A Two-stage Human Fall Detection Model Based on Rule-Based Algorithm and CNN-LSTM

Aman Kumar Patel[1], Sneha Barmaiya[2], Megha Patidar[3] and Anand Singh Jalal*[4]

[1,2,3,4] School of Computer Science and Information Technology,
Devi Ahilya Vishwavidyalaya, Indore, India
[1]amanp4631@gmail.com, [2]snehaabarmaiya@gmail.com,
[3]meghapatidar13388@gmail.com, [4]anandsinghjalal@gmail.com*

Abstract. Human fall detection is an important area of concern in the context of healthcare monitoring systems and is a significant issue. Automatic fall detection systems help in the prevention of fatal injuries and rapid medical care for senior citizens living alone, children left alone, as well as in various other such instances. Fall detection models regardless of their precisions are struggling with fall detection in uncontrolled environments with respect to pose variations, lighting variations, fall instances with pose occlusions, and fall instances with high similarities among activities. In this paper, we propose a two-stage fall detection model that combines a rule-based fall detection approach with a CNN–LSTM model. A public fall detection dataset called Le2i is used for training of the fall detection model that contains information about fall instances, fall instances with their boundary box values, as well as instances with their time fall boxes. Experimental results indicate that the proposed fall detection model would significantly reduce the computational cost while providing comparable performance to existing fall detection models.

Keywords: Fall detection, CNN-LSTM, Healthcare.

1 Introduction

Human fall detection models have become a crucial component in healthcare monitoring, elderly care, and smart-home surveillance, especially because falls are a leading cause of injury-related hospitalization among older adults. The global elderly population faces increasing risk of injury, disability, and mortality due to falls, making automated monitoring essential [1]. Many older adults live alone or remain left unattended for long periods, thus creating hazardous delays in obtaining medical assistance after a fall [2]. Therefore, the design of a trustworthy automatic fall detection model that functions effectively without many of the controls typical of laboratory conditions is still an important research goal [3].

Traditional fall-detection models are normally categorized into wearable-sensor-based methods and vision-based methods. Wearable models often use accelerometers, gyroscopes, or sensors to identify sudden changes in motion, characteristic of falls [4].

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_24

However, their practical applications are restricted by the necessity of continuous user compliance, correct sensor placement, and battery availability [5].

Most of these constraints are overcome by vision-based models using RGB, depth, or multi-modal cameras to extract the posture features appearance, or motion cues without requiring the user to wear any device [3]. In controlled environments, methods relying pose estimation, silhouettes, and body-geometry analysis have shown convincing fall recognition performance. However, this can be a challenge when working in uncontrolled indoor environments. Challenges include pose variation, illumination changes, and background clutter, occlusion, and similarity between fall and non-fall actions such as fast sitting [6]. Deep learning has considerably increased the accuracy of activity recognition, but those models require large datasets and high computational resources which reduce their viability in real-time embedded models [7].

The Le2i [8] benchmark fall detection dataset includes a wide range of realistic scenarios. The benchmark includes multiple indoor environments - home, office, coffee room, lecture room, and is comprised of both fall and daily activity sequences. However, applying deep models to every video frame is computationally expensive, and purely rule-based models may fail in ambiguous cases such as slow falls or gradual collapse.

To address these limitations, we propose a two-stage fall-detection model where a lightweight rule-based module serves as a first-stage filter to quickly eliminate obvious non-fall activities, and a CNN-LSTM module acts as a second- stage classifier for more ambiguous cases. This two-stage design aims to achieve high accuracy, robustness, and real-time efficiency.

The remainder of the paper is structured as follows: Section 2 presents the related work, Section 3 describes the proposed two-stage fall detection model, Section 4 reports the experimental results, and finally, Section 5 concludes the paper.

2 Related Work

This section reviews existing approaches categorized into sensor-based, vision-based, deep-learning-based, and hybrid models.

Early research on fall detection was related to wearable sensors, such as accelerometers or gyroscopes, that measured sudden acceleration peaks occurring in falls. Kangas et al. [9] evaluated thresholding-based fall detection techniques in comparison with other approaches, concluding that thresholding-based solutions work satisfactorily in laboratory settings. Wearable sensors pose issues of compliance.

Vision-based approaches emerged as a powerful alternative. Classic approaches rely on background subtraction, silhouette extraction, and optical flow to determine abnormal motion patterns. Zhang et al. [10] provided a comprehensive analysis of vision-based fall detection methods and highlighted the challenges of pose and illumination variations. These handcrafted techniques struggle with robustness under real-world conditions.

However, with the advent of deep learning, models that incorporate CNNs became prominent for the extraction of spatial features. Ali et al. [11] used models of CNN for

fall activity classification on multi-camera video data. Later, temporal concepts were modelled using LSTM and 3D-CNN architectures. Donahue et al. [12] showed that a combination of CNN–LSTMs is effective for sequence classification.

Pose-based fall detection techniques also gained prominence. Chaudhuri et al. [13] extracted skeletal key points using a temporal CNN–LSTM network for human activity recognition. These approaches work adequately regardless of the background but are severely reliant on pose estimation.

Hybrid approaches combining multiple sensors or multiple algorithms have also been explored. Ahmed et al. [14] proposed a sensor-fusion fall detector integrating accelerometer data with CNN features, achieving high accuracy. However, these models require additional hardware and may not scale well.

In recent years, transformer-based temporal models have also drawn interest. A vision transformer (ViT) in conjunction with temporal attention mechanisms was proposed by Zhang et al. [15] for the detection of falls in indoor environments. Transformers are useful for capturing long-range temporal dependencies, but their applicability for small to medium-sized datasets like Le2i is limited by their need for large-scale training data and computational resources.

Current benchmarking research reveals enduring difficulties in practical implementation. Many cutting-edge fall detection models exhibit significant performance degradation when tested on unseen environments, as Li et al. [16] showed. Their results highlight the fact that robust real-world performance is not always correlated with high accuracy on carefully selected datasets, especially when there are occlusions, lighting variations, and a variety of camera viewpoints.

Recent studies emphasize the need for efficient models that balance computational cost with robustness in uncontrolled environments. Deep models alone are computationally expensive, while purely rule-based models produce high false-positive rates. This motivates the development of a two-stage model that leverage the strengths of both paradigms.

3 Proposed Model

The proposed two-stage fall detection model consists of four major components: Frame extraction, Pre-processing, Rule-based fall detection module, and CNN–LSTM-based classification module. Fig. 1 depicts the overall workflow of the proposed model.

3.1 Frame Extraction

Each video in dataset is split into individual frames at its native frame rate. Let a video sequence be denoted as:

$$V = \{F_1, F_2, \ldots, F_n\} \tag{1}$$

Where, F_t denotes frame at time stamp t, and n total number of frames in the video.

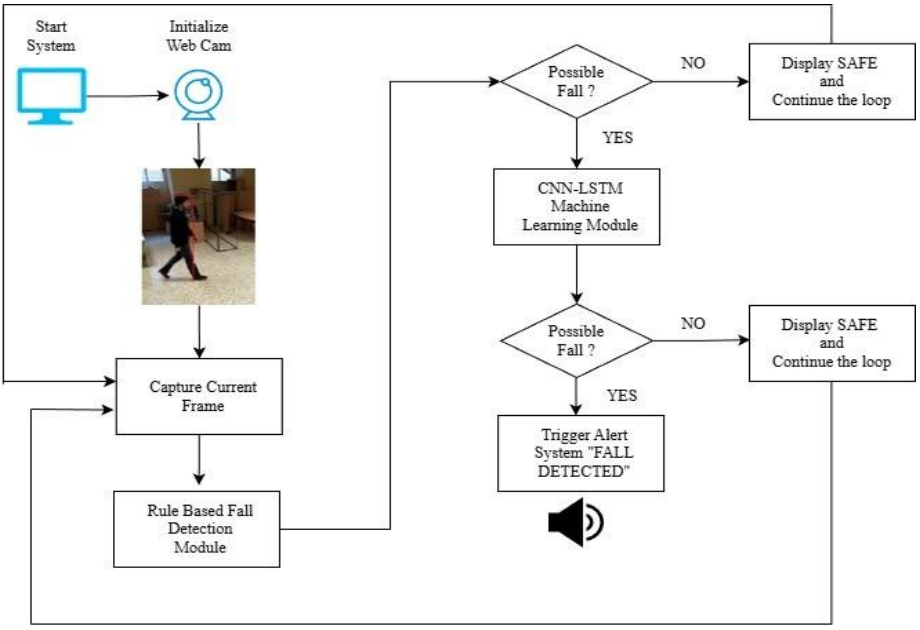

Fig. 1. The proposed two-stage fall detection model.

3.2 Preprocessing and Localization

The Le2i [8] dataset provides RGB videos at 25 FPS with 320×240 resolution. For each frame, the dataset includes bounding-box details such as height, width, and center coordinates. During training, these bounding boxes are used directly for person localization. In real-time inference, a person detector such as YOLO [17] is utilized. Each frame is resized to 64×64 to create normalized frame inputs and bounding box is cropped. A sequence length of 16 frames is used to capture short-term temporal dynamics.

3.3 Rule-Based Fall Detection Module (Stage 1)

The rule-based module as shown in fig. 2, is designed to capture rapid dynamic changes indicating a potential fall based on:

- Vertical velocity of the centroid
- Aspect ratio changes of the bounding box (standing → lying)
- Area change indicating sudden collapse
- Temporal consistency over consecutive frames

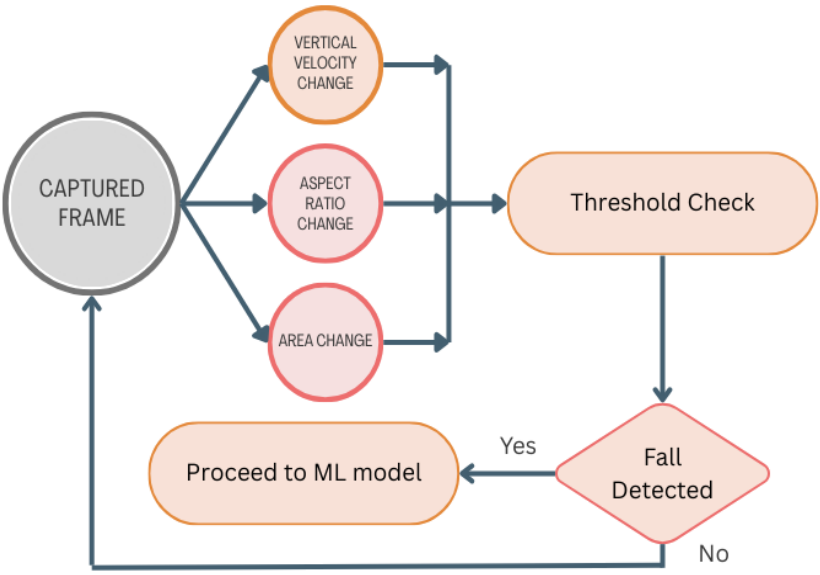

Fig. 2. Rule-based feature extraction and decision flow

Let $c_t = (x_t, y_t)$ be the centroid at time t. Then normalized vertical velocity is computed as:

$$V_t = (y_t - y_{t-1})/H \tag{2}$$

Where:
y_t = vertical centroid position in the current frame,
y_{t-1} = vertical centroid in the previous frame,
V_t = normalized vertical velocity,
H = frame height

If:

$$V_t > \theta_v \tag{3}$$

This means that the person is falling down quickly
For each frame:

$$AR_t = h_t / w_t \tag{4}$$

Where:
h_t = bounding box height,
w_t = bounding box width
During a fall:
Height decreases,
Width increases,
Aspect Ratio becomes smaller

We assess:

$$AR_t / AR_{t-1} < \theta_{AR} \tag{5}$$

If true, posture is collapsing, hence, person is falling.

Once fall is detected, the corresponding frame sequence is forwarded to the deep learning module for verification.

3.4 CNN-LSTM Machine Learning Module (Stage 2)

To enhance accuracy in ambiguous cases, a two-stage deep learning classifier is used. A CNN encoder extracts spatial features from each cropped frame and the LSTM network models temporal dynamics across the sequence. The CNN includes convolutional layers, ReLU activation, and max-pooling, creating a dense feature vector. The LSTM processes the sequence and predicts fall or non-fall through a sigmoid-activated dense layer.

Spatial Feature Extraction using CNN. From each preprocessed frame, spatial features are extracted using a CNN. The CNN transforms each input frame into a compact feature vector that captures posture and appearance information.

Formally, for each frame F_t, the CNN produces a representation of feature:

$$F_t = CNN(F_t) \tag{6}$$

Temporal Modeling using LSTM. A Long Short-Term Memory (LSTM) network is used to process the extracted feature sequences in order to capture temporal dependencies across successive frames. The LSTM learns discriminative temporal patterns linked to falls by modeling how human motion changes over time.

At each time stamp t, the hidden state of the LSTM is updated as:

$$h_t = LSTM(f_t, h_{t-1}) \tag{7}$$

Where, h_t represents the hidden state that is encoding temporal information upto time t.

Sequence-Level Classification. To determine class probabilities, the sequence-level representation is run through a fully connected layer and then a softmax activation function.

$$\hat{u} = softmax(Wh_T + b) \tag{8}$$

Where, W and b denote the learnable parameters of the classifier.

4 Result and Discussion

4.1 Dataset

Experiments are conducted using the Le2i [8] dataset containing videos of staged falls (forward, balance loss, from sitting). The videos are recorded in four distinct indoor scenes: home, office, lecture room, and coffee room, with varying lighting, clothing, and textures. Each video is accompanied by a ground-truth file specifying fall start-end frames and bounding-box coordinates for each frame.

4.2 Evaluation Metrics

To evaluate the performance of the proposed model, we have computed Accuracy, Precision, and Recall (Sensitivity) [18].

4.3 Baseline Methods for Comparison

We have compared the proposed two-stage model against Rule-based detection only [19], CNN–LSTM only [20], Optical Flow + SVM [21] and HOG Features + SVM [22] models from literature.

Table 1. Comparison between existing approaches and proposed model

S. No.	Model	Accuracy (%)	Precision (%)	Recall (%)
1.	Optical Flow + SVM	94.7	93.4	95.1
2.	HOG + SVM	91.3	89.2	90.1
3.	Rule Based Only	88.4	82.1	85.6
4.	CNN-LSTM Only	96.1	95.6	96.8
5.	Proposed Model	97.2	98.1	97.5

4.4 Discussion

As illustrated in table 1 and shown in Fig. 3, the proposed two-stage model demonstrates, reduced ML inference load (~40% fewer sequences processed by CNN-LSTM), better robustness to lighting and rapid motion, lower false positives from non-fall activities and higher temporal consistency in classification.

The introduction of the rule-based pre-filtering phase is shown to be effective in removing a significant number of frames that are then processed by the deep learning-based model. Around 60-65% of the non-fall frames are filtered out at the rule-based phase. This is the advantage of the proposed model over the state-of-the-art methods

where the deep models process all the frames irrespective of their relevance to the motion.

The proposed two-stage fall detection scheme clearly verifies that a combination of a rule-based pre-filtering step with a confirmation step using a deep learning approach is an effective and computational resource-efficient scheme. The rule-based scheme efficiently removes a vast number of frames that are not fall instances at an initial phase, which helps minimize unnecessary processing of the deep learning module. This is beneficial since it decreases the computational latency of the scheme with minimal loss of fall instances. Moreover, the subsequent confirmation step using a CNN efficiently suppresses false positives triggered by sudden non-fall activities.

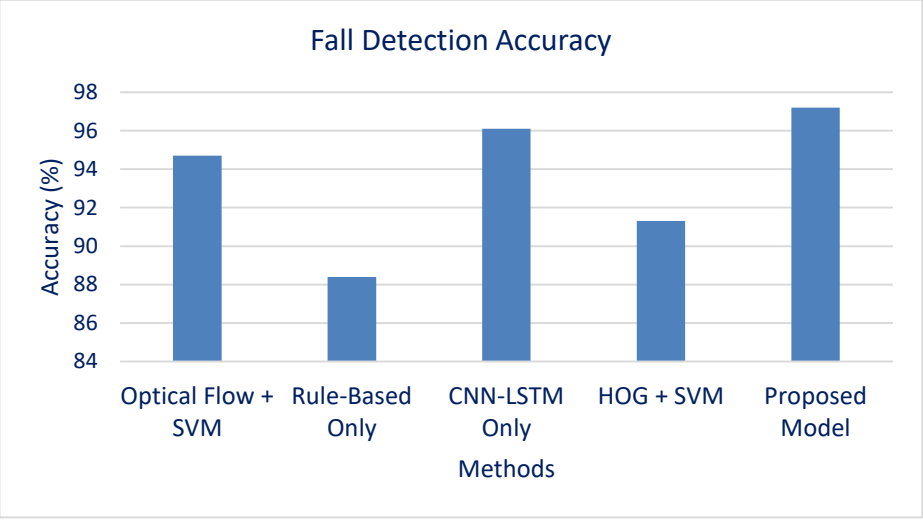

Fig. 3. Comparison of the proposed model with existing models

5 Conclusion

This work proposes a two-stage fall detection that embodies both rule-based heuristics and deep learning CNN-LSTM model. The proposed model also handles fall detection under uncontrolled indoor environments. The rule-based logic provides a fast, interpretable, and lightweight detection. While the LSTM-based classifier ensures the robustness in fall scenarios. The two-stage model attains superior performance compared to standalone, rule-based or deep learning approaches. Rule-based logic alone provides lightweight architecture but lacks robustness and deep learning approach alone is robust but is too heavy to be used for real-time deployment. Future work involves extending the model to multi-person environments, model deployment on real-time embedded platforms.

References

1. J. Gutiérrez, V. Rodríguez, and S. Martín, "Comprehensive review of vision-based fall detection systems," Sensors, vol. 21, no. 3, p. 947, 2021.
2. H. Sun et al., "Long-lie after falls and its medical consequences in elderly populations," PubMed, 2022.
3. X. Wang, J. Ellul, and G. Azzopardi, "Elderly fall detection systems: A literature survey," Frontiers in Robotics and AI, vol. 7, p. 71, 2020.
4. Y. Delahoz and M. Labrador, "Survey on fall detection and fall prevention using wearable and external sensors," Sensors, vol. 14, no. 10, pp. 19806–198042, 2014.
5. M. Kangas, A. Konttila, I. Winblad, and T. Jämsä, "Fall detection using accelerometers in real-world scenarios," Journal of Telemedicine and Telecare, vol. 14, no. 7, pp. 367–373, 2008.
6. A. Mubashar et al., "Vision-based human fall detection using body geometry and pose estimation," Sensors, vol. 21, no. 3, p. 947, 2021.
7. F. Ahmed et al., "Robust vision-based fall detection using multi-stream 3D CNNs," Applied Sciences, vol. 13, no. 12, p. 6916, 2023.
8. I. Charfi, J. Mitran, J. Dubois, M. Atri, and R. Tourki, "Optimised spatio-temporal descriptors for real-time fall detection: Comparison of SVM and Adaboost based classification," Journal of Electronic Imaging, vol. 22, no. 4, Oct. 2013.
9. M. Kangas, A. Konttila, I. Winblad, and T. Jämsä, "Determination of simple thresholds for accelerometer-based fall detection," IEEE Engineering in Medicine and Biology Magazine, vol. 27, no. 1, pp. 66–70, Jan. 2008.
10. Z. Zhang, J. Wang, and J. Liu, "A review on vision-based fall detection," Healthcare, vol. 7, no. 2, p. 37, 2019.
11. N. Ali, I. S. Bajwa, M. K. Sabir, and T. Hussain, "An intelligent real-time human fall detection system based on deep learning for healthcare monitoring," Journal of Ambient Intelligence and Humanized Computing, Springer, 2020.
12. J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 2625–2634.
13. S. Chaudhuri, L. Borzi, and A. Ramaswamy, "Skeleton-based human activity recognition using CNN-LSTM networks," Pattern Recognition Letters, vol. 144, pp. 128–135, 2021.
14. F. Ahmed et al., "Robust vision-based fall detection using multi-stream CNNs and sensor fusion," Applied Sciences, vol. 13, no. 12, p. 6916, 2023.
15. Z. Zhang, Y. Li, and J. Wang, "Vision transformer based fall detection with temporal attention for indoor surveillance," Sensors, vol. 24, no. 3, pp. 1–17, 2024.
16. Y. Li, X. Chen, and H. Zhou, "Cross-environment evaluation of vision-based fall detection systems," IEEE Access, vol. 12, pp. 45621–45633, 2024.
17. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.
18. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006.
19. S. Katamneni and G. Jeyakumar, "A computer vision based fall detection technique for home surveillance," in Advances in Intelligent Systems and Computing, Springer, 2020.
20. S. S. Jeong, N. H. Kim, and Y. S. Yu, "Fall detection system based on simple threshold method and long short-term memory: Comparison with hidden Markov model and extraction of optimal parameters," Applied Sciences, vol. 12, no. 21, p. 11031, 2022.

21. S. Chhetri et al., "Deep learning for vision-based fall detection system: Enhanced optical dynamic flow," Computational Intelligence, vol. 37, no. 1, pp. 578–595, 2021.
22. X. Kong, X. Meng, Z. Li, and Z. He, "A HOG-SVM based fall detection IoT system for elderly persons using deep sensor," Procedia Computer Science, vol. 147, pp. 276–282, 2019.