Preventing Elderly Injuries: A Pose-Based Framework for Fall Detection

Preventing Elderly Injuries: A Pose-Based Framework for Fall Detection

Lorenzo Longarini
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
s1110740@studenti.univpm.it

Alessandro Rongoni
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
s1114661@studenti.univpm.it

Aldo F. Dragoni
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
a.f.dragoni@univpm.it

Geremia Pompei
Dept of Computer Science
University of Pisa
Pisa, Italy
geremia.pompei@di.unipi.it

Mariorosario Prist
Dept. of Information Engineering (DII)
Polytechnic University of Marche
Ancona, Italy
m.prist@staff.univpm.it

***Abstract*—Automatic fall detection systems that can rapidly identify dangerous events are increasingly important for elderly care. This paper presents a novel fall detection methodology based on a modular four step pipeline: dataset preparation, skeletal keypoint extraction, advanced feature engineering, and temporal pattern analysis. The feature engineering process transforms human pose data into rich representations capturing ground proximity, body orientation, and motion dynamics-key biomechanical signatures that differentiate falls from daily activities. These engineered features are then analyzed through a temporal framework that models the distinctive sequential patterns characteristic of falling events, including LSTM, GRU, and Bidirectional LSTM architectures. For implementation, You Only Look Once (YOLO11m-pose) estimation was employed in the keypoint extraction step, while the temporal classifier module tests the above networks. Evaluations on the GMDCSA24 dataset demonstrate the approach's effectiveness. By utilizing skeletal pose data and a plug-and-play design, the approach offers multiple advantages, for example, robust performance across different lighting conditions, and adaptability to various deployment scenarios including healthcare facilities, homes for the elderly, and public spaces where continuous monitoring is necessary.**

***Index Terms*—Fall Detection, Pose Estimation, You Only Look Once, Keypoints, Feature Extraction, Deep Learning, Healthcare Monitoring**

I. Introduction

Falls represent one of the main causes of injuries and disabilities among the elderly, leading to serious health consequences and significant social and healthcare costs [1]. Timely detection of these events can drastically reduce negative consequences, enabling rapid interventions by healthcare providers or family members. Traditionally, fall detection has been approached through wearable sensors, such as accelerometers or gyroscopes, which classify events based on empirical thresholds [2]. Despite their simplicity and low cost, these systems have difficulty distinguishing daily activities from true falls, causing frequent false alarms. With the recent development of deep learning and computer vision techniques, research has shifted towards video-based systems, particularly using pose estimation techniques. In this area, methods based on models such as OpenPose [3] and variants of the You Only Look Once (YOLO) [4] algorithm adapted for real-time extraction of human body keypoints [5], [6] have been proposed. More recent frameworks that combine pose estimation with temporal attention or graph-based reasoning have further improved detection accuracy [7]. In particular, models like YOLOv7-W6-Pose have demonstrated high performance in detecting anomalous poses, thus enabling the identification of postures indicative of falls. However, these models often analyze single frames or short video sequences without explicit modeling of temporal dynamics, limiting their ability to correctly identify complex events or transient movements. To overcome this limitation, some studies have explored the use of Recurrent Neural Networks (RNNs) [8] such as Long Short-Term Memory (LSTM) [9], Gated Recurrent Unit (GRU) [10], and Bidirectional LSTM (BiLSTM) [11], capable of analyzing complex temporal sequences and learning dynamic movement patterns. These approaches, however, have primarily been applied to data from wearable sensors, showing the effectiveness of the temporal component in more accurately distinguishing real falls from daily activities [12]. A parallel line of research has investigated convolutional–recurrent hybrids on video data [13], yet without an explicit feature-engineering stage capable of emphasising biomechanical cues. Despite extensive research efforts, the explicit integration of pose estimation with comprehensive feature engineering and subsequent temporal modelling has been only marginally explored. To tackle this gap, this paper introduces a modular, pose-based framework composed of four main modules: Dataset, Keypoints Extraction, Feature Extraction, and Temporal Classifier. The GMDCSA24 dataset [14]–[16], specifically designed for human fall detection and containing annotated videos of simulated falls and activities of daily living (ADL), is employed as the experimental dataset. Skeletal keypoints extracted using the YOLO11m-pose model

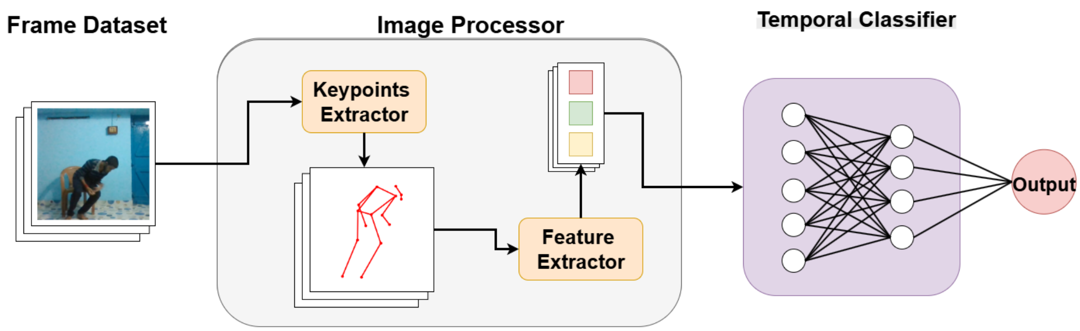

Fig. 1. The workflow is composed of four modules: Dataset, Keypoint Extractor, Feature Extractor, Temporal Classifier

[17] are first temporally associated through the Hungarian algorithm [18]. Subsequently, the proposed Feature Extraction module transforms the raw coordinates into high-level biomechanical descriptors, capturing critical aspects such as ground proximity, body orientation changes, joint-angle dynamics, and abrupt velocity variations—features often diluted or omitted in end-to-end approaches. These engineered descriptors are then used to train and evaluate three recurrent neural network classifiers (LSTM, GRU, and BiLSTM), allowing a direct and systematic comparison of unidirectional, gated, and bidirectional recurrent mechanisms using the same enriched feature representation.

The main contributions of this work are:

- A novel modular framework for fall detection composed of independent and interchangeable modules (Dataset, Keypoints Extractor, Feature Extractor, and Temporal Classifier), enabling flexible experimentation and facilitating improvements.
- A tailored feature extraction module that transforms raw skeletal coordinates into robust biomechanical descriptors specifically designed to capture critical indicators of falls, such as ground proximity, body orientation, joint-angle dynamics, and motion abruptness.
- A systematic comparative evaluation of three temporal recurrent architectures (LSTM, GRU, BiLSTM) within the same enriched feature representation, providing insights into the impact of unidirectional, gated, and bidirectional temporal modeling.
- Experimental validation demonstrating superior performance of the proposed feature extraction approach compared to raw skeletal keypoints alone, using the publicly available GMDCSA24 dataset.

The paper is organised as follows: Section 2 provides a detailed description of the proposed methodology, illustrating the four independent modules comprising the framework (Dataset, Keypoints Extractor, Feature Extractor, and Temporal Classifier), as well as the evaluation metrics employed. Section 3 presents the results from experiments conducted on the GMDCSA24 dataset, including a detailed comparative analysis of the temporal models tested and the different data representations used. Finally, Section 4 summarises the main conclusions and outlines potential directions for future developments.

II. Materials and Methods

This section describes the adopted methodology, illustrated by the pipeline shown in Fig. 1, which is divided into four independent and functional modules: Dataset, Keypoint Extractor, Feature Extractor, and Training. This modular architecture promotes flexibility and adaptability, as each block can be updated or replaced separately as long as the standardised interface with the adjacent modules is maintained. The pipeline enables the isolation and evaluation of the effectiveness of each module independently, facilitating comparative analyses and future expansions of the proposed framework.

A. Dataset

The dataset module is crucial because it provides the annotated video sequences on which the entire analysis and training process is based. For this experiment, we used the GMDCSA-24 dataset, which is specifically designed for the automatic detection of falls in domestic environments. This corpus comprises 160 low-resolution video sequences (0.92 MP, 720p, 30 FPS uncompressed) divided into 81 real falls and 79 activities of daily living (ADLs). The video clips were recorded in three different domestic contexts with different lighting conditions (daylight and low light) and the actors wore different clothes. Each video is accompanied by manual annotations organised in a CSV file containing information such as the file name, the activity class (either Fall or a specific ADL), the clip duration, the subject ID and the time slot of the recording. Such metadata facilitate integration into the training process. Fig. 2 illustrates representative examples of the sequences used. The dataset was split into 70% for training, 10% for validation and 20% for testing, for model training and evaluation.

B. Keypoints extractor

The keypoint extraction module identifies the key joints that describe the subject's posture in each video frame. The Hungarian algorithm was utilised to ensure correct traceability of individuals between frames. Applying this phase enables us to identify the keypoints in a single frame, as shown in Fig. 3. In this implementation, the YOLO11m-pose model was

Fig. 2. Fall and ADL frames of GMDCSA dataset compared.

chosen. This model can detect 17 distinct keypoints for each subject, ensuring a complete representation of the human body structure (see Fig. 4). The keypoints are divided as follows:

- Head and face: nose (1), left and right eyes (2–3), left and right ears (4–5).
- Trunk: left and right shoulders (6–7) and left and right hips (12–13).
- Upper limbs: left and right elbows (8–9), left and right wrists (10–11).
- Lower limbs: left and right knees (14–15) and left and right ankles (16–17).

The modular architecture enables this model to be easily replaced with other pose estimation models, such as lighter versions (YOLOv8-Pose) or more accurate ones (HRNet), while preserving the structure of the workflow and compatibility with subsequent modules.

C. Feature Extractor

The Feature Extractor module is fundamental to the proposed methodology, converting the raw coordinates of the keypoints into a richer, more descriptive representation of human movement. This transformation enables predictive models to more effectively identify characteristic patterns of falls compared to ordinary movements. The extracted features fall mainly into three categories: postural, geometric-spatial, and

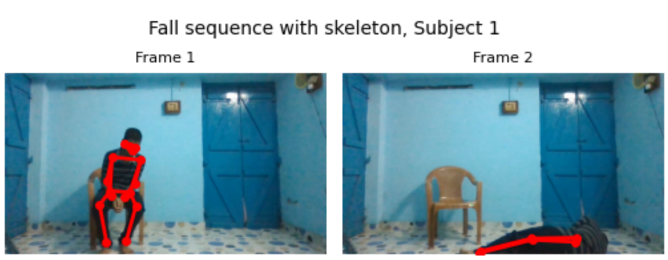

Fig. 3. Fall frames of GMDCSA with skeleton keypoints.

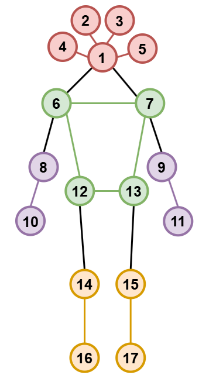

Fig. 4. Skeleton keypoints used by YOLO11m-pose.

dynamic. One of the most relevant postural features is the centroid of body posture, representing the average position of all detected keypoints. This is useful for identifying and monitoring sudden movements or postural anomalies. From a geometric and spatial point of view, critical joint angles (such as those between the trunk and lower limbs) are considered to detect unnatural positions or sudden changes in posture. The height-to-width ratio of the bounding box is also considered to identify when the subject transitions from a vertical (upright) posture to a horizontal one (typical in falls). Finally, a series of dynamic features are added to the static features described above to capture the temporal evolution of movement. In particular, the velocity and acceleration of the most critical keypoints (e.g. the head, torso, and centre of gravity) are computed to capture sudden and abrupt movements associated with falls. These are further complemented by the cumulative displacement of the keypoints computed in short time windows and by a summary of the local movement statistics (mean, standard deviation and range) computed on moving windows. These measures summarise the variability and dynamic nature of the movement, enabling the models to distinguish critical events more clearly from daily activities. This integrated combination of postural, geometric, and dynamic features ensures the model has a complete and discriminative representation, guaranteeing high performance in automatic fall detection.

D. Model Architecture

The final module is responsible for training and classifying activities using recurrent neural networks. Descriptors generated by the previous module are used as input for sequential classification models, such as LSTM, BiLSTM and GRU. These models treat feature sequences as temporal data and recognise distinctive patterns in body motion dynamics. The training phase is optimised using techniques such as dropout, batch normalisation and early stopping criteria, which prevent overfitting problems and maximise the classifier's generalisation ability. The recurrent architecture adopted in this module can be easily replaced with other temporal classifier structures (e.g. temporal Transformer networks or spatial-temporal graph networks), provided the input data (fixed-length sequences) and binary output interfaces are standardised. This section describes the architectures used for the experiments.

1) LSTM: The LSTM architecture is our baseline, consisting of two stacked layers of Long Short-Term Memory that learn the temporal sequence in a single direction. The first layer returns the full sequence of internal states, while the second layer compresses the information into a final vector. Dropout and batch normalisation, applied after each layer, help prevent overfitting and stabilise learning. This traditional setup is useful to evaluate how well the unidirectional analysis can recognise the onset, development and termination phases of a drop event.

2) BiLSTM: In the BiLSTM variant, each LSTM layer is duplicated in a forward-backward fashion, allowing the model to simultaneously integrate both past and future information within each time window. While increasing the number of parameters, bidirectionality provides richer context and improves the distinction between dynamic transitions typical of a fall and less abrupt everyday movements.

3) GRU: The GRU version, uses the Gated Recurrent Unit, a simplification of the LSTM that reduces the number of internal gates and therefore the total number of parameters. Again, the two-stage structure follows the same sequence return scheme in the first layer and final vector in the second, with batch normalisation and dropout to control overfitting. This model proves to be an excellent compromise between temporal modelling capabilities and computational efficiency.

E. Metrics

In order to assess the performance of the models in a complete and balanced way, we consider some indicators that capture both the overall accuracy and the behaviour with respect to the fall and non-fall classes:

- Accuracy: percentage of correct predictions out of the total, useful for a general overview.
- Precision: percentage of correct falls out of the total number of positive reports, which indicates the control of false alarms.
- Recall: percentage of real falls captured by the model, measures sensitivity to positive cases.
- F1-score: harmonic mean between precision and recall, balancing the two aspects, especially in case of imbalance.

III. Results and Discussion

This section describes the experiments performed to evaluate the ability of the models to distinguish fall events, comparing the different chosen architectures. All of the experiments reported in this section were carried out using a computer equipped with an AMD Ryzen 9 7945HX processor running at 2.50 GHz and with 32.0 GB of RAM. First, a grid search was conducted to identify the models' optimal hyperparameters. Then, the models were trained using these optimised parameters. Finally, the performance of two data representations was compared: one containing only the keypoints and the other containing the extracted features.

TABLE I
Hyperparameters tested during Grid Search.

Units	32	64	128
Dropout rate	0.4	0.6	0.8
Learning rate	1×10^{-3}	5×10^{-4}	1×10^{-4}
Batch size	8	16	32

TABLE II
Optimal hyperparameters selected after Grid Search.

Model	**Units**	**Dropout rate**	**Learning rate**	**Batch size**
LSTM	32	0.3	1×10^{-4}	8
GRU	64	0.5	1×10^{-4}	16
BiLSTM	32	0.3	1×10^{-4}	8

A. GridSearch

Grid search was employed to systematically explore the combinations of the four main hyperparameters: the number of units per layer, the dropout rate, the learning rate and the batch size. These are reported in Tab. I. This procedure enabled the objective evaluation of the influence of each parameter on the performance of the models, and the selection of the most effective configuration for each of the three architectures (LSTM, GRU and BiLSTM). A dedicated portion of the dataset was used as a validation set for each combination, thus ensuring a robust and non-arbitrary choice of hyperparameters. The results of the hyperparameter search, i.e. the optimal configurations obtained for each model, are reported in Tab. II.

B. Results

In this section, we compare the performance of obtained by the models trained using two different representations of the input data:

- Only Keypoints: only the normalized coordinates of the extracted skeletal points.
- Keypoints & Features: the coordinates of the keypoints enriched by the biomechanical and dynamic features described in the previous section.

The results presented in Tab. III confirm the effectiveness of the proposed approach, highlighting in particular the significant contribution provided by the feature extraction module compared to the use of raw keypoint coordinates alone. Across all tested architectures (LSTM, GRU, and BiLSTM), the inclusion of engineered biomechanical and dynamic features consistently results in higher accuracy, precision, recall, and

TABLE III
Experimental results for considered models.

Data	**Model**	**Acc. (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
Only Keypoints	LSTM	76.47	76.84	76.47	76.38
	GRU	76.47	80.23	76.47	75.71
	BiLSTM	85.29	85.41	85.29	85.28
Keypoints & Features	LSTM	88.23	90.47	88.23	88.07
	GRU	85.29	88.63	85.29	84.97
	BiLSTM	91.17	92.50	91.17	91.11

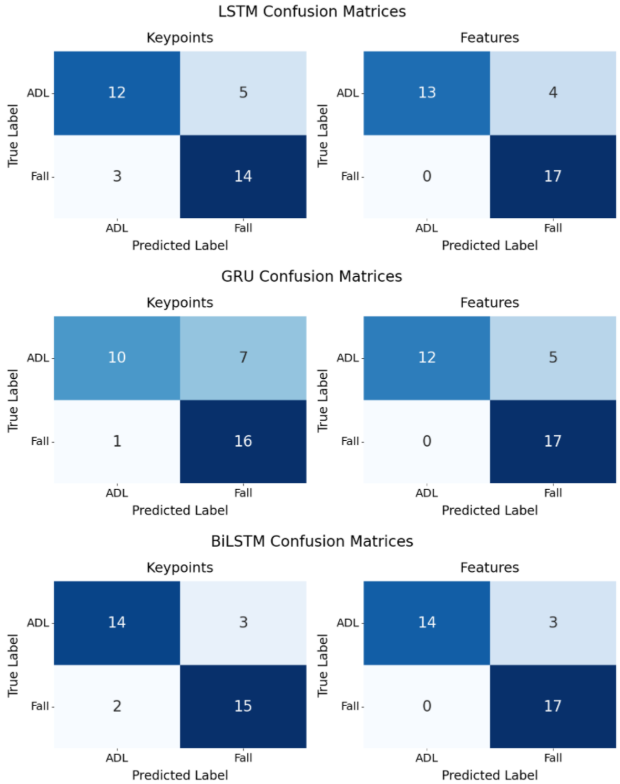

Fig. 5. Confusion matrices for the LSTM, GRU and BiLSTM models.

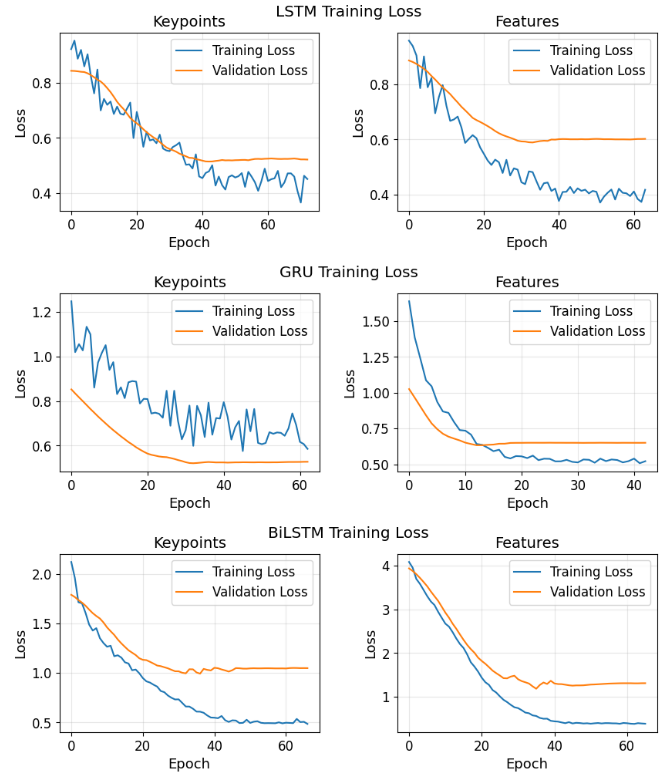

Fig. 6. Training and validation loss curves for LSTM, GRU and BiLSTM models.

F1-score, demonstrating enhanced discrimination between real falls and normal daily activities. Further insights can be drawn from the confusion matrices associated with each model (Fig. 5). These matrices illustrate how incorporating biomechanical descriptors reduces both false positives (daily activities mistakenly classified as falls) and false negatives (falls that are missed). Such improvement is particularly evident when directly comparing the matrices obtained under the two experimental conditions (keypoints only versus keypoints enriched by engineered features), confirming that the richer feature representation enables the temporal models to more effectively distinguish between subtle yet critical differences in human motion. Finally, the loss curves during training and validation phases (Fig. 6) provide additional information regarding the learning behavior of the models. Although the models trained with the proposed features generally converge more rapidly compared to those trained solely with raw keypoints, there is evidence of slight overfitting, particularly noticeable in the validation curves. This behavior is likely attributable to the limited size of the dataset used in the experiments, suggesting that further generalization could be achieved with additional training data or regularization techniques. Nevertheless, the overall improvement in model performance clearly indicates the effectiveness and robustness of the proposed feature extraction methodology. Among the architectures tested, BiLSTM consistently outperforms the LSTM and GRU models, owing to its capability of integrating both past and future temporal contexts. This result underscores the relevance of adopting bidirectional temporal analysis when detecting complex motion patterns typical of fall events.

C. Quantitative Comparison with the State of the Art

As shown in Tab. IV, performance reported in the literature is closely tied to the complexity of the evaluation setup. Among the references discussed in the introduction, we selected three works for direct comparison as they adopt similar input modalities and provide sufficient implementation details and metrics for a meaningful evaluation. Tirziu *et al.* [5] achieve 95.7% accuracy on the Le2i dataset using a single-person, posture-rule-based approach. Jeong *et al.* [13] report 100% on a custom dataset with highly controlled and scripted activities. In contrast, Liu *et al.* [6] evaluate on the more realistic MPFDD dataset and obtain 77.5% accuracy due to crowded scenes and small targets. Beyond numerical results, the methods differ in both architecture and scope. Tirziu *et*

TABLE IV
COMPARISON BETWEEN PROPOSED METHOD AND REFERENCE WORKS.

Reference	**Dataset**	**Accuracy (%)**	**F1 Score (%)**
Tirziu *et al.* [5]	Le2i fall dataset	95.70	95.50
Liu *et al.* [6]	MPFDD	77.50	77.40
Jeong *et al.* [13]	Custom	100.00	-
Current Work	GMDCSA24	91.17	91.11

al. use a fixed-threshold geometric model on single-person data without temporal reasoning. Jeong *et al.* apply LSTMs to raw key-points in controlled settings. Liu *et al.* combine a modified YOLOv8s and AlphaPose with a frame-wise hybrid classifier aimed at edge deployment. Our system adopts a modular structure with pose estimation, biomechanical feature extraction and bidirectional RNNs, allowing flexibility and interpretability. The proposed method is evaluated on GMDCSA24, a challenging dataset with variable lighting, 720p resolution, and unscripted daily activities. Moreover, the use of biomechanical descriptors increases robustness but also expands the learning space. For this reason, the slightly lower performance compared to highly controlled settings is expected and reflects a more realistic evaluation context.

IV. Conclusions

This work introduces an innovative, modular framework for automatic fall detection. It is based on advanced biomechanical feature extraction techniques and the application of recurrent temporal models. The framework's modularity is its main strength, as it allows different combinations of pose extraction models, feature engineering techniques, and temporal classifiers to be easily tested, facilitating the identification of the optimal configuration based on application context and specific operational requirements. The proposed feature extraction module is a particular highlight of this work, providing a richer, more discriminative representation of human motion and significantly improving performance compared to approaches based solely on raw skeletal keypoint coordinates. Integrating ground proximity descriptors, joint angles, and movement dynamics provides the system with highly discriminative biomechanical information, substantially improving its reliability and robustness. Future developments will include investigating the generalisation capability of the proposed framework in real-world settings characterised by variable visibility conditions, occlusions and less structured behaviours also by extending the evaluation to additional public datasets and comparing results across diverse acquisition contexts. Another possible development would be to integrate multimodal techniques, such as wearable sensor fusion, to strengthen detection robustness further and reduce the number of false alarms. A further line of future work concerns the systematic evaluation of latency and processing speed on embedded hardware, with a view to assessing the system's responsiveness under real-time constraints. Moreover, the utilisation of a modular design has the potential to facilitate the streaming of the sequence of poses to an augmented reality module. One such example is smart glasses, in which the fallen person is delineated within the scene and recovery instructions, informed by context, are displayed. Finally, the modularity of the framework enables the pipeline to be easily optimised for edge-computing scenarios, with the aim of bringing this technology into real-world environments where stringent requirements exist in terms of privacy, fast response times and low energy consumption.

References

[1] W. H. Organization, "Who global report on falls prevention in older age," *WHO Press, Geneva, Switzerland*, 2007.

[2] A. K. Bourke and G. M. Lyons, "A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor," *Medical engineering & physics*, vol. 30, no. 1, pp. 84–90, 2008.

[3] Z. Huang, Y. Liu, Y. Fang, and B. K. P. Horn, in *International Conference on Universal Village*, 2018.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[5] E. Tîrziu, A.-M. Vasilevschi, A. Alexandru, and E. Tudora, "Enhanced fall detection using yolov7-w6-pose for real-time elderly monitoring," *Future Internet*, 2024.

[6] L. Liu, Y. Sun, Y. Li, and Y. Liu, "A hybrid human fall detection method based on modified yolov8s and alphapose," *Scientific Reports*, p. Article 2636, 2025.

[7] H. Miawarni, T. A. Sardjono, E. Setijadi, M. H. Purnomo *et al.*, "Fall detection system for elderly (fds-e) using low-cost camera based on lstm and openpose," in *2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2023, pp. 1–6.

[8] A. C. Tsoi, "Recurrent neural network architectures: An overview," *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pp. 1–26, 1997.

[9] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 07 2019. [Online]. Available: https://doi.org/10.1162/neco_a_01199

[10] S. Nosouhian, F. Nosouhian, and A. K. Khoshouei, "A review of recurrent neural network architecture for sequence learning: Comparison between lstm and gru," 2021.

[11] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International conference on artificial neural networks*. Springer, 2005, pp. 799–804.

[12] J. Wu, J. Wang, A. Zhan, and C. Wu, "Fall detection with cnn-casual lstm network," *Information*, vol. 12, no. 10, p. Article 403, 2021.

[13] S. S. Jeong, N. H. Kim, and Y. S. Yu, "Comparison of fall detection systems based on yolopose and long short-term memory," 2024.

[14] E. Alam, A. Sufian, P. Dutta, M. Leo, and I. Hameed, "Gmdcsa24: A dataset for human fall detection in videos," *Data in Brief*, 2024.

[15] E. Alam, "ekramalam/GMDCSA24-A-Dataset-for-Human-Fall-Detection-in-Videos: 2.0," Jul. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.12921216

[16] E. Alam, A. Sufian, P. Dutta, and M. Leo, "Real-time human fall detection using a lightweight pose estimation technique," in *International Conference on Computational Intelligence in Communications and Business Analytics*. Springer, 2023, pp. 30–40.

[17] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2637–2646.

[18] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.