with movement disorders, which ultimately improves their motor skills and overall quality of life [8]. Using technology to efficiently evaluate and forecast human movements, motion prediction models can be used in healthcare to improve patient care, treatment results, and rehabilitation procedures [9].

Although significant progress has been made in fall detection systems, accurately forecasting falls before they occur presents several fundamental challenges. The first indicators of imminent falls, subtle changes in balance, slight stumbles, or minor gait disturbances are significantly harder to detect than the dramatic motions characteristic of actual falls. Current datasets further complicate this challenge, as most focus exclusively on post-fall detection and are collected in controlled laboratory environments that do not reflect real-world conditions. Our preliminary experiments clearly revealed these limitations: When testing conventional image classification approaches in which data was resized, split into

individual frames, and shuffled, the models struggled to understand the crucial temporal relationships needed for accurate forecasting. This frame-by-frame processing, while effective for detection tasks, destroys the sequential motion patterns essential for prediction. The current research landscape reveals important gaps that our work addresses. Most existing studies concentrate solely on fall detection after the fact, providing alerts only after a fall has already occurred. Furthermore, available datasets suffer from notable limitations, and they often feature scripted falls performed by young, healthy participants rather than the elderly population most at risk, and labeling tends to be subjective. In our initial dataset of six subjects performing various fall directions along with non-fall activities, we observed how these constraints affect model performance and generalizability. To overcome these challenges, we propose a novel dataset specifically designed for fall motion forecasting. The dataset includes carefully annotated fall and non-fall scenarios across diverse real-world environments, moving beyond the artificial constraints of laboratory settings. We introduce evaluation metrics focused on practical forecasting performance, such as Mean Per Joint Position Error (MPJPE), which focus on practical forecasting performance, provide a more realistic assessment of a model's ability to predict motion trajectories in advance essential aspect for time-sensitive tasks like fall prevention. By addressing both data limitations and methodological gaps in current research, this dataset supports meaningful progress toward truly preventive fall intervention systems.

Despite advances in fall detection, forecasting falls poses unique technical challenges. First, early pre-fall cues such as subtle imbalances or irregular gait patterns are significantly harder to capture than overt falls, requiring high-resolution motion analysis. Second, existing datasets predominantly focus on post-fall detection, lacking annotated pre-fall sequences or non-fall scenarios. Many are also collected in controlled lab environments, limiting their applicability to real-world settings. Third, unlike detection, forecasting demands temporal modeling over extended windows (e.g., >500ms), which existing methods struggle to achieve due to their reliance on short-term features. Current approaches remain overwhelmingly detection-centric. For instance, widely used datasets like URFD [10] and SisFall [11] provide labeled fall events but omit the critical pre-fall phase, rendering them unsuitable for forecasting. In addition, these data sets often suffer from selection biases, such as limited elderly participants or scripted falls performed by healthy adults, which do not represent the diversity of real-world fall scenarios.

To address these gaps, we introduce a new dataset and an experimental study explicitly designed for falling motion forecasting. Our data set captures data from real-world and simulated environments, with annotations for falls and non-fall instability across diverse scenarios. Preliminary testing revealed limitations in earlier versions of our dataset, where frame-level shuffling and resizing for image classification disrupted temporal coherence, complicating model training. To mitigate this, we refined the dataset to preserve sequential integrity and standardized labeling protocols to reduce subjectivity. The current dataset comprises six subjects, each performing fall actions (forward, backward, left, right, sitting, and standing falls) and non-fall activities (jumping, laying, walking, picking object, squatting, and stretching), ensuring a balanced representation of motion dynamics. By offering diverse and temporally annotated data, our dataset supports the development of models that can anticipate falls rather than simply react to them, marking a crucial advancement toward proactive healthcare and assistive technologies.

2 Related Works

2.1 Fall Forecasting System

Research on fall detection systems has advanced significantly, with diverse approaches developed to address this critical healthcare challenge. In healthcare fields, predicting the motion of fall can be helped by implementing preventive measures to improve emergency response and reduce the risk of major repercussions. A notable study designed a smart fall detection system using a fuzzy adaptive threshold algorithm integrated with a smartwatch accelerometer. This approach not only detects falls, but also supports indoor positioning, offering a dual benefit to aging populations while minimizing healthcare costs [12]. Another study proposes a low-cost fall detection system for the elderly using pyroelectric infrared (PIR) sensors, which demonstrates feasibility for elderly monitoring [13].

Beyond wearable devices, IoT-based systems have become versatile tools for predicting falls. One such system combines accelerometer and gyroscope sensors with deep learning to classify falls alongside routine activities, enhancing contextual awareness [14]. Another study uses wearable data of multiple sensors combined with eXplainable AI (XAI) for the detection of interpretable falls, improving the detection accuracy while maintaining the interpretability of the model, which is a crucial feature for clinical adoption [15]. Although sensor-based methods dominate current research, their implementation varies. For example, a study applied a deep learning model to wearable sensor data, achieving robust classification of activity and fall [14]. Another introduced a dual-channel feature integration method with sliding windows, refining fall detection by distinguishing between "falling-state" and "fallen-state" perspectives [16]. To address hardware limitations, a 360-degree camera system was developed, expanding the field of view for more reliable monitoring in complex environments [17].

While these studies demonstrate significant progress in fall forecasting systems, their practical implementation still faces challenges related to real-world variability, user compliance, and system integration. The diversity of sensor-based approaches from wearable devices to ambient IoT systems highlights the need for adaptable solutions tailored to different environments and populations. As research continues to refine these technologies, the focus must now shift toward optimizing their reliability and accessibility for broader adoption. This foundation of fall detection methodologies naturally leads to an examination of human motion analysis, which plays a pivotal role in advancing predictive accuracy, as discussed in the following subsection.

2.2 Human Motion in Human Activity Analysis

Human motion forecasting analyzes body postures and movements from video data to predict future actions, with fall detection being one of its most critical applications. By examining body postures and acceleration patterns in human motion data, Human Fall Motion Prediction aims to prevent injuries [18]. In order to ascertain whether a fall has happened, a study created the Human Torso Motion Model (HTMM), which compares the rates at which the torso angle and centroid height change with predetermined criteria. Comparing this method to other fall detection techniques, it was discovered to be very accurate in differentiating between falls [19]. This approach highlights how motion analysis serves as the foundation for reliable fall prediction systems.

Since self-attention models are thought to be able to capture intricate spatial-temporal connections, they have also been applied, building on the advancements made in motion forecasting by the RNN-based approach. Because these methods can simultaneously address the right time sequences and the right portions of the feature space, motion modeling has improved. For instance, the high-resolution spatial-temporal attention network (HR-STAN) in Zhang et al.'s work improves the model representation of spatial-temporal relations by combining attention with spatial-temporal convolutional networks (STConv) [20]. Such innovations have significantly improved short-term motion prediction, particularly for precise, localized movements. Further progress came with frameworks that integrate predictions across multiple time steps, using a fusion system to provide a prediction framework for the head, upper arm, and lower arm. Prediction frames created in previous time steps are combined with the frames created in the present phase in this framework. This fusion procedure improves the motion's coherence and lowers prediction errors. The spatiotemporal transformer graph convolutional network (STTG-Net) was created as a result of this method, and it not only reduces error accumulation but also makes motion prediction considerably more fluid and reliable [21]. Additionally, to overcome the drawbacks of traditional encoder schemes, peripheral motion had to be encoded in a way that was more suitable for the situation at hand. The context-sensitive motion recognition method used by the STTG-Net systematically focuses attention on previous motions. In this procedure, similar historical subsequences are extracted using a motion-attention model, spatial and temporal correlations are extracted using a Graph Convolutional Network, and historical motion patterns are accurately modeled using a representation based on the Discrete Cosine Transform [9].

The widespread application of various machine learning architectures, including self-attention, has led to advances in the field of predicting human motion or falls. New ideas that employ privacy-preserving sensors and human posture estimation, as well as computer vision-based methods, have also advanced the field by filling in some of the gaps in its successful application. Models based on self-attention have shown promise in modeling spatio-temporal scenes and predicting short-range mobility. These advances demonstrate the potential for developing systems that are more precise, adaptable, and scalable. Future developments must address computational efficiency, dataset variability, and the integration of privacy-conscious sensors to create practical, scalable solutions for fall prevention and other motion analysis applications.

2.3 Challenge in Fall Forecasting

Despite significant advances in fall prediction systems, several key challenges remain in achieving robust real-world performance. First, environmental and behavioral variability complicates accurate detection, as systems must distinguish intentional motions (e.g. sitting or crouching) from actual falls in diverse settings (e.g. cluttered homes or public spaces) [9, 15]. Second, real-time processing constraints limit the deployment of computationally intensive models, particularly for wearable devices with limited power resources [16, 22]. Third, data scarcity and bias in existing datasets often reduce generalization, as models trained in controlled laboratory environments may not adapt to unpredictable real-world scenarios [23, 24]. In addition, privacy concerns arise with vision-based methods, requiring trade-offs between accuracy and ethical data usage [17, 25]. Finally, integrating multimodal sensor data (e.g. inertial sensors, cameras, and environmental IoT devices)

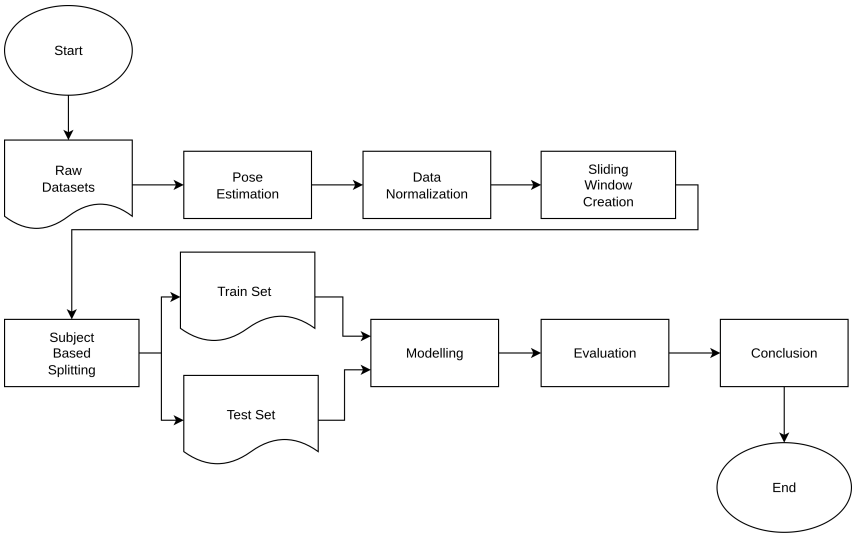

Figure 1: Research Workflows

while maintaining user comfort and system reliability remains an open problem [**?**, 26]. Addressing these challenges requires balancing algorithmic complexity, practical deployability, and ethical considerations to create inclusive and adaptive solutions.

3 Proposed Method

This section presents the general approach used in this study to perform human motion forecasting and fall classification. The process begins with raw video data, from which pose information is extracted using YOLOv11 to obtain 17 keypoints per frame. These keypoints are normalized and organized into sequences using a sliding window technique. The resulting sequences are then split into training and testing sets according to the subject. Finally, the data are used to train and evaluate multitask deep learning models for simultaneous forecasting and classification. The complete workflow is illustrated in Figure 1.

3.1 TelUP HumanFall Forecasting Dataset

The TelUP HumanFall Forecasting Dataset is a curated collection of annotated video sequences designed for human movement prediction and fall detection. It consists of 72 videos collected from six participants who performed both fall and non-fall activities with varied speeds and styles to increase diversity and realism. Each frame is annotated with a binary label indicating fall (1) or non-fall (0). The fall category includes six types of motion: backward, forward, left, right, sitting, and standing. The non-fall category covers daily activities such as jumping, lying down, picking up objects, squatting, stretching, and walking. Statistically, the dataset contains video clips ranging from 4.77 to 11.1 seconds in length, with an average duration of 6.56 seconds, providing a balanced and quantitative overview

of its composition. These details highlight the structure of the data set and strengthen its suitability for classification and prediction tasks in human fall motion research. This structure enhances the generalizability of models trained on the data. Furthermore, the temporal layout of the data set supports the analysis of motion over time, allowing it to be used for both classification and prediction tasks to predict future human movements and assess the probability of falls. The general setup and recording environment used to capture the data set are illustrated in Figure 2, showing how the placement of the camera and the positioning of the actors were arranged to ensure consistent data acquisition. In this experiment, we set the training set with subject numbers 1, 2, 3, and 4. While, subject number 5 and 6 are for testing set.

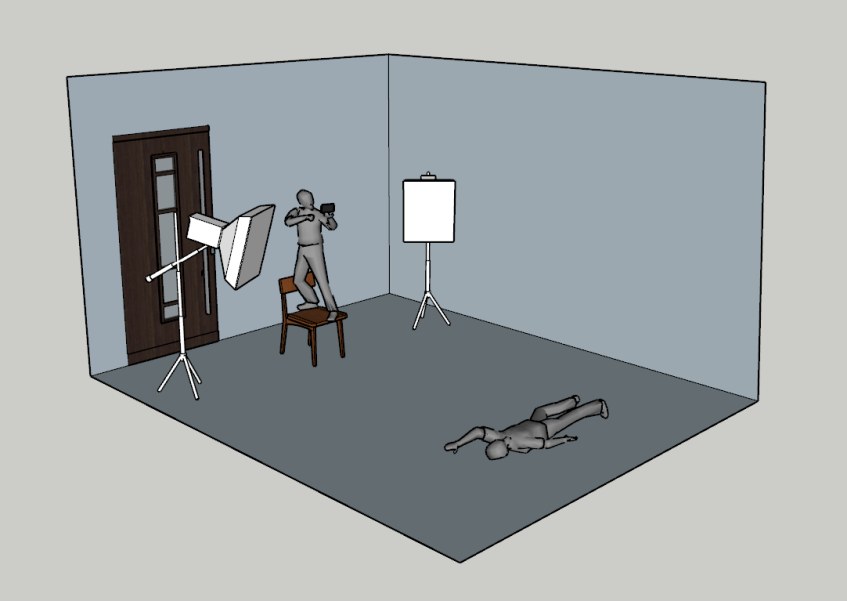

Figure 2: Illustration of Data Collection Process

3.2 Data Preprocessing

Data preprocessing plays an essential role in the preparation of raw video inputs for model training and evaluation. The preprocessing pipeline in this study involves the transformation of video frames into structured numerical representations suitable for deep learning. The details of each step in the data preprocessing are described in Sections 3.2.1, 3.2.2, and 3.2.3, respectively. These steps include pose estimation using YOLOv11, conversion of keypoint sequences into time series format, and normalization of keypoint coordinates.

3.2.1 Video to Keypoints Extraction

In the first stage, we use a pose estimation method based on YOLOv11 to extract key human skeletal points from each frame of the video. The model detects all individuals present in the scene and estimates 17 body keypoints, with each point represented as two-dimensional

(x, y) coordinates. These keypoints capture the essential motion features of the human body, including joints such as the shoulders, elbows, hips, and knees, and serve as the fundamental input for subsequent processing and model training.

3.2.2 Keypoints Normalization

After extracting 17 keypoints for each subject in every frame, the (x, y) coordinates are normalized to ensure spatial consistency in the data set. This normalization minimizes the effect of variations in frame size or subject position and helps the model learn motion patterns independent of scale or location. In this study, all frames are resized to a fixed resolution of 500×500 pixels. Normalization is performed using Min-Max scaling to rescale the keypoints in a fixed range between 0 and 1. Given a keypoint coordinate $p = (x, y)$ in the original frame, the normalized coordinate $p' = (x', y')$ is computed as follows:

$$x' = \frac{x}{W}, \quad y' = \frac{y}{H} \tag{1}$$

where W and H denote the width and height of the frame, respectively. Since all frames are resized to $W = 500$ and $H = 500$, the normalized coordinates fall within the range $[0, 1]$.

3.2.3 Sequence Generation with Sliding Window

To prepare the normalized keypoint sequences for training, a sliding window technique is applied to segment the data into overlapping input-output pairs. Both the input and output windows consist of 15 consecutive frames, and the window moves forward with a step size of 5 frames. This means that for each segment, the model receives 15 frames as input and is tasked with forecasting the following 15 frames. In addition, the classification label for each output window is determined by analyzing the labels of the frames within that window. The majority class within the 15-frame output window is selected and encoded to represent the overall class for that segment.

3.3 Predictive Models

In this study, we propose a multitask learning architecture capable of simultaneously performing two tasks. The first task focuses on classifying types of human motion, with particular emphasis on falling detection. The second task involves forecasting future sequences of human poses. By combining these tasks within a unified framework, the model can learn shared representations that enhance both performance and robustness, as supported by previous research [27, 28]. To evaluate the effectiveness of this approach, we implement and assess three types of neural network architecture such as Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), and Long-Short-Term Memory (LSTM). Each model processes sequential input data and generates outputs that include future pose predictions and motion classification probabilities.

3.3.1 Multi Layer Perceptron

The MLP-based model treats the input sequence as a flattened vector and processes it through a series of fully connected layers. This architecture learns a shared hidden representation that is passed into two branches. One branch is responsible for pose forecasting,

while the other performs motion classification. Although MLP does not explicitly model temporal relationships, it can still learn sequential patterns when time-based features are carefully encoded. However, its performance may be limited for highly dynamic sequence data compared to temporal models such as RNNs or LSTMs [29].

3.3.2 Recurrent Neural Network

The RNN model is designed to process sequential data by maintaining a hidden state that evolves over time. At each time step, the model updates its internal representation based on the current input and the previous state. The hidden state from the final time step is used for classification, while the outputs of all steps are used for pose forecasting. Despite its ability to model temporal sequences, standard RNN suffers from the vanishing gradient problem, which makes it difficult to learn long-term dependencies in sequences [30]. This limitation reduces its effectiveness for tasks that involve extended motion patterns.

3.3.3 Long Short-Term Memory

LSTM improves on the RNN by introducing memory cells and gating mechanisms to control information flow. These gates help the model retain important signals across longer sequences, which is essential for capturing complex human motion over time [31]. In the proposed method, LSTM processes the input sequence and generates outputs for both tasks. Forecasting is performed using the outputs at each time step, while classification is based on the hidden state at the final time step. LSTM is well suited for modeling temporal dependencies in sequential human pose data [32].

3.3.4 Loss Function

To train the multi-task models effectively, we adopt a joint loss function that combines two individual loss components: Mean Squared Error (MSE) for the forecasting task and Cross-Entropy Loss for the classification task. This combination allows the model to simultaneously learn the regression of future keypoint positions and the classification of motion categories, particularly for detecting falls. The total loss used during backpropagation is formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forecast}} + \mathcal{L}_{\text{classify}} \tag{2}$$

The forecasting loss, $\mathcal{L}_{\text{forecast}}$, is computed using the Mean Squared Error (MSE), which penalizes the squared differences between the predicted and actual keypoints:

$$\mathcal{L}_{\text{forecast}} = \frac{1}{N} \sum_{i=1}^{N} \|y_i - \hat{y}_i\|^2 \tag{3}$$

where y_i is the ground truth keypoint, $\hat{y}_i$ is the predicted keypoint, and N is the total number of keypoints in the output window. The classification loss, $\mathcal{L}_{\text{classify}}$, is calculated using the Cross-Entropy Loss, a standard choice for binary classification problems:

$$\mathcal{L}_{\text{classify}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{4}$$

where C is the number of classes (in this case, 2), y_i represents the true label, and $\hat{y}_i$ denotes the predicted probability for each class.

3.4 Evaluation Methods

In this research, we separate the model evaluation based on the forecasting task and the classification task. The predictive model has two tasks which are to forecast the human motion in 10 frames ahead and then classifying the forecasted motion into fall or non-fall classes. Performing two different functions at the same time. We formulate the following proposition. evaluation method to balance performance metrics on forecasting and classification problems.

3.4.1 Forecasting Task

To measure the accuracy of these predictions, we use the Mean Per Joint Position Error (MPJPE). This metric calculates the average distance between the predicted keypoints and the actual keypoints.

$$\text{MPJPE} = \frac{1}{N} \sum_{i=0}^{N} ||y_i - \hat{y}_i||_2 \tag{5}$$

In this formula, y_i is the true position of the keypoint, $\hat{y}_i$ is the predicted position, and N is the total number of joints. A lower MPJPE value means better forecasting performance.

3.4.2 Classification Task

To evaluate the performance of classification, we use a confusion matrix, which summarizes the number of correct and incorrect predictions for each class. From this we compute the accuracy, which shows how often the model makes correct predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{6}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives. Higher accuracy means that the model performs better in detecting falls correctly.

4 Experimental Setup

We conducted the experiment running on a machine with AMD Ryzen 7 CPU, NVidia RTX GeForce 4090. The evaluation included benchmarking the models based on the number of trainable parameters. Meanwhile, we set the hyperparameter for the training process with 1000 epochs, batch size of 32, trained with Adam optimizer with 1×10^{-4} learning rate.

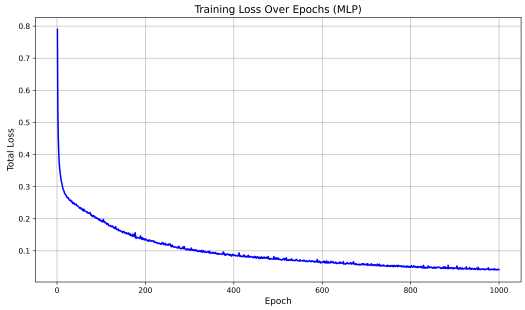

Figure 3: Training loss over epochs for the MLP model

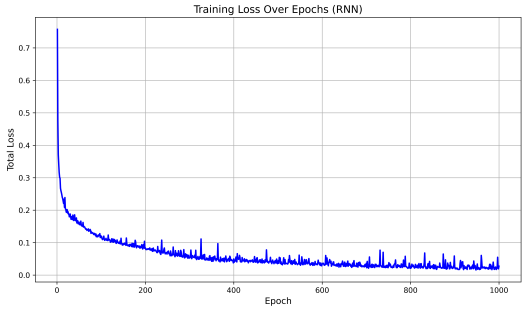

Figure 4: Training loss over epochs for the RNN model

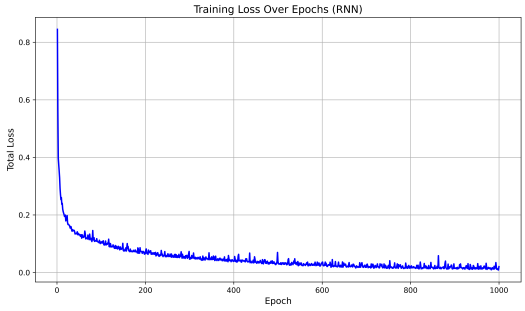

Figure 5: Training loss over epochs for the LSTM model

Figures 3, 4, and 5 illustrate the training loss over 1000 epochs for the MLP, RNN, and LSTM models, respectively. In the early training phase (approximately epochs 0–200), all models exhibit a steep decrease in loss, indicating rapid learning and adaptation to the data. As training progresses, the loss continues to decrease more gradually. Notably, the loss curves begin to flatten out starting around epoch 600, suggesting that the models are approaching convergence. This plateau indicates that further improvements in loss are

minimal and that the models have reached a relatively stable state. Although the LSTM model shows occasional fluctuations, it still follows an overall downward trend, reflecting stable training behavior.

5 Experimental Result and Discussion

5.1 Quantitave Evaluation

The results of the experiment of the proposed multitask models are summarized in Table 1. The evaluation focuses on the Mean Per Joint Position Error (MPJPE), classification accuracy, and model complexity in terms of trainable parameters. The MLP model achieves the most accurate pose prediction with the lowest MPJPE of 0.2630 (131.5 pixels), but it also records the lowest classification accuracy at 85.08%. It is the most complex model, with 131,456 trainable parameters. The RNN model, although it has the smallest number of parameters (25,636), offers a balanced trade-off, achieving an MPJPE of 0.3252 and an improved classification accuracy of 88.37%. The LSTM model, with 88,612 parameters, delivers the highest classification accuracy at 92.89%, although it has a slightly higher MPJPE of 0.3371 (168.6 pixels). These results demonstrate a trade-off between pose forecasting precision and motion type classification, with the LSTM model showing the strongest generalization in recognizing human motion patterns. The evaluation was carried out on a subject-independent test set comprising two participants excluded from the training, ensuring that the models were evaluated on previously unseen individuals.

Table 1: Experiment Results with Model Parameters

Model	MPJPE	Accuracy (%)	Params
MLP	0.2630 (131.5 pixels)	85.08	131,456
RNN	0.3252 (162.6 pixels)	88.37	25,636
LSTM	0.3371 (168.6 pixels)	92.89	88,612

5.2 Qualitative Evaluation

To further examine forecast performance, we conducted a qualitative evaluation using the MLP model, which achieved the lowest Mean Per Joint Position Error (MPJPE) of 0.2630. This model was selected for visual analysis as it represents the most accurate forecasting model quantitatively. The motion sequence chosen for visualization is categorized under *forward falls*, a dynamic and challenging class that tests the model's ability to anticipate rapid, full-body movements.

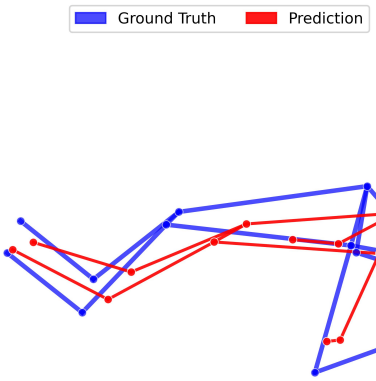

Figure 6: Prediction Comparison in Frame of MLP Model

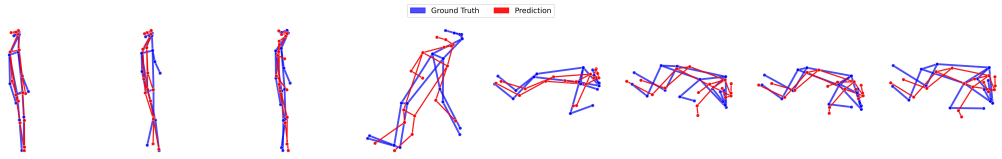

Figure 7: Prediction Comparison in Frame of MLP Model

As shown in Figure 6, the MLP model is able to approximate the general body configuration at a single point in time. However, some deviations can be observed from the truth of the ground, particularly in the lower extremities and arm positions. The predicted pose (red) appears slightly contracted compared to the ground truth (blue), indicating difficulty in accurately capturing limb extension during the impact phase of the fall.

Figure 7 illustrates the prediction performance in a sequence of frames. Although the model successfully captures the overall motion trajectory and body orientation as the subject transitions from standing to falling, notable mismatches are present. For instance, the predicted leg positions during mid-fall frames lag behind the actual motion, and the upper body orientation appears overly rigid. This suggests that although the MLP model performs well numerically, it lacks sensitivity to the more subtle joint dynamics required for accurate prediction of the fall motion.

These visual discrepancies highlight that a low MPJPE does not fully guarantee high perceptual or physical accuracy, especially in complex temporal contexts. Therefore, qualitative analysis remains essential for identifying practical limitations and guiding future improvements in forecasting models.

6 Conclusion

This research presented a comparative study of three multitask learning models, MLP, RNN, and LSTM, for simultaneous pose forecasting and motion classification tasks. The

experimental results show that while the MLP model achieved the lowest Mean Per Joint Position Error (MPJPE) of 0.2630, it lagged in classification performance. In contrast, the LSTM model attained the highest classification accuracy of 92.89%, but produced a slightly higher forecast error. These findings underscore a trade-off between forecast precision and classification performance, suggesting that no single model excelled universally on both tasks. A qualitative evaluation using the MLP model revealed that even the most quantitatively accurate model struggled to capture certain complex joint dynamics, particularly in high-velocity motions such as forward falls. Discrepancies in the predictions for the lower limb and upper body indicate that existing models, despite their low MPJPE scores, still face challenges in producing perceptually accurate forecasts over time.

In addition, the results suggest that the current forecast pipeline could benefit from further refinement. Although the models were trained efficiently and reached convergence in a short time frame, their accuracy, especially in temporally dynamic sequences, remains suboptimal. One contributing factor may be the limited expressiveness of the architectures used and the reliance on raw keypoint sequences obtained via YOLO based pose estimation, which may introduce noise or inconsistencies. To improve forecasting performance, future work should consider several directions: (1) applying systematic hyperparameter tuning to optimize learning dynamics, (2) adopting more modern and expressive architectures such as Transformer-based or graph convolutional networks that better capture spatio-temporal dependencies, and (3) incorporating more structured and semantically enriched keypoint extraction pipelines that do not solely depend on YOLO but integrate body structure priors or use high-precision pose estimation frameworks.

In conclusion, while this study demonstrates the feasibility of using multitask learning for joint motion forecasting and classification, it also highlights the current limitations in motion realism and generalization, emphasizing the need for more robust modeling and data processing approaches in future research.

Data and Source Code Availability

The data set and the source code are open to the public to ensure the transparency and sustainability of the research. The source code is available on github https://github.com/AndiDemonLab/HumanFallForecasting/.

References

[1] World Health Organization, "Falls," 2024. Accessed: 2025-07-11.

[2] World Health Organization, "Ageing and health," Oct. 2024. Accessed: 2025-07-11.

[3] J. R. Ehrlich, S. E. Hassan, and B. C. Stagg, "Prevalence of falls and fall-related outcomes in older adults with self-reported vision impairment," *Journal of the American Geriatrics Society*, vol. 67, no. 2, pp. 239–245, 2019.

[4] L. Liu, Y. Sun, Y. Li, and Y. Liu, "A hybrid human fall detection method based on modified yolov8s and alphapose," *Scientific Reports*, vol. 15, no. 1, p. 2636, 2025.

[5] 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, 2021.

[6] S.-H. Jung, J.-M. Hwang, and C.-H. Kim, "Inversion table fall injury, the phantom menace: Three case reports on cervical spinal cord injury," *Healthcare*, vol. 9, no. 5, 2021.

[7] R. G. Candraningtyas, A. P. Yunus, and Y. H. Choo, "Human Fall Motion Prediction - A Review," 1 2024.

[8] M. A. Khatun, M. A. Yousuf, S. Ahmed, M. Z. Uddin, S. A. Alyami, S. Al-Ashhab, H. F. Akhdar, A. Khan, A. Azad, and M. A. Moni, "Deep cnn-lstm with self-attention model for human activity recognition using wearable sensor," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, p. 2700316, 2022.

[9] W. Mao, M. Liu, M. Salzmann, and H. Li, "Multi-level motion attention for human motion prediction," *International journal of computer vision*, vol. 129, no. 9, pp. 2513–2535, 2021.

[10] M. Kepski and B. Kwolek, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[11] A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, "Sisfall: A fall and movement dataset," in *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pp. 393–396, IEEE, 2017.

[12] G. J. Horng and K. H. Chen, "The smart fall detection mechanism for healthcare under free-living conditions," *Wireless Personal Communications*, vol. 118, no. 1, pp. 715–753, 2021.

[13] C. A. U. Hassan, F. K. Karim, A. Abbas, J. Iqbal, H. Elmannai, S. Hussain, S. S. Ullah, and M. S. Khan, "A cost-effective fall-detection framework for the elderly using sensor-based technologies," *Sustainability*, vol. 15, no. 5, p. 3982, 2023.

[14] S. K. Bhoi, S. K. Panda, B. Patra, B. Pradhan, P. Priyadarshinee, S. Tripathy, C. Mallick, M. Singh, and P. M. Khilar, "Fallds-iot: a fall detection system for elderly healthcare based on iot data analytics," in *2018 International Conference on Information Technology (ICIT)*, pp. 155–160, IEEE, 2018.

[15] H. Mankodiya, D. Jadav, R. Gupta, S. Tanwar, A. Alharbi, A. Tolba, B.-C. Neagu, and M. S. Raboaca, "Xai-fall: Explainable ai for fall detection on wearable devices using sequence models and xai techniques," *Mathematics*, vol. 10, no. 12, p. 1990, 2022.

[16] B.-H. Wang, J. Yu, K. Wang, X.-Y. Bao, and K.-M. Mao, "Fall detection based on dual-channel feature integration," *IEEE Access*, vol. 8, pp. 103443–103453, 2020.

[17] S. Saurav, R. Saini, and S. Singh, "A dual-stream fused neural network for fall detection in multi-camera and 360 videos," *Neural computing and applications*, vol. 34, no. 2, pp. 1455–1482, 2022.

[18] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias, "Fall detection and activity recognition using human skeleton features," *Ieee Access*, vol. 9, pp. 33532–33542, 2021.

[19] L. Yao, W. Min, and K. Lu, "A new approach to fall detection based on the human torso motion model," *Applied Sciences*, vol. 7, no. 10, p. 993, 2017.

[20] O. Medjaouri and K. Desai, "Hr-stan: High-resolution spatio-temporal attention network for 3d human motion prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2540–2549, 2022.

[21] L. Chen, R. Liu, X. Yang, D. Zhou, Q. Zhang, and X. Wei, "Sttg-net: a spatio-temporal network for human motion prediction based on transformer and graph convolution network," *Visual Computing for Industry, Biomedicine, and Art*, vol. 5, no. 1, p. 19, 2022.

[22] G.-J. Horng and K.-H. Chen, "The smart fall detection mechanism for healthcare under free-living conditions," *Wireless Personal Communications*, vol. 118, no. 1, pp. 715–753, 2021.

[23] K.-L. Lu and E. T.-H. Chu, "An image-based fall detection system for the elderly," *Applied Sciences*, vol. 8, no. 10, p. 1995, 2018.

[24] N. Worrakulpanit and P. Samanpiboon, "Human fall detection using standard deviation of c-motion method," *Journal of Automation and Control Engineering*, vol. 2, no. 4, 2014.

[25] A. Y. Alaoui, Y. Tabii, R. O. H. Thami, M. Daoudi, S. Berretti, and P. Pala, "Fall detection of elderly people using the manifold of positive semidefinite matrices," *Journal of Imaging*, vol. 7, no. 7, p. 109, 2021.

[26] G. Diraco, G. Rescio, P. Siciliano, and A. Leone, "Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing," *Sensors*, vol. 23, no. 11, p. 5281, 2023.

[27] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, Jul 1997.

[28] S. Liu and W. Hao, "Forecasting the scheduling issues in engineering project management: Applications of deep learning models," *Future Generation Computer Systems*, vol. 123, pp. 85–93, 2021.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016. Available at http://www.deeplearningbook.org.

[30] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

[32] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," 2015.