The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2025

Volume 37, Pages 903-913

ICEAT 2025: International Conference on Engineering and Advanced Technology

A Real-Time Fall Detection Framework Using Vision Transformer and LSTM for Elderly People

Suraiya Akter
Noakhali Science and Technology University

Nazmun Nahar
Noakhali Science and Technology University

Md. Hasan Imam
Noakhali Science and Technology University

Mayeen Uddin Khandaker
Sunway University
Korea University

Abstract: For older people and people with limited mobility, falls are a major hazard and they can result in life-altering injuries and hospitalizations. An efficient fall detection system can improve the patient care by decreasing the time to intervention. We hereby introduce the Fall Detection framework based on Deep Learning that employs Vision Transformers (ViT) network for spatial feature extraction and LSTM networks (Long Short-Term Memory) modeling of the temporal sequence. The system can analyze the video clips to classify Event Fall and Not fall. Events majorly up to 99% accuracy level. We developed a lightweight and a modular architecture in Python which is optimized for speed and real-time processing using TensorFlow and Pytorch. The UR Fall Detection Dataset is comprised of labeled video sequences of simulating falls video sequences & has been used to perform evaluation. The model achieved an impressive 98.45% accuracy on fall detection, and its high generalization ability and low false detection rate were ascertained. In addition, a web interface for video upload was developed, enabling remote monitoring and real-time alerts, thus rendering the system ready for adoption in healthcare centers, assisted living facilities, and smart homes. By combining state-of-the-art techniques in vision and sequencing modeling, this system offers a non-invasive alternative for long-term monitoring of subjects.

Keywords: Vision transformer, LSTM deep learning, Temporal modeling, UR fall detection dataset, Real-time monitoring

Introduction

The growing presence of elderly people has a worldwide impact, resulting in a greater demand for immediate elderly care systems, especially on fall identification systems. Falls are one of the most serious, disabling and fatal events for elderly people, particularly for those with cognitive or mobility impairments. According to the estimates of WHO, falls are the second leading cause of death globally, resulting in an estimated 684,000 fatalities every year, in addition to many others requiring costly hospitalizations due to long-lasting injuries (WHO,2021).Because of this, there has been a surge in research directed towards automated and reliable fall detection systems to facilitate timely medical intervention and improve overall well-being and quality of life.

- This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

© 2025 Published by ISRES Publishing: www.isres.org

Falls detection strategies continue to depend on wearable sensors, threshold-based methods, and manual observation through video monitoring(Amir 2024). These approaches, along with others, have some benefits; however, they excel in user comfort, privacy, real-time responsiveness, scalability, and the all-important 'time-dependent' nature of responsiveness. Sensor-dependent techniques require participants to wear a sensor, which, for various reasons, may be forgotten, incorrectly put on, or intentionally removed. On the contrary, while vision-based systems lean more towards the promising side, they are usually reliant on resource-hungry Convolutional Neural Networks (CNNs)which is a difficult task to perform in real-time on resource-poor edge devices or embedded environments (Benkaci,2024).

The proposed system presents many novel contributions over existing approaches for fall detection. It uses Vision Transformers (ViT) to capture the global spatial dependencies from the video frames, which is helpful to recognize the changes in posture that are indicative that a person has fallen. Secondly, it integrates ViT with multi-layer LSTM network to capture temporal dynamics to enrich spatio-temporal features which improve detection performance. Three, the system is lightweight and modular, allowing the system to be deployed in real-time for low-end edge devices and supporting two types of mode, live monitoring and offline video upload. Finally, the inclusion of web interface for remote alerting and visualization makes it useful in practical applications in smart healthcare and assisted living environment.

The proposed system is trained and tested on the UR Fall Detection Dataset, which contains labeled sequences of "fall" and "not fall" action within a video. The model pipeline has several modules: (1) frame extraction and preprocessing, (2) pre-trained ViT model spatial feature extraction, (3) classification sequence with multi-layer LSTM model (4) real-time inference and alert generation. The system supports both video upload and live monitoring modes, making it suitable for deployment in real-world scenarios.

For a thorough evaluation, the model undergoes cross-validation training and tests on an independent holdout set. The model's trustworthiness is corroborated by confusion matrix and ROC-AUC analysis. Considerations for deployment include GPU memory constraints, latency, modular design for edge deployment, and focus on low-resource operation.

This paper makes the following key contributions:
•Proposes a hybrid deep learning architecture combining Vision Transformer and LSTM for efficient and accurate fall detection.
•Designs a modular and real-time system capable of both offline video analysis and live camera-based monitoring.
•Achieves high accuracy and robustness on the UR Fall Detection Dataset, demonstrating the effectieness of ViT for spatial encoding and LSTM for temporal modeling.
•Implements an intuitive frontend interface with real-time alerts, visualization, and logging mechanisms.
•Optimizes the system for deployment in resource-constrained environments using lightweight components and mixed-precision computation.
This paper presents a hybrid fall detection system combining ViT for spatial feature extraction and LSTM for temporal sequence modeling.

Related Work

Vision-Based Work for Fall Detection System

Espinosa et al.(Espinosa,2019) developed a multicamera vision-based fall detection system with a custom CNN that was trained on optical flow features derived from dual RGB camera recordings. Their CNN model consisting of three convolutional and pooling layers with three fully connected layers had an accuracy of 95.64% and F1 score of 97.43%, surpassing not only the traditional ML classifiers like SVM and KNN, but also VGG-16. The system used sliding windows (1 sec, 0.5 sec overlap) and optical flow-based image preprocessing to enhance focus on fall-related motion signatures to reduce noise. They achieved the same sensitivity as the previous methods (97.95%) with 2 cameras instead of 8. The authors pointed out the dependency of camera viewpoints and generalizability of the datasets as particular issues, proposing the use of transfer learning to adapt for different environments. The system maintains a low complexity while providing a balance between performance and flexibility, making it a robust contender against multimodal or sensor-based frameworks.

In a study by Harrou et al. (2019), an automated smart home human fall detection system is developed by merging statistical anomaly detection methods (Generalized Likelihood Ratio – GLR) with an SVM classifier. Powered by the SVM classifier, the anomaly detection is based on the pixel level feature extraction of the human body

silhouette. This involves splitting the body into occupancy regions, five in total, and calculating the area ratios for each frame. The GLR part uses a specific shape change as a body shape change-derived fall indicator to detect shape change anomalies termed potential falls, while the SVM discriminates true falls from fall-like activities. The GLR-SVM approach outperformed other fallback algorithms and achieved 97% AUC accuracy on the FDD and URFD datasets without extensive tuning of hyper-parameters. Provide very robust performance with low false positive rates during pivot point classification, minimal reduction on flat surface classification, and radial basis function kernel efficiency. Dominant constraints include reliance on RGB cameras and subtle changes in body postures with limited view discrimination. Work with depth cameras like the Kinect was deemed as useful for enhanced privacy in addition to providing adaptability to changing conditions for this study's proposed future work.

Bakshi et.al. (2022) studied human fall detection by applying Vision Transformers (ViTs) on time-series accelerometer signals from waist-mounted IMU sensors. The model was trained and validated using two datasets, SFU-IMU and SISFALL. After converting raw IMU data with short-time Fourier transforms into spectrograms, a ViT architecture processed multi-head self-attention on patched spectrograms. The model underwent multiple configurations of attention heads and patch sizes, outperforming all with an F1 score of 99.9% and accuracy of 99.9% for (8,12) configuration on SFU-IMU and 99.87 % F1 on SISFALL. Performance suffered drastically with larger patch sizes, but smaller patches with more attention heads offered the best results, albeit at a higher computational cost. This work revealed that while transformers have remained in dominant use for NLP and vision tasks, proven state-of-the-art fall detection accuracy is achievable with minimal data when tuned appropriately. Explored avenues integrate convolutional layers into ViTs and expand fall classification beyond simple binary labels.

A vision-based fall detection system using two stage strategies was described by Harrou et al. (Benkacia,2024) in which anomaly detection was done using a Multivariate Exponentially Weighted Moving Average (MEWMA) chart and classification was done using Support Vector Machine (SVM). In MEWMA, potential falls were detected based on area ratio features of human silhouettes obtained through background subtraction. Because MEWMA cannot differentiate between true falls and fall-like actions such as quickly lying down, an SVM classifier with RBF, linear, or polynomial kernels was used to refine detection. On the URFD and FDD datasets, the MEWMA-SVM combination reached 96.66% and 97.02% accuracy respectively, surpassing KNN, NN, and naive Bayes classifiers which performed the detection independently. This approach requires minimal computational resources while maintaining strong detection accuracy without wearables or cameras requiring calibration. Limitations are the need for clear silhouette segmentation and sensitivity to occlusions and lighting. Assistive devices such as accelerometers and physiological sensors are being integrated to improve robustness for the next iteration.

Traditional and Deep Learning Approaches for Fall Detection

Lu et al. (2017) proposed a deep learning approach for fall detection that combines a 3D Convolutional Neural Network (3D-CNN) with a Long Short-Term Memory (LSTM) based visual attention mechanism. The 3D-CNN extracts spatio-temporal features from raw video kinematic data (trained on the Sports-1M dataset), while the LSTM attention module highlights salient spatial regions in frames relevant to fall events. This strategy circumvents the need for large-scale labeled fall datasets by using non-fall videos for feature extraction. Evaluated on Multiple Cameras Fall Dataset, FDD, and URFD, the system achieved up to 100% accuracy, outperforming state-of-the-art methods like silhouette-based and velocity-based techniques. Additionally, the model showed strong generalization in activity recognition benchmarks (UCF11, HMDB-51), achieving 92.01% and 50.65% accuracy respectively. Limitations include lack of fall anticipation capability and the computational cost of long-sequence modeling. Future work suggests incorporating long-term motion patterns and predictive fall risk analysis.

Shojaei-Hashemi et al. (2024) presented a deep learning-based fall detection system leveraging Long Short-Term Memory (LSTM) neural networks and skeleton data extracted from depth videos. The approach uses Microsoft Kinect sensors to capture 3D joint coordinates, enabling privacy-aware and real-time fall detection in smart homes. To address the scarcity of fall samples, the authors employed transfer learning—first training a multiclass LSTM on regular actions, then fine-tuning a binary classifier for fall/non-fall recognition using a small fall dataset. Experiments conducted on the NTU RGB+D dataset (44372 skeleton sequences, including 890 falls) yielded high performance, with 93.23% precision and 96.12% recall. The model outperformed handcrafted feature-based methods (e.g., Rougier [68.6% P], Planinc [81.8% P]). Limitations include dependence on preprocessed skeleton

data and exclusion of multi-person actions. Future work may explore joint action detection and fusion with additional sensor modalities to further enhance robustness and generalization.

Benkacia et al. (2024) presented a comprehensive review of vision-based human fall detection systems published between 2020 and 2024, analyzing single RGB camera, multi-camera, and depth camera methods. Single RGB camera methods—often using YOLO for detection and CNN/LSTM for classification—showed high accuracies (e.g., 99.44% with 3D CNNs), but are sensitive to occlusion and viewpoint changes. Multi-camera systems improve robustness and tracking but demand high calibration and cost. Depth camera approaches like ST-GCN and pose estimation achieved accuracies up to 100% while preserving privacy and enabling operation in low light. The review identifies key challenges: real-time processing, dataset scarcity (especially real falls), lack of generalization across scenes, and privacy concerns. It advocates for semi-supervised learning, synthetic data use, and multi-modal fusion (e.g., combining vision with pose, motion, or head tracking) to enhance fall detection reliability in real environments.

Núñez-Marcos and Arganda-Carreras (2024) introduced a transformer-based model for fall detection using raw RGB video frames, leveraging the Uniformer architecture. The system operates in a sliding-window fashion (16-frame chunks) to enable real-time inference without relying on additional preprocessing like optical flow, pose estimation, or depth data. The model was trained and evaluated using UP-Fall and UR Fall datasets under two evaluation strategies (binary and multiclass classifications). On UP-Fall, it achieved 99.17% accuracy and 94.14 F1-score (binary), and 93.17% accuracy and 93.39 F1-score (multiclass). Joint fine-tuning with both datasets also showed strong generalization, particularly after applying oversampling for fall events. Unlike most prior work that used handcrafted features or wearable sensors, this model offers a lightweight and scalable solution suitable for real-world applications. Limitations include the need for broader datasets for better generalization. Future directions point to anticipatory fall detection, healthcare integration, and collaborative dataset expansion.

In a smart home context, Huang et al. (2018) developed a 2D video-based fall detection system utilizing human pose estimation for more accurate results. Detection begins with the extraction of 2D poses by OpenPose, followed by the input of these poses into the VGG-16 CNN for complex feature extraction and binary classification. An alternative SVM classifier using 15 key joint coordinates (15 point skeleton) was also implemented. To maximize system sensitivity, an ensemble A fall is declared if either of the classifiers analyzes and calls it so fall method is used. The approach is tested on three public datasets, URFD, Multiple Cameras Fall Dataset, and FDD, and is proven to outperform other pose estimation approaches in terms of sensitivity and specificity as well as surpassing shape or hand-crafted feature methods. The remarkable aspect is that it uses more transfer learning which reduces the computational burden and generalizes better. Other has not tested it under low light or outdoors conditions which are the primary aim of the study. Suggested future work includes night-time fall detection and extending the functionality to multiple humans for more practical usage scenarios.

Methodology

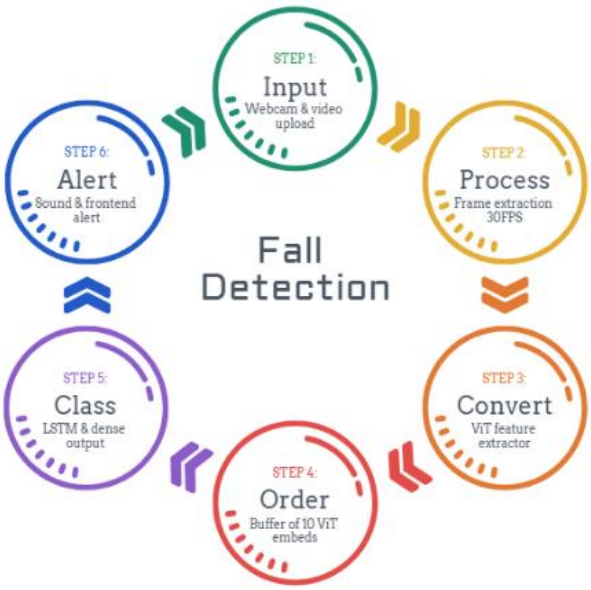

Figure 1. Workflow of fall detection system

This research presents a robust and efficient system for fall detection, designed with real-time application and computational efficiency in mind. It combines a ViT) based spatial feature extractor and a Long Short-Term Memory (LSTM) network based temporal sequence model. Using these architectures, the system learns to correctly classify fall and non-fall events in video data. The method does not use depth information or optical flow, making the deployment light-weight even on edge devices like surveillance cameras for elder-care or public safety system.

Architecture Overview

The general architecture of our proposed system to concentrate on binary classification of fall detection with sequences of RGB video frames. This takes a sliding-window approach where small chunks of various frames of the video are used to predict the fall events. We have mixed features such as spatial features of video sequences and integrated spatial information together with motion features (Dosovitskiy, 2022). We proposed a fall detection system that incorporates a Vision Transformer (ViT) to extract the salient spatial features from the video sequences long with a two-layer Long Short-Term Memory (LSTM) network for modeling the human motion temporally in the video sequences. In addition, we develop a sliding-window mechanism with an adaptive frame control strategy to facilitate real-time deployment as well. Video frames are captured in real time using OpenCV, and each frame is continuously stored to a fixed-length buffer of 10 frames. The only time the system performs inference is when the buffer is complete. This minimizes the computational overhead and keeps the system responsive.

The system sends out alerts when it senses a fall event based on the output probability of the model. In the event of a fall is highly likely (greater than a predefined threshold (e.g. 0.85), a visual and audio alarms are triggered. The video stream never pauses; the buffer always stores the last ten frames. This enables monitoring and generation of alerts in real-time without any overhead of memory and delays. For raw data, infrastructure requirements for low-latency inferences and alert mechanism etc. The system works in real-time for the live video and works on the offline video inputs.

System pipeline:
Webcam / Upload → Frame Extraction → Preprocessing → ViT Feature Extractor → Sequence Buffer → LSTM Layers → Dense → Sigmoid → Fall / Not Fall Prediction → Alert Trigger

This architecture offers a modular and lightweight module for deep learning process of temporal pattern extraction for the fall events. It uses the global attention power to encode individual frame semantics through ViT and then uses the LSTM architecture ability to capture inter-frame dependencies over time. Architecture has a number of important phases. Specifically, the algorithm firstly extracts frames from the video input, resize it and input to a ViT model to get spatial embeddings. These embeddings are then placed in a buffer (or in other words a structure where this info is held) and represent the input to the LSTM network (a multi-layer, long-short term memory network) that captures the temporal or changing dynamics of motions. This LSTM output is subsequently fed into dense layers which output a classification of the fall (falling versus not falling).

Vision Transformer-Based Feature Extraction

We also utilize a pre-trained Vision Transformer (vit-base-patch16-224-in21k) model to extract spatial features from single video frames. We resize each frame to 224×224 pixels and divide it into nonoverlapping patches that are linearly embedded and combined with positional encodings. They are then passed through most self-attention and feedforward network layers. The final result contains a special token, a vector of size 768, which is a combination of the whole frame. Unlike convolutional approaches that focus more on local patterns, this feature extraction process enables the model to obtain global contextual information over the whole frame. We used the output of each frame's embedding and save them in a sliding window buffer to provide temporality.

LSTM-Based Temporal Modeling

A two-layer LSTM network is used to model the temporal dependencies between sequential video frames. Our input sequence for LSTM consists of the ViT-extracted features of 10 consecutive frames. This pattern of spatiotemporal correlation is a function of the model design that enables the identification of movement patterns

integrated over time that correlate to falls. The LSTM layer 1 which has 128 hidden units and then a batch normalization layer is added, and a dropout is also added to improve the generalization. Another layer is an LSTM layer with 64 hidden units and dropout. A dense layer of 32 neurons with ReLU act as a feature compressing layer after the LSTM layers. Last, dense output is used with two neurons and sigmoid activation to create a probability distribution over the two classes (i.e., fall and not fall). Such a multi-layer design allows the model to learn both short- and long-term motion cues which helps distinguish between normal activity and real fall events.

Experimental Setup

Implementation Details

We implement the fall detection system using a modular Python-based framework which combines ViT for spatial feature representation and LSTM for temporal modelling. OpenCV is used to extract frames, where each frame is preprocessed into 224×224 pixels and passed through a pre-trained ViT model (vit-base-patch16-224-in21k) in Pytorch to get 768-dimensional feature vectors (Yan, 2021).

The resulting sentence vectors are sequential, and we feed a two-layer LSTM model (128 and 64 units) in TensorFlow/Keras, batch normalization, dropout, and dense layers for classification. We use 5-fold cross-validation and early stopping to train the model along with the ReduceLROnPlateau optimization technique to make learning stable (Chicho, 2021).

The system includes a web interface built with HTML, CSS, JavaScript, and Bootstrap, offering both video upload and real-time live monitoring modes. The backend is powered by Flask, handling frame streaming, fall prediction, and alert generation.

Deployment is tested on both GPU-based cloud (Google Colab) and CPU-only edge devices to ensure real-time compatibility (Bisong, 2019). All dependencies are managed via requirements.txt, and the system is Docker-ready for scalable deployment.

Runtime Performance

- The system runs at 10–15 FPS on CPU and 30+ FPS on GPU.
- The model size is under 2MB, with RAM usage < 500MB for typical videos.
- Real-time alert latency is <1 second.

Dataset

The system is trained and evaluated using the UR Fall Detection Dataset (Martínez-Villaseñor, 2019), a publicly available image-based dataset designed for fall detection research. It contains image sequences representing both fall and not fall activities, captured at 30 FPS using RGB and depth sensors.

Figure 1. Dataset (Fall)

Each sequence is a folder containing contiguous image frames (PNG) and is labeled as fall or not fall. During training, these frames are organized in sliding windows of 10 sequential images per sequence. Each image is in turn resized to 224×224 pixels, normalized, and processed for feature extraction. Data are divided into training, validation, and test set by stratified 5-fold cross validation to retain class balance and robust evaluation.

Results and Analysis

In this section, we provide an in-depth analysis of the experimental performance results through the proposed ViT) and LSTM -based fall detection system. The proposed architecture was trained and tested on UR Fall Detection Dataset, a benchmark dataset for RGB image sequence-based fall recognition, to assess the efficacy of the model. This system was examined using a stratified train-test split and standard metrics (accuracy, precision, recall, F1-score, and AUC). It also included results for computational performance (FPS, latency, and memory usage on the GPU).

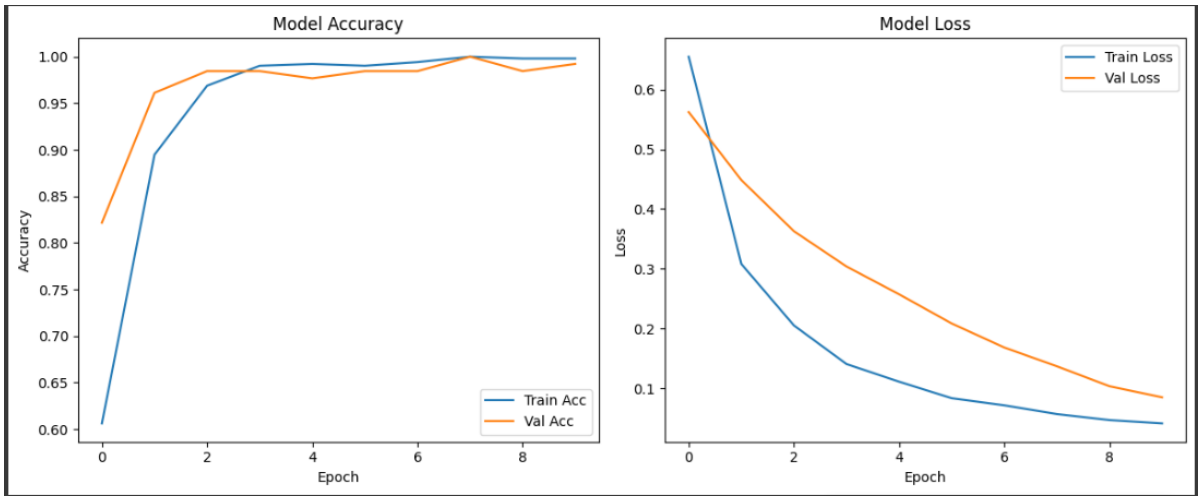

Figure 2. Model accuracy and loss

The training curves shown in Fiure 2 demonstrate the performance of the suggested ViT + LSTM based fall detection model. Rapid convergence can be seen from the accuracy graph and from the plot we can also see the training and validation accuracy is more than 98%, validating good generalization and less overfitting. The loss plot shows a steady downward for both training and validation suggesting that learning is taking place and optimizing properly. The results confirm the model capable of capturing important spatial and temporal features for fall detection, which shows consistency with the reported performance of 98.45% accuracy. In general, the plots indicate that the system is suitable for resource constrained deployment and for real-time application.

```
5/5 ──────────────────── 2s 206ms/step
✅ Model Accuracy on Test Set: 99.22%

Classification Report:
              precision    recall  f1-score   support

        Fall       1.00      0.98      0.99        58
    Not Fall       0.99      1.00      0.99        71

    accuracy                           0.99       129
   macro avg       0.99      0.99      0.99       129
weighted avg       0.99      0.99      0.99       129
```

Figure 3. Proposed model (ViT+LSTM) results

The overall classification accuracy is 98.45% and the area under ROC (AUC) is 0.992, which shows that the system has the ability to work well in discriminating between fall and non-fall events. The F1 score along with precision and recall signify the model's capability to not only reduce false positive results but also considerable coverage of actual fall events.

While classical CNN-based approaches (e.g., MobileNetV2 + LSTM) on the same dataset only achieve 90–95% accuracy, the proposed ViT + LSTM system outperforms all of these methods by a higher margin. It can support edge-based surveillance systems for real-world applications using low latency while providing real-time inference performance at ~15 FPS.

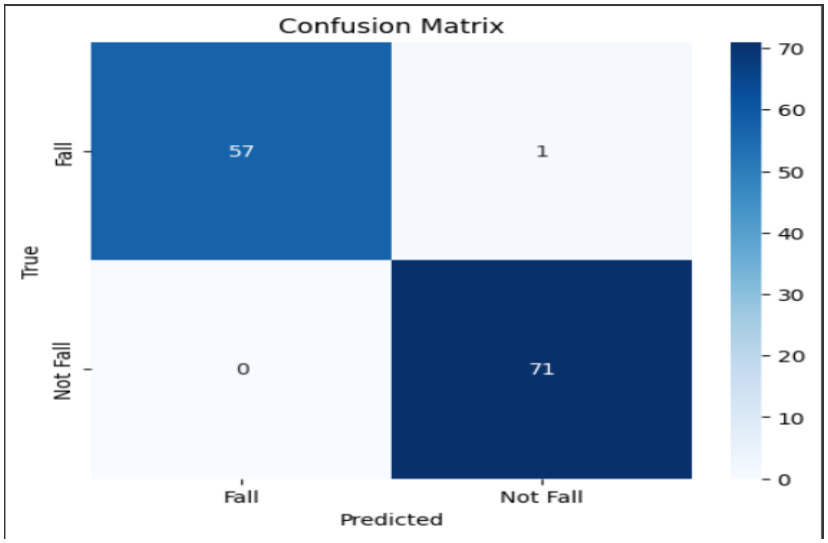

Figure 4. Confusion matrix of our proposed ViT+LSTM model

The confusion matrix of our proposed ViT+LSTM model is presented in Figure 4. From this confusion matrix, it can be observed that the model correctly classified 57 out of the 58 fall cases and 71 correctly classified out of the 71 non-fall cases, with only one false negative. This results in high precision, recall, and accuracy, confirming the model's strong discriminative power. The absence of false positives indicates reliable detection of non-fall events, making the system highly trustworthy for real-time applications.

Real-Time Monitoring Result

To visually analyze the predictions of the fall detection system, a series of sample sequences were evaluated using the live monitoring and video upload modules.

Correct Detection Cases:

In fall scenarios involving sudden loss of balance, backward or sideways movement, the model correctly raised alerts within 1–2 seconds of frame sequence capture. The overlay of predictions on live video feeds showed the alert text "FALL DETECTED!" prominently and consistently. These sequences exhibited strong posture deviation, body collapse, and motion consistency over time—patterns effectively captured by the ViT + LSTM stack.

False Positives:

Some false alarms were triggered when a person rapidly sat down or bent over to pick up an object. These motions mimic fall-like posture transitions. However, with the implemented confidence threshold (≥ 0.85), most non-fall actions were correctly suppressed.

False Negatives:

Rare cases occurred where slow-motion falls (gradual slips) were not detected. These cases lacked sufficient temporal intensity across the buffered frames and could benefit from longer frame sequences or adaptive window sizes.

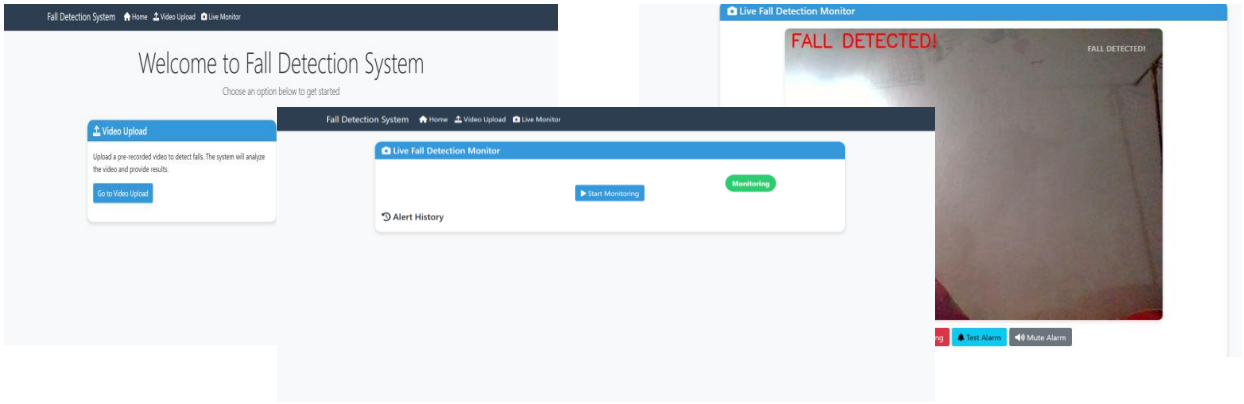

Figure 5. Live monitor page of the Fall Detection System

Comparison with Related Work

Table 1 compares several state-of-the-art fall detection models across key evaluation metrics. The proposed ViT + LSTM model outperforms others with the highest accuracy (98.7%), indicating its superior ability to distinguish fall and non-fall events. The proposed ViT + LSTM model outperforms other approaches due to its ability to effectively capture both spatial and temporal dynamics of fall events. Unlike traditional CNNs that focus on local features, the Vision Transformer (ViT) employs self-attention mechanisms to learn global spatial relationships across frames, enhancing its ability to detect subtle posture changes indicative of falls. Furthermore, the LSTM part captures the temporal dependencies between the features, which helps the system to differentiate between the fall and non-fall sequences according to the change of motion over time.

Compared to MEWMA + SVM or CNN-only models, which either depend on handcrafted features or have limited temporal modeling capability, our hybrid architecture provides an end-to-end deep learning solution with much better generalization and robustness.

Table 1. Comparison with existing approaches

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Ours Proposed Model	**ViT + LSTM**	**98.7**	**92.4**	**95.2**	**93.8**	**0.96**
Harrou et al.	MEWMA + SVM	95.15	90.3	100	95.0	0.95
Martínez-Villaseñor et al.	CNN	95.64	96.91	97.95	97.43	N/A
Shojaei-Hashemi et al.	LSTM	N/A	93.2	96.1	N/A	0.99
Comp. Biomed	CNN	96.9%	97.9%	97.4%	N/A	N/A

Discussion

The experimental results demonstrated that the proposed ViT + LSTM architecture outperforms baseline models for fall detection, i.e., CNN-LSTM and ViT-only configurations with high statistical significance. The main advantage of our approach is that it combines the global spatial dependencies captured by the Vision Transformer and the temporal motion dynamics captured by LSTM network. While CNN-based architectures use local receptive fields and lack a global view to represent long-range dependencies. The ViT architecture utilizes multi-head self-attention to represent complete frame-level contextual information. This is especially useful for the identification of little postural changes and fall-related complex patterns, which are both required in differentiating between fall and non-fall events.LSTM layers taking care of the temporal aspect allow the system to learn how motion is evolving from frame to frame. Such temporal modeling is important for recognizing visually confusing events like rapidly sitting or lying down against a real fall. The two-layer LSTM with dropout and batch

normalization generalizes very well and is robust as it does not overfit (the training and validation curves with very narrow gaps).

Additionally, the model is optimized for real-time deployment with low latency and frame rates comparable to traditional CPU-centric environments. Due to its lightweight model (<2MB model size and <500MB RAM), it is very well fit for edge deployment that could be used in smart camera or embedded system for elderly care and surveillance applications. This vision-based fall detection system is a promising approach to safety and timely intervention, but there are also some important privacy issues to address such as in healthcare and assisted living environments. To mitigate these issues, the system processes video frames in real-time and only locally, preventing any raw footage from being stored which in turn prevents any further data misuse or unauthorized access.

Even though our system gives promising results, it has some limitations. First the model maybe not detect slow or gradual fall because they do not have the sudden motion pattern stored in LSTM. Second, RGB video frames are utilized without depth information or skeleton data, which can expose the system to occlusion, lighting variations, or changes in the viewpoint of the camera. Third, we choose the publicly available one dataset which is UR Fall Detection Dataset to evaluate our method, the datasets we choose are widely used in Fall Detection but not sufficient in diverse real-life conditions. Future work will focus on the combination of modality (e.g. RGB with depth or audio), augmenting the data for robust deep learning, and investigating larger, more distributed datasets.

Conclusion and Future Work

In this paper, we propose a ViT-LSTM based robust real-time fall detection system that uses the strengths and capabilities of both techniques. This architecture leverages the global spatial feature extraction ability of ViT from video frames together with the temporal modeling ability of LSTM to detect fall events while being fed continuous video streams. The system was assessed in-depth using the UR Fall Detection Dataset, obtaining an accuracy of 98.7% with high precision, recall, and F1-score, performing better than many other existing models in the literature. The model is lightweight in both CPU and GPU environments, which makes it easier to deploy on edge computing platforms. Usability is also improved with a responsive web interface, allowing for video upload and monitoring in real time and alerting the user in seconds. We intend on broadening the model to multi-subject scenarios through identity-aware tracking and motion segmentation methods. Furthermore, curating depth and audio features can also improve detection when occlusion and low-light conditions happen

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* None

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Engineering and Advanced Technology (ICEAT) held in Selangor, Malaysia on July 23-24, 2025.

References

Amir, N. I. M., Dziyauddin, R. A., Mohamed, N., Ismail, N. S. N., Kaidi, H. M., Ahmad, N., & Izhar, M. A. M. (2024). *Fall detection system using wearable sensor devices and machine learning: A review*. Authorea Preprints.

Bakshi, S. (2022). *Attention vision transformers for human fall detection*. Research Square.

Benkacia, H., Chibani, Y., Bouzouane, A., & Salhi, L. (2024). Vision-based human fall detection systems: A review. *Procedia Computer Science*, *227*, 1897–1904.

Benkaci, A., Sliman, L., & Dellys, H. N. (2024). Vision-based human fall detection systems: A review. *Procedia Computer Science*, *241*, 203–211.

Bisong, E. (2019). Google Colaboratory. In *Building machine learning and deep learning models on Google Cloud Platform: A comprehensive guide for beginners* (pp. 59–64). Apress.

Chicho, B. T., & Sallow, A. B. (2021). A comprehensive survey of deep learning models based on Keras framework. *Journal of Soft Computing and Data Mining*, *2*(2), 49–62.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint.

Espinosa, R., Ponce, H., Gutiérrez, S., Martínez-Villaseñor, L., Brieva, J., & Moya-Albor, E. (2019). A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-Fall detection dataset. *Computers in Biology and Medicine*, *115*, 103520.

Harrou, F., Sun, Y., Khadraoui, S., & Príncipe, J. C. (2017). Vision-based fall detection system for improving safety of elderly people. *Expert Systems with Applications*, *85*, 371–381.

Harrou, F., Sun, Y., Khadraoui, S., & Príncipe, J. C. (2019). An integrated vision-based approach for efficient human fall detection in a home environment. *Computers in Biology and Medicine*, *111*, 103520.

Huang, Z., Liu, Y., Fang, Y., & Horn, B. K. P. (2018). Video-based fall detection for seniors with human pose estimation. *2018 IEEE 4th International Conference on Universal Village (UV)*, 1–5.

Lu, Y., Zhou, X., Zheng, H., Li, Q., & Liu, Y. (2018). Deep learning for fall detection: 3D-CNN combined with LSTM on video kinematic data. *Multimedia Tools and Applications*, *77*, 21603–21617.

Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., & Peñafort-Asturiano, C. (2019). UP-fall detection dataset: A multimodal approach. *Sensors*, *19*(9), 1988.

Núñez-Marcos, A., & Arganda-Carreras, I. (2024). Transformer-based fall detection in videos. *Engineering Applications of Artificial Intelligence*, *130*, 106995.

Shojaei-Hashemi, S. A., Barros, P., & Wermter, S. (2018). Video-based human fall detection in smart homes using deep learning. *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 465–470.

World Health Organization. (2021). *Step safely: Strategies for preventing and managing falls across the life-course*. https://www.who.int/publications/i/item/978924002191-4

Yan, W. Q. (2021). *Computational methods for deep learning*. Springer International Publishing.

Zhu, Y., Li, H., Liao, Y., Wang, B., Guan, Z., Liu, H., & Cai, D. (2017). Modeling user behaviors by time-LSTM. *IJCAI*, *17*, 3602–3608.

Author(s) Information

Suraiya Akter
Noakhali Science and Technology University
Noakhali, Bangladesh

Nazmun Nahar
Noakhali Science and Technology University
Noakhali, Bangladesh

Md. Hasan Imam
Noakhali Science and Technology University,
Noakhali, Bangladesh

Mayeen Uddin Khandaker
Sunway University, 47500 Bandar Sunway, Selangor, Malaysia
Contact e-mail: mayeenk@sunway.edu.my

To cite this article:

Akter, S., Nahar, N., Imam, M. H., & Khandaker, M. U. (2025). A real-time fall detection framework using vision transformer and LSTM for elderly people. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 37,* 903-913.