

# Deep Multimodal Learning for Human Fall Detection Using RGB Cameras and Wearable Inertial Sensors

Pulkit Gandhi<sup>1</sup>, Ritik Gandhi<sup>2</sup>

<sup>1</sup>Jayshree Periwal International School, India, <sup>2</sup>University of Alberta, Canada.

\*\*\*

**ABSTRACT**-Human fall detection has become an important research area due to the increasing elderly population and the severe health risks associated with undetected falls. Conventional **vision-based** systems are often affected by occlusions, illumination variations, and viewpoint changes, whereas **wearable-sensor-based** approaches are susceptible to false alarms caused by noisy measurements and activities of daily living that resemble falls. To address these limitations, this paper presents a **deep multimodal fusion framework** that integrates visual information from RGB cameras with inertial data acquired from wearable sensors for robust human fall detection. The proposed architecture employs convolutional neural networks to independently learn spatial features from camera streams and temporal motion features from inertial sensor signals. The extracted representations are subsequently combined through **feature-level fusion** and classified using fully connected layers. The proposed framework was evaluated on the publicly available **UP-FALL** dataset and achieved an accuracy of **99.04%** and an **F1-score** of **99.03%**. Experimental results indicate that multimodal feature fusion can improve the reliability of fall detection by effectively leveraging complementary visual and inertial information. These findings highlight the potential of deep multimodal learning for intelligent healthcare monitoring and assisted living systems.

**Key words:** Deep learning, Fall detection, Human activity recognition, Inertial measurement unit (IMU), Multimodal fusion

## 1. INTRODUCTION

### 1.1 Motivation

Falls are a major healthcare concern, particularly among elderly individuals, as they can result in severe injuries, reduced independence, and increased healthcare costs. The risk associated with falls increases significantly when an individual remains unattended after an incident. Therefore, automated fall detection systems capable of providing timely alerts have become an important area of research in healthcare monitoring, assisted living, and smart home environments [1].

Existing fall detection approaches can be broadly classified into vision-based and wearable-sensor-based methods. Vision-based approaches analyze human posture

and motion using camera data. Traditional techniques such as background subtraction, optical flow, skeleton tracking, and depth-based analysis have been widely investigated for fall detection [2], [3]. These methods provide rich spatial information and allow continuous monitoring without requiring users to wear additional devices. However, their performance can be affected by illumination changes, occlusions, camera viewpoints, and background complexity. In addition, continuous video monitoring raises privacy concerns that may limit deployment in healthcare settings.

Wearable sensor systems use inertial measurement units (IMUs), including accelerometers, gyroscopes, and magnetometers, to capture body motion through acceleration, angular velocity, and orientation measurements. Compared to camera-based methods, they are less sensitive to environmental conditions and can operate reliably in different settings. However, they may generate false positives because several activities of daily living (ADLs) exhibit motion patterns similar to falls. Furthermore, users are required to wear sensing devices continuously, which may reduce long-term usability and user compliance [4].

### 1.2 Problem Statement

Although both vision and wearable sensor approaches have shown promising results, each modality has inherent limitations. Vision-based methods provide valuable information about body posture and surrounding context but may fail under challenging visual conditions. In contrast, wearable sensors accurately capture body motion but lack contextual information about the environment.

To overcome these limitations, recent research has focused on combining multiple sensing modalities for fall detection. The **UP-FALL** dataset introduced by Martínez-Villaseñor *et al.* provides synchronized RGB camera and wearable sensor recordings, enabling the development of multimodal approaches for human activity recognition and fall detection [5].

Despite recent progress, effectively combining heterogeneous data remains a challenging task because visual and inertial data differ in both representation and temporal characteristics. Many existing approaches either rely on handcrafted feature engineering or use complex network architectures that increase computational cost.

There is therefore a need for deep learning frameworks that can automatically learn meaningful representations from both modalities while effectively combining their complementary information.

In this work, we propose a deep vision-and-sensor fusion framework for human activity recognition and fall detection. The proposed architecture processes the two camera streams and the wearable sensor data using three independent neural-network branches. Two two-dimensional convolutional neural networks extract spatial features from the camera images, while a one-dimensional convolutional neural network learns temporal patterns from the inertial sensor signals. The extracted representations are subsequently combined through feature-level fusion and classified into the activity categories available in the selected dataset.

### 1.3 Contributions

The main contributions of this work are summarized as follows:

- A deep multimodal framework is presented that combines two RGB camera streams with wearable inertial sensor measurements for multiclass human activity recognition and the identification of fall-related activities.
- A three-branch neural-network architecture is developed in which two two-dimensional CNN branches learn visual representations from separate camera viewpoints, while a one-dimensional CNN branch learns temporal representations from wearable sensor measurements.
- A feature-level fusion strategy is employed to concatenate the learned camera and sensor representations before classification, enabling complementary spatial and motion information to be processed within a unified model.
- The framework is evaluated using the recordings of one participant from the publicly available UP-FALL dataset and obtains a testing accuracy of 99.04% and a testing F1-score of 99.03%.
- The limitations of the subject specific evaluation are explicitly discussed, providing a basis for future subject independent evaluation using the complete dataset.

## 2. RELATED WORK

### 2.1 Vision-Based Fall Detection

Vision-based approaches have been widely investigated for automated fall detection because they enable continuous monitoring without requiring individuals to

wear sensing devices. Early methods relied on traditional computer vision techniques such as background subtraction, optical flow, silhouette extraction, and shape analysis to identify changes in human posture associated with falls. Although these approaches produced encouraging results in controlled environments, their performance was often affected by illumination changes, occlusions, complex backgrounds, and variations in camera viewpoint [2], [3].

The rapid development of deep learning has significantly improved vision-based fall detection. Convolutional neural networks automatically learn discriminative spatial features from image data, reducing the dependence on handcrafted feature extraction [2], [3]. Temporal information can be incorporated by analysing consecutive frames or by combining convolutional feature extractors with sequence-modelling architectures. More recent work has also investigated transformer-based models for learning long-range spatial and temporal relationships in video sequences [8]. Despite these advances, the performance of vision-based systems remains dependent on image quality, camera placement and environmental conditions. Privacy concerns associated with continuous video surveillance may also limit their adoption in practical healthcare applications [2], [3].

### 2.2 Wearable-Sensor-Based Fall Detection

Wearable sensor systems detect falls by monitoring body motion using inertial measurement units (IMUs) that typically consist of accelerometers, gyroscopes, and magnetometers. These sensors measure changes in acceleration, angular velocity, and body orientation, making them suitable for identifying sudden movements associated with fall events. Initial research primarily focused on threshold-based algorithms that classified falls by comparing sensor measurements against predefined acceleration limits. While computationally simple, these methods frequently produced false alarms because several activities of daily living exhibit motion patterns similar to falls [4].

To improve reliability, researchers have increasingly applied machine learning techniques such as Support Vector Machines, Decision Trees, Random Forests, and Artificial Neural Networks to classify inertial sensor data. More recently, deep learning models, including one-dimensional CNNs and recurrent neural networks, have demonstrated the ability to learn temporal patterns from sequential sensor measurements with reduced dependence on manually designed features [4], [7]. Recurrent architectures are particularly useful for modelling relationships across successive observations in continuous activity streams [7]. Although wearable sensor systems are less affected by environmental conditions and generally provide greater privacy than camera-based systems, their

effectiveness depends on continuous user compliance and appropriate sensor placement [1], [4].

## 2.3 Multimodal Fall Detection

The complementary characteristics of vision-based and wearable-sensor-based approaches have motivated the development of multimodal fall-detection systems. Camera data provide information about body posture and the surrounding environment, whereas inertial sensors capture motion characteristics that are less dependent on lighting conditions and camera viewpoint. Combining these modalities can provide a more comprehensive representation of human activities than using either type of information alone.

The UP-FALL dataset introduced by Martínez-Villaseñor et al. provides synchronized visual and wearable sensor recordings for investigating multimodal human activity recognition and fall detection [5]. Subsequent research using this dataset has examined methods that combine information from cameras and heterogeneous sensors to improve activity classification [6]. Such approaches seek to exploit the complementary characteristics of visual posture information and inertial motion measurements.

Although multimodal methods have achieved promising results, several challenges remain. Visual and inertial inputs differ in dimensionality, sampling structure and representation, making their alignment and fusion difficult. Some multimodal architectures also introduce substantial computational complexity or depend on extensive preprocessing and handcrafted feature construction. Motivated by these challenges, the proposed framework uses lightweight modality-specific convolutional branches followed by feature-level concatenation to construct a unified representation for multiclass activity recognition and the identification of fall-related activities.

## 3. PROPOSED METHODOLOGY

### 3.1 Framework Overview

The proposed framework employs a multimodal deep learning architecture for human activity recognition and fall detection by jointly processing visual and wearable sensor data. The system integrates image sequences captured from two synchronized camera viewpoints with corresponding wearable inertial sensor measurements, allowing complementary spatial and motion information to be learned simultaneously.

The framework consists of four main stages: data preparation, modality-specific preprocessing, feature extraction, and multimodal classification. During preprocessing, the visual and sensor data are processed independently according to their respective characteristics.

Feature extraction is then performed using two identical two-dimensional convolutional neural network branches for the camera inputs and a one-dimensional convolutional neural network branch for the wearable sensor data. The extracted feature representations are subsequently combined through feature-level fusion and passed to fully connected layers for activity classification.

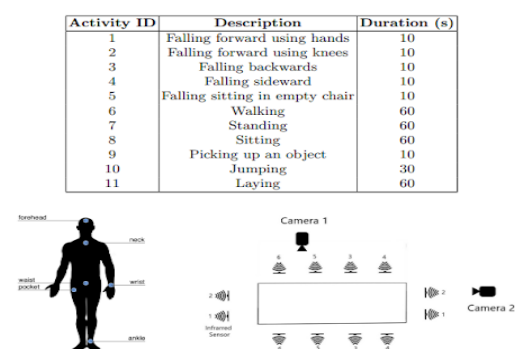
Unlike decision-level fusion, where each modality is classified independently, the proposed approach performs feature-level fusion before classification. This enables the network to learn the relationships between visual posture and body movement simultaneously, resulting in a unified representation for recognizing both activities of daily living and fall events.

The proposed architecture is formulated as a multi-class activity recognition problem, where fall events and activities of daily living are treated as separate activity classes. Consequently, fall detection is achieved by identifying the predicted fall class during the final classification stage.

### 3.2 UP-FALL Dataset

The proposed framework was evaluated using the **UP-FALL Detection Dataset**, a publicly available multimodal dataset developed for human activity recognition and fall detection [5]. The dataset contains recordings from **17 participants** performing **11 different activities**, with each activity repeated across three trials. The activities comprise activities of daily living and simulated fall scenarios, providing a diverse collection of movements for evaluating activity recognition and fall detection methods.

The UP-FALL dataset incorporates multiple sensing modalities, including two RGB cameras, wearable inertial sensors, infrared sensors, and physiological sensing devices. In this work, only the visual data captured by the two cameras and the wearable sensor measurements were utilized for model development.



**Figure 1.** Experimental setup of the UP-FALL dataset showing the placement of RGB cameras and wearable sensing devices [5].

The visual recordings provide two complementary viewpoints of each activity, allowing the model to observe body posture from different perspectives. The wearable sensor data captures changes in body movement that may not be visually distinguishable, particularly during sudden transitions associated with fall events. The combination of these two modalities provides a richer representation of human activities than either modality alone.

Although the complete dataset contains recordings from 17 participants, the experiments reported in this paper were conducted using recordings from a single participant. Consequently, the reported results represent a subject-specific evaluation rather than a subject-independent assessment. All available activities and trials corresponding to the selected participant were used during preprocessing and model development. This subject-specific evaluation allows the proposed multimodal architecture to be investigated under controlled experimental conditions while maintaining consistency with the original implementation.

### 3.3 Data Preparation and Synchronization

The original UP-FALL dataset is organized according to participant, activity, trial, and sensing modality. Before being supplied to the proposed network, the visual and sensor data were reorganized into a format suitable for multimodal learning.

For the visual modality, the recorded video sequences from both cameras were decomposed into individual image frames. Each extracted frame inherited the activity label associated with its corresponding recording, thereby generating labelled image samples for supervised learning.

For the wearable sensor modality, the sensor measurements corresponding to the selected participant were collected from the CSV files associated with each activity and trial. These records were consolidated into a single dataset while preserving their corresponding activity labels.

Initially, the implementation considered a binary classification problem consisting of fall and non-fall classes. During model development, this formulation was extended to classify all activity categories available in the selected dataset. Consequently, the proposed network learns to distinguish between individual human activities, while fall detection is achieved by identifying the activity classes corresponding to fall events.

Since image data and sensor measurements differ significantly in their structure, separate preprocessing procedures were applied before feature extraction. The processed outputs from both modalities were subsequently

provided to their respective network branches for multimodal learning.

### 3.4 Visual Data Preprocessing

The visual modality consists of image frames extracted from the video sequences captured by Camera 1 and Camera 2. The same preprocessing procedure was applied independently to both camera streams to ensure consistency during feature extraction.

Each RGB frame was converted to grayscale to reduce the input dimensionality while preserving the structural information required for activity recognition.



**Figure 2.** Representative preprocessed activity samples extracted from Camera 1 after grayscale conversion and activity labelling.



**Figure 3.** Additional examples of preprocessed activity samples used during model training.

The grayscale image was obtained using the following transformation:

$$I_g(x, y) = 0.299R(x, y) + 0.587G(x, y) + 0.114B(x, y)$$

where  $R(x, y)$ ,  $G(x, y)$ , and  $B(x, y)$  represent the red, green, and blue intensity values of the input image, respectively.

Following grayscale conversion, all images were resized to  $32 \times 32$  pixels to provide a uniform input dimension for the convolutional neural network. Each processed image was therefore represented as a  $32 \times 32 \times 1$  input tensor before being supplied to the corresponding visual branch. The activity labels associated with the original video sequences were preserved during frame extraction, ensuring that each processed frame retained its corresponding activity label.

### 3.5 Sensor Data Preprocessing

The wearable sensor data were stored as numerical records in CSV format, with each file corresponding to a particular activity and trial. The sensor files belonging to the selected participant were consolidated into a single dataset while preserving the associated activity labels.

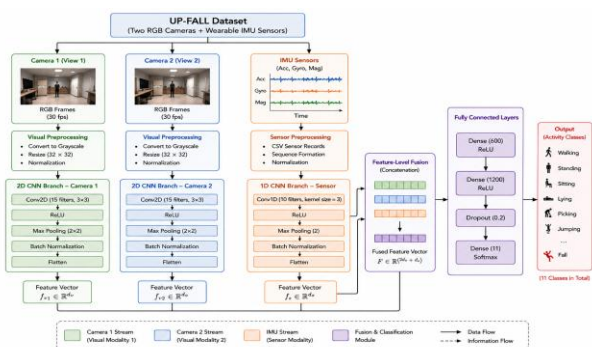
Unlike image data, the wearable sensor measurements represent sequential numerical observations. A sensor sequence is represented as

$$X_s = \{x_1, x_2, \dots, x_T\}$$

where  $T$  is the length of the sensor observation and each  $x_t$  contains the available sensor measurements at time  $t$ .

The processed sensor sequences were arranged in a format compatible with the one-dimensional convolutional neural network. No manual feature engineering or statistical feature extraction was performed. Instead, the raw sensor sequences were directly provided to the sensor branch, allowing the network to learn discriminative temporal features automatically during training.

### 3.6 Multimodal Network Architecture



**Figure 4.** Proposed three-branch multimodal deep learning architecture for human activity recognition and fall detection.

The proposed multimodal architecture consists of three parallel feature extraction branches: two visual branches corresponding to Camera 1 and Camera 2, and one sensor branch for the wearable sensor data. Each branch processes its corresponding input independently to learn modality-specific features before the extracted representations are combined through feature-level fusion.

#### 3.6.1 Visual Feature Extraction

Each camera stream is processed independently using an identical convolutional neural network. The visual branch consists of a two-dimensional convolutional layer with **15 filters** and a  $3 \times 3$  kernel, followed by a ReLU activation function,  $2 \times 2$  max pooling, batch normalization, and a flatten layer. This sequence enables the network to extract spatial features describing body posture and movement from each camera view.

For an input image  $X$ , the convolution operation is expressed as

$$H(i, j, k) = f(\sum_m \sum_n \sum_c W_{m,n,c,k} X_{i+m,j+n,c} + b_k)$$

where  $H(i, j, k)$  denotes the output feature map,  $W_{m,n,c,k}$  represents the convolution kernel weight associated with the  $k$ -th filter and  $b_k$  is the corresponding bias term, and  $f(\cdot)$  is the activation function.

The final feature maps generated by each visual branch are flattened into one-dimensional feature vectors before multimodal fusion.

#### 3.6.2 Sensor Feature Extraction

The wearable sensor data are processed using a one-dimensional convolutional neural network designed to learn temporal characteristics from sequential motion data. The sensor branch consists of a Conv1D layer with **10 filters** and a **kernel size of 3**, followed by ReLU activation, max pooling, Batch Normalization, and a Flatten layer.

The one-dimensional convolution operation is defined as

$$H(t, k) = f(\sum_{i=0}^{K-1} W(i, k) X_s(t+i) + b_k)$$

where  $X_s$  denotes the input sensor sequence,  $W(i, k)$  represents the convolution kernel,  $b_k$  is the bias corresponding to the  $k$ -th filter, and  $H(t, k)$  is the resulting feature-map.

The extracted sensor feature maps are subsequently flattened into a one-dimensional feature vector before being combined with the visual representations.

### 3.6.3 Feature Fusion

After feature extraction, the flattened feature vectors obtained from the two visual branches and the sensor branch are combined using feature-level fusion. Let  $F_{C1}$ ,  $F_{C2}$ , and  $F_s$  denote the feature vectors extracted from Camera 1, Camera 2, and the wearable sensor branch, respectively. The fused representation is obtained by concatenating these vectors as

$$F_{fusion} = [F_{C1}; F_{C2}; F_s]$$

where  $[\cdot]$  denotes the concatenation operation.

The fused feature vector combines complementary information from the three modalities into a single representation, enabling the classifier to jointly learn spatial characteristics from the camera views and temporal motion information from the wearable sensor data.

### 3.6.4 Classification Network

The fused feature vector is passed through two fully connected layers containing **600** and **1200** neurons, respectively. A dropout layer with a dropout rate of **0.2** is applied before the output layer to reduce overfitting during training.

The final classification is performed using a Softmax output layer, which computes the probability of each activity class as

$$P(y=i) = e^{z_i} / \sum_{j=1}^C e^{z_j}$$

where  $z_i$  represents the output score of the  $i^{th}$  class and  $C$  denotes the total number of activity classes.

The predicted activity corresponds to the class with the highest probability. Since fall events are included among the activity classes in the UP-FALL dataset, the proposed network performs fall detection by identifying the corresponding fall class during classification.

### 3.7 Data Partitioning and Model Training

After preprocessing, the aligned multimodal samples were divided into training, validation and testing subsets using a ratio of 70:15:15. The corresponding inputs from Camera 1, Camera 2 and the wearable sensor branch were partitioned together to preserve the relationship among the three modalities and their associated activity labels. A fixed random seed of 42 was used during data partitioning to improve the reproducibility of the experiments.

The proposed multimodal network was implemented using TensorFlow and Keras. The model was trained using the Adam optimizer with an initial learning rate of 0.001.

Categorical cross-entropy was employed as the loss function because the proposed framework performs multiclass activity classification. The network was trained for 30 epochs using a batch size of 32.

During each training epoch, the model parameters were updated using the training subset, while the validation subset was used to monitor the learning performance and identify potential overfitting. The testing subset was kept separate from the training process and was used only for the final evaluation of the trained model.

The performance of the proposed framework was evaluated using accuracy, precision, recall and F1-score. Accuracy represents the proportion of correctly classified samples, whereas precision measures the proportion of predicted samples belonging to a class that were correctly classified. Recall measures the proportion of actual samples belonging to a class that were correctly identified, while the F1-score represents the harmonic mean of precision and recall.

For the multiclass classification task, the reported precision, recall and F1-score values were calculated by aggregating the results across the activity classes. The same evaluation procedure was applied to both the validation and testing subsets to enable a consistent comparison of model performance.

## 4. RESULTS AND DISCUSSION

### 4.1 Performance of the Proposed Multimodal Framework

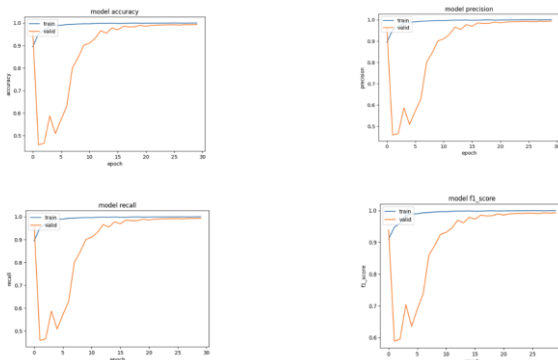
The proposed framework was evaluated as a multiclass human-activity-recognition model using visual information from Camera 1 and Camera 2 together with wearable sensor measurements. The three input streams were independently processed by their corresponding feature-extraction branches and subsequently combined through feature-level fusion. The fused representation was used to distinguish among the activity classes included in the selected UP-FALL recordings. Fall detection was performed by identifying predictions corresponding to the fall-related activity classes.

The proposed multimodal framework achieved an accuracy of 99.07% on the validation dataset and 99.04% on the testing dataset. The corresponding validation precision, recall, and F1-score were 99.09%, 99.05%, and 99.05%, respectively. On the testing dataset, the model obtained a precision of 99.05%, a recall of 99.03%, and an F1-score of 99.03%.

**Table 1.** Validation and testing performance of the proposed multimodal framework.

| Performance Metric | Validation (%) | Testing (%) |
|--------------------|----------------|-------------|
| Accuracy           | 99.07          | 99.04       |
| Precision          | 99.09          | 99.05       |
| Recall             | 99.05          | 99.03       |
| F1-score           | 99.05          | 99.03       |

The small difference between the validation and testing results indicates that the model maintained consistent performance on samples that were not used during training. The testing accuracy was only 0.03 percentage points lower than the validation accuracy. Similarly, the testing F1-score differed from the validation F1-score by only 0.02 percentage points. This consistency suggests that the learned multimodal representation was not limited to the validation data and remained effective during testing.



**Figure 5.** Training and validation performance of the proposed multimodal framework over 30 epochs in terms of (a) accuracy, (b) precision, (c) recall, and (d) F1-score.

The similarity among accuracy, precision, recall, and F1-score also indicates balanced classification performance. The high precision suggests that only a small proportion of the predicted activity labels were incorrect, while the high recall indicates that the model correctly identified most of the actual activity samples. The resulting F1-score confirms that the framework maintained a strong balance between precision and recall.

## 4.2 Comparison with Existing Methods

The performance of the proposed framework was compared with the methods reported in the original UP-FALL study. The earlier study evaluated conventional machine learning approaches, including Random Forest, Support Vector Machine, Multilayer Perceptron, and K-Nearest Neighbour, using multimodal information from the UP-FALL dataset.

The best-performing conventional machine learning method reported in the original UP-FALL study achieved an F1-score of 97.28%. In comparison, the proposed multimodal framework achieved a testing F1-score of 99.03%, representing an improvement of 1.75 percentage points. This improvement demonstrates the effectiveness of learning modality-specific feature representations from the two RGB camera views and wearable sensor data before feature-level fusion.

The performance improvement can be associated with the use of modality-specific convolutional branches. Instead of directly applying a conventional classifier to the combined input data, the proposed framework learns separate representations from the two camera views and the wearable sensor measurements. The visual branches extract spatial information related to body posture and appearance, while the sensor branch captures patterns present in the wearable measurements. These representations are combined before classification, allowing the network to use information from all three sources.

The proposed method also differs from approaches that depend on manually designed features. The convolutional layers learn relevant representations directly from the preprocessed visual and sensor inputs. This reduces the dependence on manually selected statistical or motion features and allows the complete network to learn the feature extraction and classification stages jointly.

## 4.3 Effect of Multimodal Feature Fusion

The obtained results demonstrate the potential benefit of combining multiple sources of information for multiclass activity recognition, including the recognition of fall-related activities. A single camera view may not always provide complete information about an activity because the subject's posture can appear differently depending on the viewing direction. In the proposed framework, Camera 1 and Camera 2 observe the activity from different positions. The two visual branches therefore provide complementary information about the subject's posture and movement.

The wearable sensor data further supports the visual information by representing changes in body motion. Sensor measurements can capture movement patterns that may not be clearly visible in an individual image frame. Conversely, the camera inputs provide spatial information that may not be available from the wearable measurements alone.

The feature-level fusion process combines these complementary representations into a single feature vector. As a result, the final classifier does not depend entirely on one camera or on the wearable sensor stream.

The high testing accuracy and F1-score indicate that the combined representation was effective for distinguishing between the activities considered in the experiment.

The results also show that the fusion process did not cause a noticeable reduction in generalization performance. The validation and testing values remained close across all four evaluation measures. This suggests that the classification network was able to use the fused information without becoming strongly dependent on the validation samples.

#### 4.4 Limitations

Although the proposed framework achieved high classification performance, the results should be interpreted with consideration of the experimental scope. The original UP-FALL dataset contains recordings from 17 subjects, whereas the present experiment used data corresponding to one subject. Therefore, the reported results demonstrate the model's performance on the selected participant's recordings and do not establish subject-independent generalization.

Human movement, posture, body structure, and fall patterns can vary considerably between individuals. A model evaluated using recordings from multiple participants may therefore produce different results. Future experiments should include additional subjects and use a subject-independent evaluation strategy in which the individuals included in the testing data are not included in the training data.

The visual inputs were also reduced to grayscale images with a spatial resolution of  $32 \times 32$  pixels. This preprocessing reduced the computational requirements of the visual branches, but it also removed colour information and fine spatial details. Higher-resolution images or temporal sequences of consecutive frames may provide additional movement information, although they would increase the computational requirements of the model.

Finally, the experiments were conducted using the recorded UP-FALL dataset. The reported results therefore represent offline classification performance under the conditions in which the dataset was collected. Further evaluation would be required before the framework could be considered suitable for real-time deployment in uncontrolled environments.

#### 5. CONCLUSION

This paper presented a multimodal deep learning framework for multiclass human activity recognition and the identification of fall-related activities using synchronized visual and wearable sensor data from the UP-FALL dataset. The proposed approach employs two

convolutional neural network branches to extract spatial features from Camera 1 and Camera 2, while a one-dimensional convolutional branch processes the wearable sensor measurements. The extracted features from all three modalities are combined through feature-level fusion and classified using fully connected layers to recognize human activities.

The experimental results demonstrate the effectiveness of the proposed multimodal framework. The model achieved a validation accuracy of 99.07% and a testing accuracy of 99.04%, with corresponding testing Precision, Recall, and F1-score of 99.05%, 99.03%, and 99.03%, respectively. These results indicate that combining complementary information from multiple camera views and wearable sensors enables accurate activity recognition while maintaining consistent performance on unseen data. Furthermore, the proposed framework outperformed the conventional machine learning approaches reported for the UP-FALL dataset, highlighting the advantages of convolutional feature extraction and multimodal feature fusion.

Although the proposed framework achieved promising results, the present study was conducted using data from a single participant in the UP-FALL dataset. Future work will focus on evaluating the framework using data from all participants to investigate its ability to generalize across different subjects. Additional research may also explore real-time deployment, the incorporation of temporal modelling techniques, and the evaluation of the proposed framework in more diverse indoor environments to further improve its practical applicability for intelligent healthcare and assisted living systems.

#### REFERENCES

- [1] Anuradha Singh, Saeed Ur Rehman, Sira Yongchareon and Peter Han Joo Chong, "Sensor Technologies for Fall Detection Systems: A Review", IEEE Sensors Journal, Vol. 20, No. 13, pp. 6889–6919, 2020.
- [2] Ekram Alam, Abu Sufian, Paramartha Dutta and Marco Leo, "Vision-Based Human Fall Detection Systems Using Deep Learning: A Review", Computers in Biology and Medicine, Vol. 146, Article 105626, 2022.
- [3] Ricardo Espinosa, Hiram Ponce, Sebastián Gutiérrez, Lourdes Martínez-Villaseñor, Jorge Brieva and Ernesto Moya-Albor, "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, Vol. 115, Article 103520, 2019.
- [4] Francisco Luna-Perejón, Manuel Jesús Domínguez-Morales and Antón Civit-Balcells, "Wearable Fall

Detector Using Recurrent Neural Networks", *Sensors*, Vol. 19, No. 22, Article 4885, 2019.

[5] Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez and Carlos Peñafort-Asturiano, "UP-Fall Detection Dataset: A Multimodal Approach", *Sensors*, Vol. 19, No. 9, Article 1988, 2019.

[6] Thao V. Ha, Hoang Nguyen, Son T. Huynh, Trung T. Nguyen and Binh T. Nguyen, "Fall Detection Using Multimodal Data", *Proceedings of the International Conference on Multimedia Modeling*, Springer, pp. 392–403, 2022.

[7] Haobo Li, Aman Shrestha, Hadi Heidari, Julien Le Kernec and Francesco Fioranelli, "Bi-LSTM Network for Multimodal Continuous Human Activity Recognition and Fall Detection", *IEEE Sensors Journal*, Vol. 20, No. 3, pp. 1191–1201, 2020.

[8] Adrián Núñez-Marcos and Ignacio Arganda-Carreras, "Transformer-Based Fall Detection in Videos", *Engineering Applications of Artificial Intelligence*, Vol. 132, Article 107937, 2024.

[9] Rafael Igual, Carlos Medrano and Inmaculada Plaza, "Challenges, Issues and Trends in Fall Detection Systems", *BioMedical Engineering Online*, Vol. 12, Article 66, 2013.