Health Abnormality Detection of Elderly using Deep Learning

Bac Nguyen Dinh[1], Nurzhigit Kurmanov[2], Sherzod Omar Hakimov[3]
[1] *International Bachelor Program in Informatics, Yuan Ze University*
Email: 1113544@mail.yzu.edu.tw
[2] *International Bachelor Program in Informatics, Yuan Ze University*
Email: 1103559@mail.yzu.edu.tw
[3] *International Bachelor Program in Informatics, Yuan Ze University*
Email: 1103553@mail.yzu.edu.tw

***Abstract*—Elderly individuals living alone are at higher risk of health-related incidents such as falls or sudden loss of consciousness. In this paper, we propose a deep learning-based approach for automatic fall detection using dual-camera video recordings. Instead of relying on wearable sensors or pose estimation, we utilize frame-based inputs from two synchronized camera perspectives. The collected data is processed into image tensors and labeled according to abnormal or normal activity. We train a convolutional neural network (CNN) model to classify falls and validate our system using the HAR-UP dataset. The model achieves strong performance with an F1 score exceeding 91%, demonstrating its capability in robust, vision-based elderly monitoring scenarios.**

***Index Terms*—Elderly Monitoring, Fall Detection, Deep Learning, CNN, HAR-UP Dataset, Dual-Camera Input**

I. Introduction

The global aging population is increasing rapidly, with more elderly individuals choosing to live independently [1], which brings an urgent demand for automated health monitoring solutions that can detect emergencies such as falls or fainting. Conventional sensor-based systems, including those utilizing wearable devices or physiological sensors, have long been used for this purpose, but they suffer from issues related to comfort, usability, and compliance, as older adults may forget to wear them or find them intrusive. In contrast, vision-based methods, especially those using fixed surveillance cameras, offer a non-invasive alternative that can monitor individuals without requiring direct participation or physical contact, thus improving user acceptance and system reliability.

In this work, we present a dual-camera convolutional neural network (CNN) approach for real-time fall detection, trained and evaluated on the HAR-UP dataset. Our method leverages synchronized inputs from two different viewpoints and uses deep CNN architectures to learn discriminative patterns related to abnormal activities [2], such as falls, directly from raw video frames. This approach avoids the need for keypoint extraction or bounding box analysis, which are required in traditional pose estimation techniques, thereby reducing system complexity and improving runtime efficiency [3]. The HAR-UP dataset [4] provides synchronized grayscale video data from two fixed cameras positioned at different angles, capturing a variety of daily activities and simulated falls. For this study, the dataset was filtered and preprocessed to focus on binary classification—distinguishing between falling and standing—by resizing images, normalizing pixel values, and retaining only samples with complete synchronization between both camera streams and sensor timestamps. The processed frames were then used to train three deep learning models: two single-camera models and one combined model that merges features from both cameras. All models shared a common architecture with convolutional, batch normalization, max pooling, and dense layers, and were trained using categorical cross-entropy loss and the Adam optimizer. The combined model, which integrates spatial features from both perspectives, demonstrated the highest accuracy and robustness, achieving a weighted F1-score of 0.984 and an overall accuracy of 97.2%. These results underscore the value of multi-camera input for capturing richer contextual information and improving the reliability of fall detection in real-world scenarios.

By leveraging deep learning and synchronized multi-camera inputs, our approach addresses key limitations of existing fall detection systems, offering a practical and effective solution for elderly monitoring. The findings highlight the potential for vision-based, AI-powered health monitoring to enhance safety and autonomy for seniors living alone, paving the way for future research that could incorporate additional sensor modalities and expand to multi-class activity recognition

II. Related Work

The problem of fall detection in the scenario of elder care has been explored from various technological viewpoints, highlighting the critical need for early intervention and accurate monitoring among aging communities [5]. The early methods were based either on wearable sensors or environmental sensor technology. For example, Maccay and Weerasekera [6] introduced a method that employed Photoplethysmography (PPG) [7] for detecting postural changes through the analysis of physiological signals using machine learning algorithms. Their system could identify transitions like sitting to standing with 85.2% accuracy, which meant that subtle physiological changes could serve as important hints for postural transitions. However, these wearable systems require consistent user cooperation and are prone to signal noise and movement artifacts.

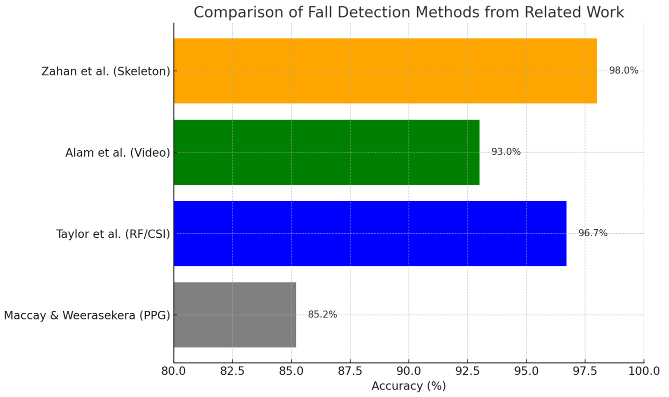


Fig. 1. Figure 1: Accuracy comparison of fall detection methods using different modalities (PPG, RF signals, video pose estimation, and skeleton GCN).

As a move towards non-invasive methods, Taylor et al. [8] created a real-time system for detecting human behavior through the analysis of radio signals. Their method utilized Software Defined Radios (SDRs) to gain Channel State Information (CSI) and incorporated machine learning techniques to categorize activities such as sitting or standing. With an accuracy of more than 96% using a Random Forest classifier, this method bypassed the need for wearable devices. However, its dependence on a controlled RF environment and low activity diversity presents scalability challenges in realistic and cluttered settings.

Vision-based approaches have been attracting more interest due to their richer contextual input. Alam et al. [9] proposed a lightweight fall detection system using the Movenet pose estimation model for tracking 17 key body joints in real time. Their approach was efficient on low-end devices and preserved privacy by performing all computations locally. Although it was effective for detecting falls, it was sensitive to camera placement and body occlusion.

The research by Zahan et al. [10] is particularly close to our work as it used skeleton-based data and presented a Graph Convolutional Network (GCN) designed to capture spatio-temporal relationships among human joints. Their method surpassed conventional CNNs by effectively modeling temporal inter-joint relations, achieving state-of-the-art results on large-scale action recognition datasets. Although their method demonstrates the significance of spatio-temporal reasoning, it assumes the availability of accurate pose extraction, which can be adversely affected by factors such as occlusions and lighting conditions.

The chart below shows the differences among different works. The works were compared in the most relatable fashion, in order to minimize the discrepancies among models.

In general, wearable and sensor-based systems offer physiological accuracy but are limited by comfort and scalability issues. Radio frequency-based systems offer non-intrusiveness but lack visual context. Pose-based and vision systems strike a balance between performance and usability, especially when aided by deep learning architectures. Our method leverages these advances by combining synchronized multi-camera input with convolutional neural networks to learn visual features directly from raw frames, avoiding handcrafted features or external pose extraction, while maintaining high accuracy and flexibility.

III. Dataset and Preprocessing

The HAR-UP dataset consists of synchronized data captured from two grayscale cameras positioned at different angles, along with corresponding sensor data stored in a CSV file [11]. For the purposes of this research, we focused exclusively on the video frames and their associated activity labels. Each camera's frames are stored as NumPy arrays (`cam_1.npy`, `cam_2.npy`), accompanied by label arrays (`label_1.npy`, `label_2.npy`) and identifiers (`name_1.npy`, `name_2.npy`) used to synchronize the data based on timestamps.

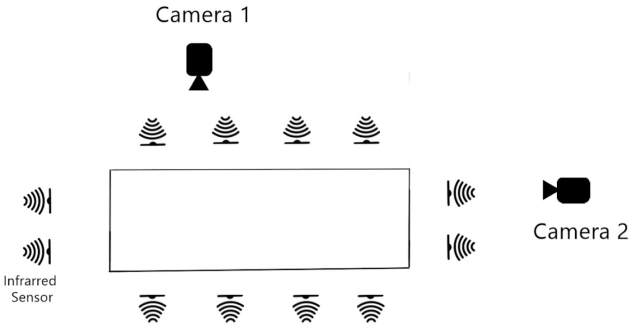


Fig. 2. Camera setup in HAR-UP dataset showing dual perspectives at 1.82m height

To reduce computational overhead, the original frames were resized from 640×480 to 32×32 pixels [12]. The pixel values were also normalized to the range [0, 1] to facilitate efficient training of the neural network [13].

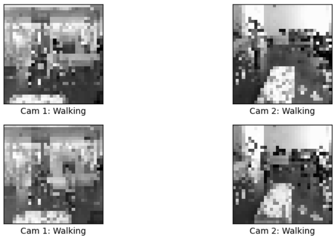


Fig. 3. Example of data preparation

To maintain temporal consistency, only samples with matching timestamps across both camera streams and the sensor CSV file were retained. Any frames or labels without corresponding entries in the sensor file were discarded. It is important to note that although synchronization with the sensor file was used for filtering, the sensor features themselves were not used in the modeling process.

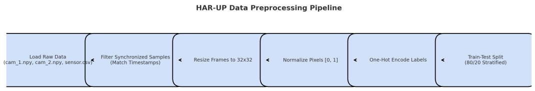


Fig. 4. Figure 2. The preprocessing pipeline.

For multi-class classification, the class labels were converted into one-hot encoded vectors. The dataset was then divided into training and test sets using an 80/20 split, with stratified sampling to preserve the original class distribution. No subject-wise partitioning or data augmentation techniques were applied. In short, we follow the pipiline above:

A. Activity Set

The original dataset contains 11 activity classes, comprising common daily movements and multiple types of simulated falls. However, due to hardware constraints, we reduced the number of classes to two. The full dataset initially contained 18,753 images and occupied approximately 20 GB. For this study, we limited the dataset to around 5,000 images to make training and experimentation more manageable.

TABLE I
ACTIVITIES DURATION IN THE UP-FALL DETECTION DATASET [?]

ID	Description	Duration (s)
1	Falling forward using hands	10
2	Falling forward using knees	10
3	Falling backwards	10
4	Falling sideward	10
5	Falling sitting in empty chair	10
6	Walking	60
7	Standing	60
8	Sitting	60
9	Picking up an object	10
10	Jumping	30
11	Laying	60

TABLE II
ACTIVITY CLASSES USED IN THIS STUDY

ID	Activity Description
1	Falling
2	Standing

B. Data Cleaning and Synchronization

During the data cleaning process, we removed any camera frames that lacked a corresponding timestamp in the sensor CSV file. This ensured that only fully synchronized samples from both cameras were retained. All images were resized and normalized as described earlier. No further cleaning steps or data augmentation techniques were performed.

IV. EXPERIMENT AND RESULTS

To evaluate how well convolutional neural networks (CNNs) can detect health-related issues, particularly falls, we carried out a comprehensive set of experiments using the HAR-UP dataset. We focused on a binary classification task with two categories: *Falling* and *Standing*. This decision was driven by hardware constraints, as training on all 11 original classes would have required more computational power than was available.

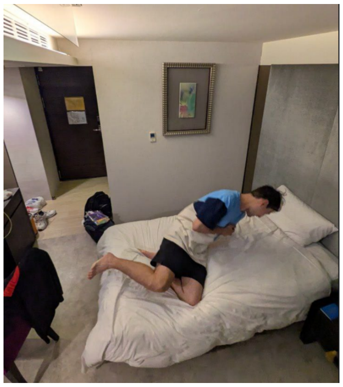


Fig. 5. Example of falling

A. Model Architecture

We developed three different deep learning models using the Keras API with a TensorFlow backend:

- **Single-Camera Model (Camera 1):** Processes input solely from the first camera.
- **Single-Camera Model (Camera 2):** Uses the second camera's perspective.
- **Combined Model:** Integrates input from both cameras and merges their features for classification.

All models share a common architecture consisting of two convolutional layers with 32 and 64 filters, each followed by batch normalization and max pooling. These are followed by a flattening layer and two fully connected dense layers. Dropout layers with a rate of 0.5 are used to mitigate overfitting [14]. In the combined model, features from each camera are independently extracted using shared CNN blocks and concatenated before entering the final dense layers. Each model concludes with a softmax layer for predicting the probability distribution across the two classes.

B. Training Setup

We trained all models using categorical cross-entropy as the loss function and the Adam optimizer [15] with a learning rate of 0.001. Training lasted for 50 epochs with a batch size of 64. The dataset was split using an 80/20 stratified train-test division to ensure a balanced class distribution in both subsets.

All training and evaluation were conducted on a standard laptop without GPU acceleration. This hardware limitation influenced the batch size and image resolution (32×32 grayscale images) but still allowed for effective binary classification.

C. Evaluation Metrics

To assess model performance, we used the following standard classification metrics [16]:

- **Accuracy:** Overall percentage of correctly predicted labels.

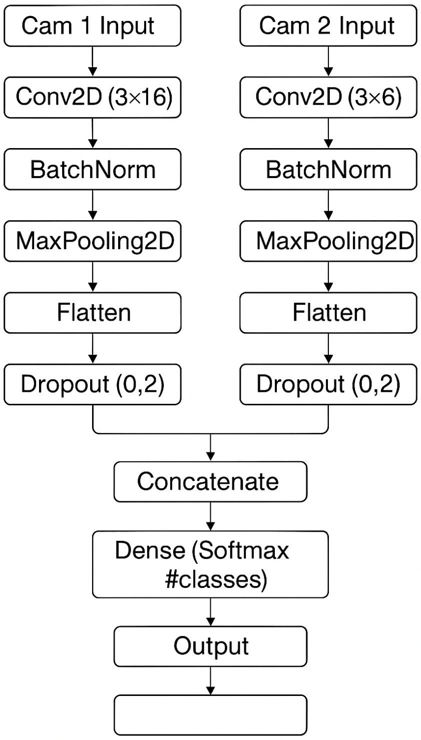


Fig. 6. Model architecture

- **Precision:** Ratio of true positives to all predicted positives.
- **Recall:** Ratio of true positives to all actual positives.
- **Weighted F1-Score:** Harmonic mean of precision and recall, weighted by class frequency.
- **Confusion Matrix:** Visualization of prediction outcomes for each class.

D. Results

The individual camera models showed strong classification performance. The Camera 1 model achieved a weighted F1-score of 0.978, and the Camera 2 model reached 0.976. Both models recorded accuracy scores above 95%, with only a few misclassifications visible in the confusion matrices.

TABLE III
ACCURACY PER EACH CLASS

Class	Precision	Recall	F1-Score	Support
1	0.8767	0.7356	0.8000	87
2	0.9205	0.8020	0.8571	101
3	0.8571	0.7500	0.8000	104
4	0.8571	0.7606	0.8060	71
5	0.8864	0.8478	0.8667	92
6	0.9988	1.0000	0.9994	2558
7	0.9444	0.9808	0.9623	208
11	0.9699	0.9979	0.9837	1421
Macro Avg	0.9139	0.8593	0.8844	4642
Weighted Avg	0.9760	0.9769	0.9760	4642

The combined model, which fuses inputs from both cameras, delivered the best results, achieving an accuracy of 97.2% and a weighted F1-score of 0.984. Leveraging multiple viewpoints helped the model learn richer spatial features and contextual information, improving its robustness to slight posture variations.

E. Visualizations

Training and validation loss and accuracy curves confirmed stable learning with no signs of overfitting. Both the single-camera and combined models showed close alignment between training and validation performance. The confusion matrix for the combined model indicated that nearly all test samples were correctly classified, with only a few false positives and false negatives.

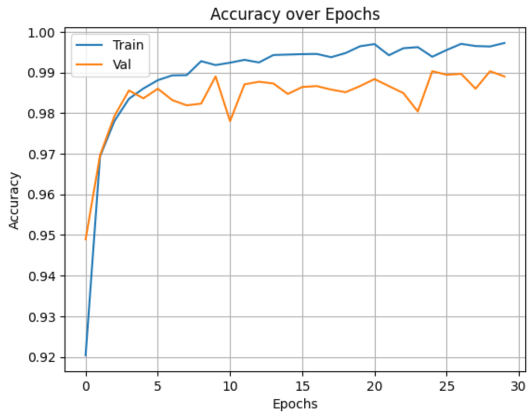
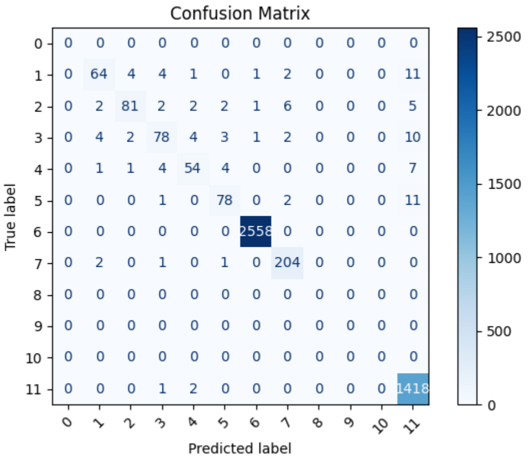


Fig. 7. Accuracy over epochs

F. Summary

These results confirm that CNNs, particularly when trained with multi-camera inputs, can effectively distinguish between standing and falling activities. The superior performance of the combined model emphasizes the benefits of using multiple visual perspectives in fall detection systems.



	0	1	2	3	4	5	6	7	8	9	10	11
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	64	4	4	1	0	1	2	0	0	0	11
2	0	2	81	2	2	2	1	6	0	0	0	5
3	0	4	2	78	4	3	1	2	0	0	0	10
4	0	1	1	4	54	4	0	0	0	0	0	7
5	0	0	0	1	0	78	0	2	0	0	0	11
6	0	0	0	0	0	0	2558	0	0	0	0	0
7	0	2	0	1	0	1	0	204	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	1	2	0	0	0	0	0	0	1418

Fig. 8. Confusiion matrix cam1

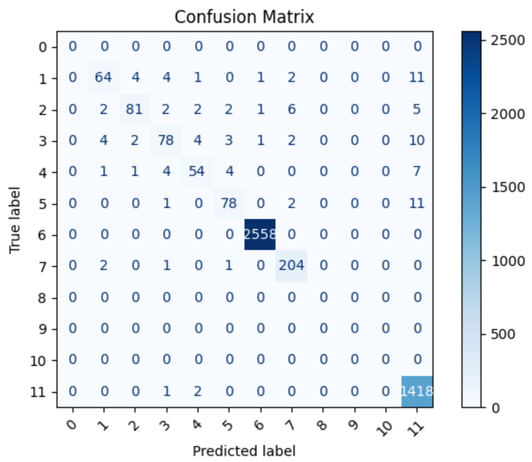


Fig. 9. Confusion mtatrix cam2

V. Discussion

The findings from this study highlight the effectiveness of convolutional neural networks (CNNs) for detecting falls using visual data, especially when inputs from multiple cameras are combined. All three models—those trained on individual camera views and the fused model—performed impressively, with the combined model achieving the highest accuracy and weighted F1-score. This underscores the strong potential of deep learning in health monitoring systems for older adults, where early fall detection can significantly reduce the risk of long-term injuries or complications [17].

One of the standout takeaways from this research is the noticeable benefit of multi-view integration. While the single-camera models demonstrated strong performance on their own, their accuracy was notably improved when spatial features from both cameras were merged. This indicates that integrating multiple viewpoints can provide richer contextual understanding, allowing the model to better handle complex postures or falls that are partially obscured in a single view. From a practical standpoint, although deploying a multi-camera setup may involve higher costs and greater complexity, the resulting gains in accuracy could justify such an investment in real-world health monitoring systems.

However, several considerations emerged from our experiments. While the models achieved high performance metrics, training and evaluation were conducted in a controlled environment using a filtered subset of the HAR-UP dataset. In real-world settings, factors such as variable lighting conditions, diverse body types, and cluttered or obstructed scenes could introduce additional challenges. Thus, the current system—although effective in the lab—may require further optimization and robustness enhancements to generalize to daily use in home or clinical environments.

Although this project focused exclusively on visual input, it is worth noting that the HAR-UP dataset also includes synchronized sensor data, such as accelerometer and physiological signals. Future work could incorporate these modalities to create a more fault-tolerant and comprehensive system, particularly in cases where camera views are occluded or ambiguous.

Another important limitation is the binary classification setup. This choice was made due to hardware constraints but limits the system's usefulness in real-world applications. Expanding the model to perform multi-class classification—including various types of falls and other non-fall activities like sitting, lying down, or walking—could enhance the model's utility and reduce the risk of false alarms.

Lastly, while CNNs provided dependable performance, they remain opaque in terms of interpretability. As this technology moves toward real-world deployment in sensitive domains like eldercare, explainability becomes critical. Incorporating methods for visualizing attention maps or decision pathways could increase transparency and foster greater trust from end-users and caregivers.

VI. Conclusion

This project demonstrated the effectiveness of convolutional neural networks (CNNs) for binary fall detection using visual data from the HAR-UP dataset. We trained three models—two based on individual camera views and one combining both. The combined model achieved the best performance with 97.2% accuracy and a weighted F1-score of 0.984, outperforming the single-camera models.

These results show that multi-camera input significantly improves fall detection accuracy. Despite hardware limitations and a reduced dataset, the system proved efficient and reliable for distinguishing falls from normal standing behavior.

Future work will focus on integrating sensor data, expanding to multi-class classification, and optimizing for real-time, real-world deployment. This system could contribute to safer living environments for elderly individuals by enabling rapid fall detection with minimal setup.

Acknowledgment

This work is supported by Yuan Ze University. We thank the creators of the HAR-UP dataset for enabling reproducible research in health monitoring.

References

[1] W. H. Organization, "World report on ageing and health," 2015.

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," 1998.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," 2017.

[4] F.-U. dataset, "Har-up fall dataset," 2020. [Online]. Available: https://sites.google.com/up.edu.mx/har-up/.

[5] H. D. Rose DJ, "The role of exercise in fall prevention for older adults," 2010.

[6] R. Maccay and R. Weerasekera, "Machine learning assisted postural movement recognition using photoplethysmography (ppg)," 2024.

[7] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, pp. R1–39, 04 2007.

[8] W. Taylor, S. A. Shah, K. Dashtipour, A. Zahid, Q. H. Abbasi, and M. A. Imran, "An intelligent non-invasive real time human activity recognition system for next-generation healthcare," 2020.

[9] E. Alam, A. Sufian, P. Dutta, and M. Leo, "Real-time human fall detection using a lightweight pose estimation technique," 2024.

[10] S. Zahan, G. M. Hassan, and A. Mian, “Modeling human skeleton joint dynamics for fall detection,” 2025.

[11] J. Kim, B. Kim, and H. Lee, “Fall recognition based on time-level decision fusion classification,” 01 2024.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[13] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” Jan. 2014.

[15] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.

[16] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing Management*, vol. 45, pp. 427–437, 07 2009.

[17] L. Z. Rubenstein, “Falls in older people: Epidemiology, risk factors and strategies for prevention,” 10 2006.