Self-supervised contrastive learning for fall detection using 3D vision-based body articulation

Guendoul O., Tabii Y., Oulad Haj Thami R.

ADMIR Lab, National School of Computer Science and Systems Analysis (ENSIAS), Mohammed V University, Rabat 10000, Morocco, oumaima_guendoul@um5.ac.ma

(Received 28 March 2025; Revised 27 September 2025; Accepted 1 October 2025)

This paper presents a mathematical modeling approach for fall detection using a 3D vision-based contrastive learning framework. Traditional models struggle with high false positives and poor generalization across environments. To address this, we propose a self-supervised contrastive learning model that maps 3D skeletal motion sequences into a low-dimensional embedding space, optimizing feature separation between falls and non-falls. Our method employs spatial-temporal modeling and a contrastive loss function based on cosine similarity to enhance discrimination. By leveraging graph-based feature representation, the model ensures robust performance even with missing or noisy data. Experimental results on benchmark fall detection datasets demonstrate a significant reduction in false positives while maintaining high accuracy, making the framework well-suited for real-world healthcare applications.

Keywords: *mathematical modeling; contrastive learning; 3D vision; skeletal motion analysis; healthcare; fall detection.*

2010 MSC: 68T05, 68U10, 68U20, 65D18 **DOI:** 10.23939/mmc2025.03.1012

1. Introduction

Falls represent a significant health risk for elderly individuals, often leading to severe injuries, long-term disabilities, and increased hospitalization rates. According to the World Health Organization (WHO), falls are the second leading cause of unintentional injury deaths worldwide, with individuals over the age of 65 at particular risk [1]. The consequences of falls can be devastating, leading to a decreased quality of life, higher healthcare costs, and a reduced ability to maintain independence. As a result, developing effective and reliable fall detection systems become a critical area of research in healthcare and assistive technologies. Existing solutions primarily fall into two categories: wearable sensor-based and vision-based approaches. Wearable sensor-based methods utilize accelerometers and gyroscopes to detect sudden changes in motion; however, they require individuals to wear devices continuously, which can be inconvenient and may lead to low adherence rates [2]. Conversely, vision-based systems offer a non-intrusive alternative by monitoring human motion through RGB cameras, depth sensors, or 3D skeletal tracking. Recent surveys, including those from our ongoing research [3], highlight the increasing adoption of vision-based methods due to their non-contact nature and potential for real-time fall detection [4].

Despite the advancements in vision-based fall detection systems, several challenges remain. One of the primary difficulties is distinguishing between falls and other common daily activities, such as sitting down abruptly, bending over, or picking up objects. These activities can resemble falls in terms of body movement, leading to false positives that compromise the reliability of the system. Furthermore, variations in lighting conditions, occlusions (e.g., furniture or other objects blocking the camera's view), and differing camera viewpoints further complicate the fall detection task [6]. In our previous work [5], we explored novel techniques aimed at improving both fall detection accuracy and predictive capabilities. These studies provide a comprehensive overview of state-of-the-art technologies while also addressing the persistent challenges associated with implementing vision-based fall detection systems in dynamic environments [7].

To overcome these limitations, one promising approach is contrastive learning, a self-supervised technique that enhances feature representation by leveraging similarities and differences between data samples [8]. By training models to contrast positive pairs (e.g., two frames depicting the same fall) against negative pairs (e.g., a fall versus a non-fall activity), contrastive learning enables the system to learn more discriminative and robust feature embeddings. This approach has been shown to significantly improve the accuracy and generalization capabilities of vision-based models, allowing them to distinguish falls from non-fall activities even in complex or ambiguous scenarios [9].

In addition to improving fall detection accuracy, precise head position estimation plays a crucial role in human motion analysis. Head movement patterns provide valuable cues for predicting falls, assessing postural stability, and enhancing pose estimation models [10]. In this work, we introduce a contrastive learning-based model that learns 3D embeddings of head position using XYZ coordinate variations. The proposed model applies contrastive loss functions to differentiate between similar and dissimilar head positions, thereby improving its ability to generalize across different individuals, environments, and camera angles. By incorporating contrastive learning for both motion classification and head position estimation, our approach aims to deliver a highly reliable, real-time fall detection system that minimizes false alarms and enhances deployment feasibility in real-world healthcare settings.

Self-supervised contrastive learning has emerged as a promising approach for fall detection, particularly when integrated with 3D vision-based body articulation techniques. This method leverages the ability to learn discriminative features from skeletal data, significantly enhancing the accuracy and efficiency of fall detection systems. Unlike traditional supervised learning, contrastive learning does not require labeled data and instead focuses on learning the underlying structure of the data through comparison. By training the model to distinguish between similar and dissimilar samples, it becomes better equipped to identify subtle yet critical differences between fall and non-fall events. This technique has shown considerable promise in recognizing falls from normal activities, where traditional methods often struggle [12]. The ability to learn such representations without relying on explicit labels is especially valuable in real-world applications where labeled data is scarce or expensive to acquire.

Furthermore, utilizing skeleton-based action recognition specifically with 3D skeleton data enhances this model can detect even the smallest movements associated with falls. Human movements like bending down, sitting abruptly, or reaching for an object can easily be mistaken for a fall if the system does not adequately capture the subtleties of body posture. By incorporating 3D skeleton data, which provides a detailed understanding of the movements of human body in three-dimensional space, fall detection systems can discern these subtle shifts with greater accuracy, thus improving detection rates even in complex and dynamic environments [13]. This is particularly important in uncontrolled real-world environments where background noise, lighting conditions, or occlusions can make it challenging to detect falls reliably.

In terms of integrating 3D vision, human pose estimation (HPE) techniques like MoveNet play a pivotal role. MoveNet, an advanced deep learning model for human pose estimation, is capable of extracting high-fidelity features from human gestures. These features are crucial for identifying fall events, as the system can track the exact position and movement of key body joints in real-time. The enhanced resolution and accuracy of pose estimation ensure that even quick or subtle falls are detected before they result in harm [12]. For instance, a fall event may involve rapid and often irregular movements, making it imperative to detect the onset of a fall as early as possible. MoveNet's capabilities allow the system to quickly adapt to different body shapes, motion speeds, and environmental conditions, making it a robust tool for fall detection.

Additionally, combining multiple image streams and applying segmentation techniques improves the model's ability to classify actions accurately and reduce false positives. By fusing images from different viewpoints or sensors, the system gains a more holistic view of the scene, mitigating issues like occlusions or partial visibility. Image segmentation helps isolate the person from the background, enhancing the model's focus on human movements rather than irrelevant environmental factors. This multi-modal approach not only increases accuracy but also helps the system make more timely and reliable fall detection decisions [14]. The ability to segment and merge multiple data sources ensures that the system is not only detecting falls but also distinguishing between different types of motions that may resemble a fall, such as sitting down or bending over [15].

Together, these approaches provide a more sophisticated and reliable framework for real-time fall detection systems. By integrating self-supervised contrastive learning, skeleton-based recognition, human pose estimation, and image fusion, the system becomes highly adaptable to diverse environments and capable of operating effectively in real-world healthcare settings. This combination of techniques ensures that the fall detection system is not only accurate but also resilient to the many challenges posed by dynamic and unpredictable conditions. As a result, the model is better equipped to minimize false alarms and provide timely alerts, improving both safety and quality of life for elderly individuals and others at risk of falls.

2. Related work

Previous research in fall detection can be categorized into two primary approaches: sensor-based and vision-based methods. Sensor-based techniques utilize accelerometers and gyroscopes to detect abrupt motion changes, offering high sensitivity but often posing practical challenges for elderly individuals who may find continuous device usage inconvenient. Additionally, these methods can generate false positives, as sudden but non-harmful movements may trigger alarms [11].

Vision-based fall detection, in contrast, relies on visual data captured by RGB cameras, depth sensors, or 3D skeleton tracking. These approaches eliminate the need for wearable devices but introduce new challenges such as variations in lighting conditions, occlusions, and similarities between fall and non-fall activities [16]. Traditional computer vision techniques relied on handcrafted feature extraction methods, which struggled to generalize across diverse real-world scenarios. The introduction of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has significantly improved fall detection accuracy by automatically extracting spatial and temporal features [17]. However, deep learning models often require large, annotated datasets, making them susceptible to overfitting, especially when trained on small or biased datasets [18].

To overcome these limitations, contrastive learning has emerged as a promising approach for improving vision-based fall detection. This self-supervised technique enhances the model's ability to learn discriminative feature representations by contrasting positive (fall) and negative (non-fall) samples. Unlike fully supervised learning, contrastive learning does not rely heavily on labeled data, making it more adaptable to diverse environments [8]. Studies have shown that integrating contrastive learning with deep learning architectures, particularly CNNs, can yield significant performance improvements, with some models achieving accuracy rates exceeding 99% in fall detection tasks [19, 20]. By learning compact feature embeddings, contrastive learning facilitates more effective differentiation between fall-related and non-fall movements, reducing false positives and improving overall system robustness [21].

Despite these advancements, contrastive learning presents certain challenges. The method typically requires large batch sizes during training to effectively contrast multiple samples, increasing computational demands. Additionally, the quality of learned representations heavily depends on the selection of positive and negative pairs inadequate selection may lead to suboptimal feature extraction. Moreover, biases in the training dataset can affect the model's ability to generalize across different environments and camera perspectives, necessitating careful dataset curation [22]. Recent advancements in contrastive learning have significantly enhanced 3D vision-based fall detection systems by improving feature representation, robustness, and generalization. These improvements stem from innovative architectures and data augmentation techniques, which allow for better discrimination between falls and regular activities, addressing the challenges posed by the complexity of human motion in diverse environments. Specifically, contrastive learning has been shown to improve feature representation by learning more discriminative features through the comparison of positive (similar) and negative (dissimilar) samples, which is crucial for accurately identifying falls. The Group3AD model, for example, uses an Intercluster Uniformity Network to cluster point cloud data, boosting the model's capacity to identify abnormal behaviors like falls from routine actions [23]. Moreover, multi-stream frameworks leveraging both spatial and temporal features have been proposed, utilizing human skeleton data and motion history, which have proven to enhance the accuracy of fall detection systems [24].

To further improve the robustness of these models, Generative Adversarial Networks (GANs) have been utilized to augment training datasets, which help the model adapt to diverse fall scenarios [25]. In addition, lightweight 3D residual networks have been developed to reduce computational complexity while maintaining high accuracy, thus improving the robustness of fall detection systems, particularly in real-time applications [26]. These lightweight models allow for more efficient processing of large-scale video data, ensuring real-time fall detection without sacrificing performance.

Furthermore, efforts to enhance the generalization capabilities of these systems have focused on creating diverse training datasets, incorporating various angles and environmental conditions. By using more comprehensive datasets, models have been able to generalize better to real-world scenarios, as demonstrated by the high accuracy rates achieved in studies conducted in varied settings [26, 27].

Despite these advancements, challenges remain, particularly in ensuring that models can adapt to new environments and variations in human behavior. For instance, variations in how individuals fall and environmental factors such as lighting conditions or occlusions can still lead to false positives or missed detections. Future research should focus on enhancing the adaptability of these systems, especially in terms of ensuring high performance across diverse settings and addressing computational constraints that may limit the scalability of these methods.

3. Method

This study utilizes a contrastive learning-based model for fall detection, leveraging 3D body articulation data from depth cameras. The task is framed as binary classification, distinguishing falls from other activities. The model's core innovation lies in the contrastive loss function, which encourages the model to cluster fall instances (positive samples) together and separate them from non-fall activities (negative samples). By learning from 3D coordinate sequences, the model builds a rich embedding space that improves the representation of human actions, enhancing both feature extraction and classification performance. We generalize the workflow in Figure 1.

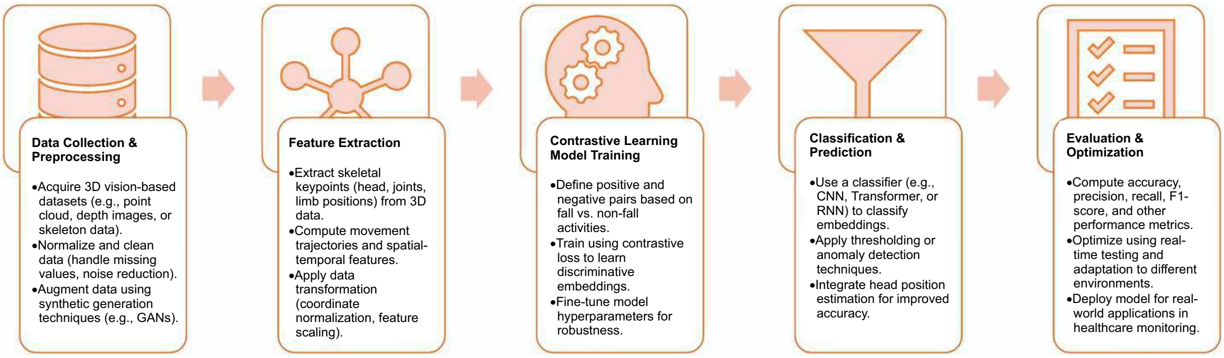
Data Collection & Preprocessing
•Acquire 3D vision-based datasets (e.g., point cloud, depth images, or skeleton data).
•Normalize and clean data (handle missing values, noise reduction).
•Augment data using synthetic generation techniques (e.g., GANs).
Feature Extraction
•Extract skeletal keypoints (head, joints, limb positions) from 3D data.
•Compute movement trajectories and spatial-temporal features.
•Apply data transformation (coordinate normalization, feature scaling).
Contrastive Learning Model Training
•Define positive and negative pairs based on fall vs. non-fall activities.
•Train using contrastive loss to learn discriminative embeddings.
•Fine-tune model hyperparameters for robustness.
Classification & Prediction
•Use a classifier (e.g., CNN, Transformer, or RNN) to classify embeddings.
•Apply thresholding or anomaly detection techniques.
•Integrate head position estimation for improved accuracy.
Evaluation & Optimization
•Compute accuracy, precision, recall, F1-score, and other performance metrics.
•Optimize using real-time testing and adaptation to different environments.
•Deploy model for real-world applications in healthcare monitoring.

Fig. 1. Workflow for Contrastive Learning-Based Fall Detection.

A Siamese network with a shared base model processes pairs of 3D data from different perspectives, increasing the accuracy of fall detection. The architecture is well explained in Figure 2. This self-supervised approach reduces the need for large labeled datasets, allowing the model to be trained on various fall scenarios captured from multiple angles. However, challenges related to computational efficiency and large batch sizes may affect real-time applications in healthcare environments.

The model processes 3D vision-based data, specifically the XYZ coordinates of human body joint movements. The dataset was collected using a ZED2 stereo camera to capture simulated actions performed by individuals of varying genders and ages. It consists of 12 videos, including 5 females and 7 males from MAScIR, who simulated both falls and normal activities of daily living (ADLs) [5].

Over 3000 frames were recorded per video in a controlled environment with consistent lighting and the camera was positioned 1.7 meters above the floor to simulate realistic indoor conditions. The data collection protocol followed guidelines from established fall detection databases, ensuring diverse fall scenarios.

To address the class imbalance, where falls (Label 1) are underrepresented compared to non-falls (Label 0), we employed the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic fall samples, thus improving model performance and ensuring a balanced dataset.

The preprocessing pipeline included noise reduction through Gaussian and median filtering, joint position normalization, and feature extraction (e.g., joint averages and movement trajectories). These steps enhanced data quality, enabling effective fall detection modeling. The simulation involved participants of various genders and ages to ensure comprehensive analysis across demographics. Each participant performed predefined actions under controlled conditions, with the ZED2 camera's optimal height and wide field of view capturing dynamic movements. Participant characteristics, such as age and physical attributes, were also documented to assess their impact on fall behavior.

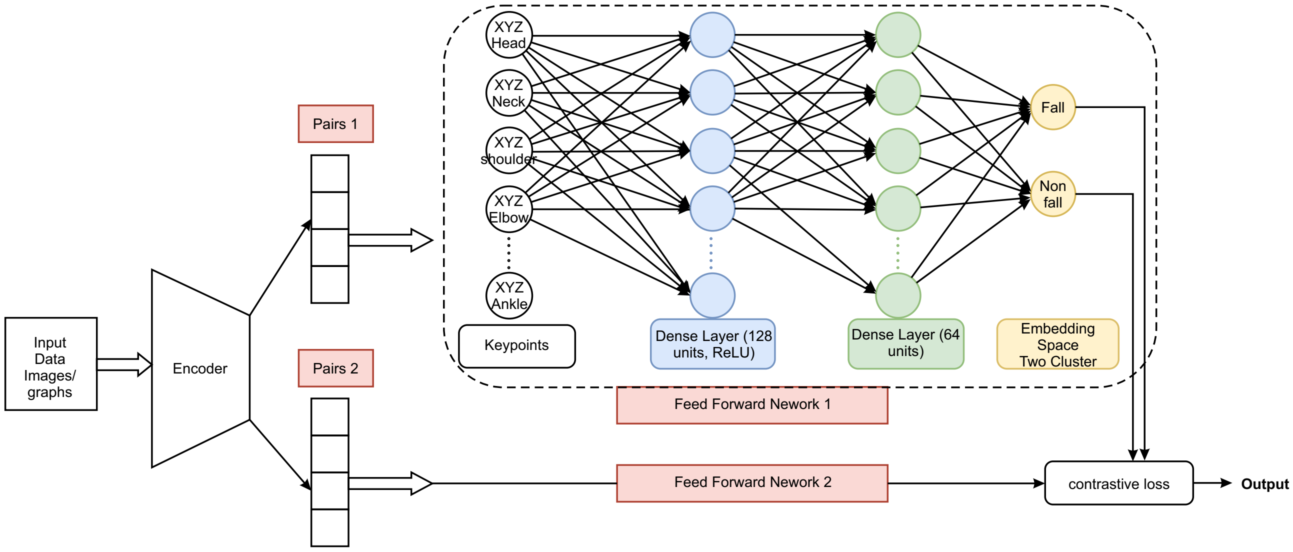

Fig. 2. Overview of the 3D Vision-Based Contrastive Learning Model for Fall Detection.

3.1. Problem formulation

Automatic fall detection using 3D skeleton-based motion analysis is a promising approach to enhancing safety in healthcare and assisted living environments. Given a set of **3D skeleton-based motion sequences**, we aim to learn an **embedding function** that effectively distinguishes **falling events** from **non-fall activities**. Let: $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ be the set of **input 3D skeletal coordinates**, where $x_i \in \mathbb{R}^{T \times J \times 3}$, with: $T =$ number of frames, $J =$ number of joints (e.g., head, shoulders, hips), Each joint has 3D coordinates (x, y, z); $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$, where $y_i \in \{0, 1\}$ represents the **fall state** (0: no fall, 1: fall).

We aim to learn a **contrastive embedding function** $f_\theta \colon \mathcal{X} \to \mathbb{R}^d$ that maps the input data into a **low-dimensional feature space** such that: **Falls (positive samples)** are mapped close together, **Non-falls (negative samples)** are pushed apart from falls.

3.2. Contrastive loss function

To ensure that the learned embedding space effectively distinguishes between **falling events** and **non-fall activities**, we employ a **contrastive loss function**. Specifically, we utilize **InfoNCE (Noise Contrastive Estimation)**, that encourages the model to bring similar motion patterns closer while pushing apart dissimilar ones. This optimization framework enhances the discriminative power of the embedding space, improving the model's ability to generalize across different motion sequences.

Formally, we define the loss function as follows:

$$\mathcal{L}_{\text{contrastive}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(f_\theta(x_i), f_\theta(x_i^+))/\tau)}{\sum_{j=1}^{M} \exp(\text{sim}(f_\theta(x_i), f_\theta(x_j^-))/\tau)}, \tag{1}$$

where: x_i^+ is a **positive pair** (fall sample close in feature space), x_j^- is a **negative pair** (non-fall sample far apart), $\text{sim}(a,b)$ is the **cosine similarity** between two embeddings:

$$\text{sim}(a,b) = \frac{a \cdot b}{\|a\|\|b\|}, \tag{2}$$

τ is a temperature hyperparameter controlling sensitivity, M is the number of negative pairs per sample.

3.3. 3D skeleton feature representation

To extract discriminative features, we use a spatial-temporal representation of skeletal motion:

$$H_t = W_1 X_t + b_1, \quad Z_t = \sigma(W_2 H_t + b_2), \tag{3}$$

where: H_t is the hidden representation at frame t, W_1, W_2 are learnable weight matrices, Z_t is the final feature embedding, σ is a non-linear activation function (e.g., ReLU).

We model the temporal dynamics using a Graph Convolutional Network (GCN) or Transformer Encoder, ensuring the embedding is robust across different body motions.

3.4. Model architecture

The main strategy that we tried to follow is explained by Algorithm 1.

Algorithm 1 Contrastive Learning for 3D Vision-Based Fall Detection.

```
Require: $\mathcal{X} = \{(X_i, Y_i, Z_i, \text{label}_i)\}$ (3D skeleton sequences), $\tau$ (temperature parameter), $\eta$ (learning rate);
Ensure: Optimized model parameters $\theta^*$;
Step 1: Data balancing
if imbalanced dataset then
    Apply to oversample for minority class (fall cases);
    Apply undersampling for majority class (non-fall cases);
Step 2: Model initialization
Randomly initialize model parameters $\theta$;
for $e = 1, \dots, E$
    Shuffle dataset and create balanced mini-batches $\mathcal{B}$;
    for all mini-batch $\mathcal{B} = \{(X_i, Y_i, Z_i, X_i^+, Y_i^+, Z_i^+, X_i^-, Y_i^-, Z_i^-)\}$
        for all sample $(X_i, Y_i, Z_i) \in \mathcal{B}$
            Compute embeddings: $Z_i = f_\theta(X_i, Y_i, Z_i)$;
            Compute positive similarity: $s^+ = \text{sim}(Z_i, Z_i^+)$;
            for all negative samples $(X_j^-, Y_j^-, Z_j^-)$ in $\mathcal{B}$
                Compute negative similarity: $s^- = \text{sim}(Z_i, Z_j^-)$;
            Compute contrastive loss:
                $\mathcal{L} = -\log \frac{\exp(s^+/\tau)}{\sum_j \exp(s^-/\tau)}$;
        Update model: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$;
return Optimized parameters $\theta^*$;
```

Feature extraction: Utilizes a Siamese network architecture with a shared base model to extract features from pairs of 3D coordinate sequences. This architecture allows the model to compare sequences of keypoint data, learning the relationship between different poses.

Contrastive loss: The model uses contrastive loss to learn embeddings that differentiate between similar pairs (e.g., two sequences from the same "falling" action) and dissimilar pairs (e.g., sequences from "not falling" actions). The contrastive loss function, being key to training siamese networks, is defined as follows:

$$L_{\text{contrastive}} = \frac{1}{2} \cdot \left(y \cdot D^2 + (1-y) \cdot \max(0, m - D)^2\right),$$

where: y is the binary label indicating whether the pair is: similar ($y = 1$), or dissimilar ($y = 0$); D is the Euclidean distance between the feature embeddings of the pair; m is a margin that defines how far apart dissimilar pairs should be in the embedding space.

This loss function encourages the network to minimize the distance D for similar pairs (falls) and to maximize the distance for dissimilar pairs (non-falls). Employs contrastive loss to learn embeddings that differentiate between similar and dissimilar pairs (falling vs. non-falling).

3.5. Training

Pair generation: During training, pairs of 3D coordinate sequences are created, where each pair is labeled to indicate whether it belongs to the same action (e.g., two "falling" actions) or different actions (e.g., "falling" vs. "walking"). If x_i and x_j represent two input sequences (from time steps i and j), the pair generation process can be represented as:

$$\text{Pair}(x_i, x_j) = \begin{cases} (x_i, x_j, y = 1) & \text{if both sequences belong to the same action class (e.g., both falls);} \\ (x_i, x_j, y = 0) & \text{if the sequences belong to different action classes} \\ & \text{(e.g., one is a fall and the other is a normal activity).} \end{cases}$$

Model training: Trains the model using these pairs with a contrastive loss function, adjusting weights to minimize the distance between similar pairs and maximize it between dissimilar pairs. This can be formalized as:

$$\min_{\theta} \sum_{(x_i, x_j, y)} L_{\text{contrastive}}(\theta, x_i, x_j, y),$$

where θ represents the parameters of the Siamese network. The training adjusts the model's weights to learn embeddings that correctly distinguish between different activities based on the 3D keypoint data.

3.6. Fall classification using nearest neighbor in contrastive space

After training, we classify falls using a simple **nearest neighbor approach** in the learned embedding space:

$$\hat{y} = \arg\min_{\hat{y} \in Y} \text{dist}(f_{\theta}(X), C_{\hat{y}}), \tag{4}$$

where $C_{\hat{y}}$ is the centroid of fall/non-fall samples in feature space, and distance metrics can be:

- Euclidean distance: $d(a, b) = \|a - b\|_2$.
- Mahalanobis distance: $d(a, b) = \sqrt{(a - b)^T \Sigma^{-1} (a - b)}$.

3.7. Evaluation metrics

To assess the performance, we use:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$.
- F1-Score: $2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.
- AUC-ROC Curve to analyze classification robustness.

4. Discussion

The application of contrastive learning to 3D vision-based fall detection has shown significant potential by improving the feature representation of human actions through the analysis of xyz coordinates of head joint movements. The model achieved an accuracy of 65.68%, demonstrating effective discrimination between non-falls and falls with high precision for non-falls as Figure 3. However, the model struggled with fall detection, evidenced by low recall and precision values for falls. This indicates that while the model effectively identifies non-fall actions, it misses many fall events, suggesting a need for further refinement. Future improvements could focus on addressing dataset imbalances and enhancing model sensitivity to falls to better meet real-world application needs.

The use of contrastive learning in vision-based fall detection systems offers significant benefits for real-world applications in elderly care settings. By reducing false positive rates, this approach enhances the reliability of fall detection, allowing for timely interventions when actual falls occur. Improved precision and recall metrics not only ensure that caregivers can respond effectively but also provide peace of mind to families, knowing their loved ones are monitored by an accurate system. Furthermore, the adaptability of this model allows it to be integrated with existing healthcare technologies, facilitating continuous learning from new data and user feedback. Notably, this approach also holds promise for the prediction of falls by identifying early indicators of risky movements, enabling a proactive stance in fall prevention. Ultimately, this innovative method not only promotes safety and independence for elderly individuals but also alleviates the workload on caregivers, leading to more efficient and effective care solutions.

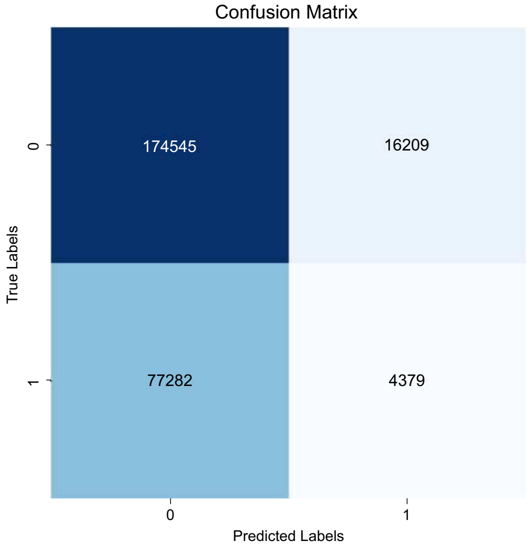

Fig. 3. Confusion matrix of the model.

In addressing the limitations related to model performance for our Contrastive Learning approach in vision-based fall detection, we acknowledge that our dataset is relatively small, comprising a total of 272 415 instances. The current metrics reflect this limitation, with an overall accuracy of 66%. Class-specific scores indicate a precision of 0.69 and recall of 0.92 for non-fall events (Class 0), but significantly lower values of 0.21 and 0.05 for fall events (Class 1). The F1-scores further illustrate this disparity, showing 0.79 for non-falls and only 0.09 for falls. The macro averages are 0.45 for precision, 0.48 for recall, and 0.44 for F1-score, while the weighted averages highlight the better performance for the non-fall class due to its larger support of 190 754 instances compared to just 81 661 for falls.

The small size of the dataset, particularly for fall events, is a significant factor contributing to the underperformance model in recognizing falls. This limitation can lead to an imbalance where the model is better trained on non-fall scenarios, resulting in missed detections and high false negatives for falls. Furthermore, the complexity and variability of fall scenarios may not be adequately captured in the available data.

To improve model performance, we plan to expand the dataset by incorporating additional fall scenarios and ensuring a more balanced representation of both classes. We will also explore techniques such as data augmentation and class weighting to enhance sensitivity to falls. By increasing the dataset size and diversity, along with refining our model architecture, we aim to enhance accuracy, precision, recall, and overall effectiveness in fall detection.

5. Results

Table 1. Extended classification report.

Class	Precision	Recall	F1-Score	Specificity	Support
0	0.69	0.92	0.79	0.21	190754
1	0.21	0.05	0.09	0.95	81661
Accuracy			0.66		272415
Macro avg	0.45	0.48	0.44	0.58	272415
Weighted avg	0.55	0.66	0.58		272415
Balanced Accuracy				0.54	
MCC				0.17	

We conducted a comprehensive evaluation of our contrastive learning-based model using a diverse and representative fall detection dataset. The model demonstrated an accuracy of 65.68%, with a precision of 21.27%, a recall of

5.36%, and an F1-Score of 8.57%, as presented in Table 1. These results highlight the model's capability in distinguishing fall incidents from non-fall activities. Despite the relatively low recall, which indicates a challenge in detecting all fall instances, the achieved precision suggests a reduction in false positives, ensuring that detected falls are indeed actual falls.

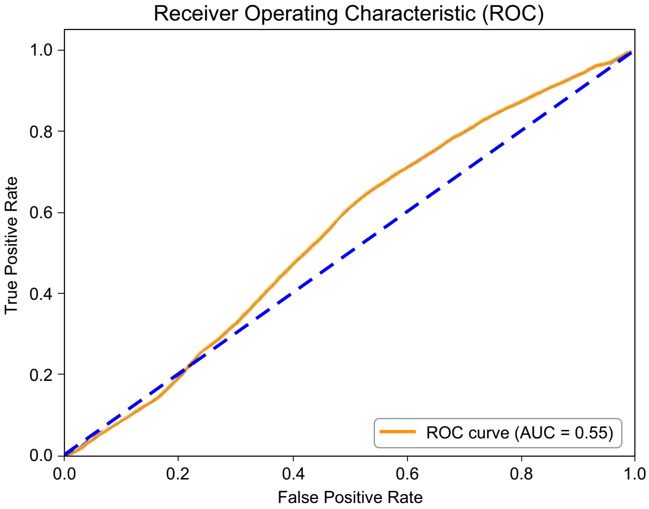

Fig. 4. ROC curve of contrastive model on our Dataset.

Furthermore, an in-depth investigation of the model's learned representations, shown using embedding space mapping, sheds light on its discriminative capacity. As seen in Figure 4, the contrastive learning strategy has considerably enhanced the model's capacity to distinguish different human movement patterns. This improvement is notably noticeable in the unique clustering of fall-related motions versus non-fall activities. A structured embedding space suggests that the model has successfully captured relevant latent properties, resulting in a more robust and interpretable decision-making process.

These findings underscore the potential of contrastive learning in enhancing fall detection models, particularly in scenarios where limited labeled data is available. Future work will focus on refining the training strategy to further improve recall while maintaining the achieved precision, thereby ensuring a more balanced performance across all evaluation metrics.

6. Conclusion

In order to address the ongoing problem of high false positive rates in conventional models, we presented a contrastive learning-based method for vision-based fall detection in this study. By utilizing contrastive learning to acquire strong feature embeddings that successfully distinguish between falls and non-falls, our approach significantly improves the accuracy and dependability of fall detection systems. We showed that this method can catch minute changes in human mobility by concentrating on 3D vision-based data, which makes it especially appropriate for real-world situations. To further improve our resilience systems, we also added a head position estimate model that uses 3D keypoints and has shown promise in processing both complete and incomplete data.

The findings of our work are consistent with recent advances in contrastive learning and 3D vision-based models, where approaches such as group-level feature grouping and multi-stream architectures have improved feature representation and classification performance. Our findings further emphasize the advantages of adopting synthetic data augmentation techniques like Generative Adversarial Networks (GANs) and lightweight deep learning architectures to improve the generalization capabilities of fall detection systems. Despite these benefits, certain obstacles persist, such as the necessity for large batch sizes during training, dataset imbalance issues, and the computing needs associated with high-dimensional contrastive learning models. These restrictions highlight the importance of additional modifications to make such models more scalable and efficient for real-time deployment.

Future research is going to emphasize on optimizing contrastive learning techniques by investigating innovative loss functions and architectural changes that can minimize computational complexity while maintaining high detection accuracy. Furthermore, using sequential models such as Recurrent Neural Networks (RNNs) or Transformer-based architectures may improve the system's ability to grasp temporal dependencies in fall events, hence increasing detection performance in dynamic settings. Furthermore, utilizing advanced data augmentation procedures and transfer learning approaches may solve dataset restrictions and increase model generalization across a variety of real-world contexts. Finally, implementing our technique in real-world healthcare settings for real-time fall detection remains an important step in determining its practical impact, ensuring its flexibility to various monitoring situations, and, ultimately, improving the safety and well-being of the elderly.

[1] World Health Organization (WHO). Falls. World Health Organization (2022). https://www.who.int/news-room/fact-sheets/detail/falls.

[2] Liu K.-C., Hung K.-H., Hsieh C.-Y., Huang H.-Y., Chan C.-T., Tsao Y. Deep-learning-based signal enhancement of low-resolution accelerometer for fall detection systems. IEEE Transactions on Cognitive and Developmental Systems. **14** (3), 1270–1281 (2021).

[3] Guendoul O., Abdelali H. A., Tabii Y., Thami R. O. H., Bourja O. Vision-based fall detection and prevention for the elderly people: A review & ongoing research. 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS). 1–6 (2021).

[4] Zobi M., Guendoul O., Tabii Y., Thami R. O. H. Vision-Based Fall Detection Systems Using 3D Skeleton Features for Elderly Security: A Survey. The International Conference on Intelligent System and Smart Technologies. 33–41 (2023).

[5] Guendoul O., Abdelali H. A., Tabii Y., Thami R. O. H., Bourja O. Enhanced Fall Detection and Prediction Using Heterogeneous Hidden Markov Models in indoor environment. IEEE Access. **12**, 187210–187219 (2024).

[6] Espinosa R., Ponce H., Gutiérrez S., Martínez-Villaseñor L., Brieva J., Moya-Albor E. 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 (2019).

[7] Gutiérrez J., Rodríguez V., Martin S. Comprehensive review of vision-based fall detection systems. Sensors. **21** (3), 947 (2021).

[8] Chen T., Kornblith S., Norouzi M., Hinton G. A simple framework for contrastive learning of visual representations. ICML'20: Proceedings of the 37th International Conference on Machine Learning. Article No.: 149, 1597–1607 (2020).

[9] He K., Fan H., Wu Y., Xie S., Girshick R. Momentum contrast for unsupervised visual representation learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 9729–9738 (2020).

[10] Chang W.-J., Hsu C.-H., Chen L.-B. A pose estimation-based fall detection methodology using artificial intelligence edge computing. IEEE Access. **9**, 129965–129976 (2021).

[11] Mutambik I. An Efficient Flow-Based Anomaly Detection System for Enhanced Security in IoT Networks. Sensors. **24** (22), 7408 (2024).

[12] Ikechukwu M. C., Wang J. Tackling Visual Illumination Variations in Fall Detection for Healthcare Applications. Proceedings of the 2024 IEEE International Conference on Advanced Computing (ICAC). 1–6 (2024).

[13] Guo T., Liu M., Liu H., Wang G., Li W. Improving self-supervised action recognition from extremely augmented skeleton sequences. Pattern Recognition. **150**, 110333 (2024).

[14] Alanazi T., Babutain K., Muhammad G. Mitigating human fall injuries: A novel system utilizing 3D 4-stream convolutional neural networks and image fusion. Image and Vision Computing. **148**, 105153 (2024).

[15] Dutt M., Gupta A., Goodwin M., Omlin C. W. An Interpretable Modular Deep Learning Framework for Video-Based Fall Detection. Applied Sciences. **14** (11), 4722 (2024).

[16] Wang Z., Ramamoorthy V., Gal U., Guez A. Possible life saver: A review on human fall detection technology. Robotics. **9** (3), 55 (2020).

[17] Salimi M., Machado J. J. M., Tavares J. M. R. S. Using deep neural networks for human fall detection based on pose estimation. Sensors. **22** (12), 4544 (2022).

[18] Yhdego H., Li J., Morrison S., Audette M., Paolini C., Sarkar M., Okhravi H. Towards musculoskeletal simulation-aware fall injury mitigation: transfer learning with deep CNN for fall detection. 2019 Spring Simulation Conference (SpringSim). 1–12 (2019).

[19] Balasubramanian R., Rathore K. Contrastive Learning for Object Detection. Preprint arXiv:2208.06412 (2022).

[20] Patel S. N., Lathigara A., Mehta V. Y., Kumar Y. A Survey on Vision-Based Elders Fall Detection Using Deep Learning Models. Lecture Notes in Electrical Engineering. **936**, 447–465 (2022).

[21] Jiang Y., Gong T., He L., Yan S., Wu X., Liu J. Fall detection on embedded platform using infrared array sensor for healthcare applications. Neural Computing and Applications. **36** (9), 5093–5108 (2024).

[22] Le-Khac P. H., Healy G., Smeaton A. F. Contrastive Representation Learning: A Framework and Review. IEEE Access. **8**, 193907–193934 (2020).

[23] Zhu H., Xie G., Hou C., Dai T., Gao C., Wang J., Shen L. Towards High-resolution 3D Anomaly Detection via Group-Level Feature Contrastive Learning. MM'24: Proceedings of the 32nd ACM International Conference on Multimedia. 4680–4689 (2024).

[24] Mobsite S., Alaoui N., Boulmalf M., Ghogho M. A Deep Learning Dual-Stream Framework for Fall Detection. 2023 International Wireless Communications and Mobile Computing (IWCMC). 1226–1231 (2023).

[25] Verma N., Mundody S., Guddeti R. M. R. An Efficient AI and IoT Enabled System for Human Activity Monitoring and Fall Detection. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). 1–6 (2024).

[26] Peng X., Li W. Fall detection algorithm based on lightweight 3D residual network. Proceedings of the SPIE Conference on Machine Vision Applications. **12602**, 126023L (2023).

[27] Priyanka K. P. M., Kumar K. A Review on Recent Developments on Detection of Fall. Journal of Trends in Computer Science and Smart Technology. **5** (2), 119–135 (2023).

[28] Yang Y., Yang H., Liu Z., Yuan Y., Guan X. Fall detection system based on infrared array sensor and multi-dimensional feature fusion. Measurement. **192**, 110870 (2022).

Самостійне контрастне навчання для виявлення падінь з використанням артикуляції тіла на основі 3D-зору

Гендоул У., Табії Ю., Улад Хадж Тхамі Р.

Лабораторія ADMIR, Національна школа комп'ютерних наук та системного аналізу (ENSIAS),
Університет Мохаммеда V, Рабат 10000, Марокко
oumaima_guendoul@um5.ac.ma

У цій статті представлено підхід до математичного моделювання виявлення падінь з використанням рамки контрастного навчання на основі 3D-зору. Традиційні моделі мають проблеми з високою кількістю хибнопозитивних результатів та поганим узагальненням у різних середовищах. Щоб вирішити цю проблему, пропонується модель контрастного навчання із самоконтролем, яка відображає 3D послідовності рухів скелета в низьковимірний простір вбудовування, оптимізуючи розділення ознак між падіннями та непадіннями. Запропонований метод використовує просторово-часове моделювання та функцію контрастних втрат на основі косинусної подібності для покращення розрізнення. Використовуючи графічне представлення ознак, модель забезпечує надійну продуктивність навіть за відсутніх або зашумлених даних. Експериментальні результати на еталонних наборах даних виявлення падінь демонструють значне зменшення хибнопозитивних результатів, зберігаючи при цьому високу точність, що робить фреймворк добре придатною для реальних застосувань у сфері охорони здоров'я.

Ключові слова: *математичне моделювання, контрастне навчання, 3D-зір, аналіз рухів скелета, охорона здоров'я, виявлення падінь.*