Human Fall Detection – Multimodality Approach

Morteza Mogharrab
Department of Computing Science
University of Alberta
Edmonton, Canada
mogharra@ualberta.ca

Ritika Ritika
Department of Computing Science
University of Alberta
Edmonton, Canada
ritika7@ualberta.ca

Sai Sarath Movva
Department of Computing Science
University of Alberta
Edmonton, Canada
saisarat@ualberta.ca

***Abstract*—Falls among the elderly pose a significant health risk, leading to severe injuries or even fatalities. Early and accurate detection of falls is crucial for enabling timely intervention and preventing adverse outcomes. This research conducts a comprehensive analysis of existing multimodal approaches for human fall detection that integrate data from diverse sources such as video cameras, wearable sensors, and ambient sensors. The key findings highlight the superiority of multimodal fusion techniques over single-modality approaches in enhancing fall detection accuracy and robustness. Through a critical review and synthesis of prior studies, it was found that among novel state-of-the-art methods, Single LSTM and CNN 1D, which rely on single sensor data, have the poorest performance. In contrast, the Graph Convolutional Network (GCN) + Transformer model outperforms other models, reaching an F1-score of 1 based on the NTU video-based dataset. However, this perfect score does not imply that the model has no limitations or problems. In this report, we also discuss why the GCN + Transformer model performs better, the process of its implementation, its limitations and problems, its strengths, and what can be done in the future to enhance its capabilities for more complex real-world scenarios that may involve occlusion, etc.**

Index terms - Human Fall Detection, Multimodal Data Fusion, Federated Learning, Knowledge Distillation, Transformers, Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), Vision Transformers (ViT)

I. Introduction

The alarming rise in annual falls among the elderly has fueled research into reliable, efficient fall detection systems, a critical need for our aging population. A significant portion, 28% to 42% of those over 65, fall each year, making falls a leading cause of serious injury and death, especially for those above 79. [7, 11] Conventional methods using video cameras, wearable sensors, or floor-mounted devices often have limitations. These limitations include the need for multiple devices, restricted data collection areas, and solely focusing on physical movements without considering individual user characteristics. To address these limitations and improve overall performance, multimodal approaches that integrate data from various sources are a promising solution. This research project will critically analyze existing multimodal fall detection methods, followed by methods for improving the best available solutions. This will be done by synthesizing current knowledge, identifying best practices, and proposing methodological improvements. The knowledge gained from this project will guide researchers and practitioners in designing innovative solutions that leverage the strengths of multimodal data while addressing challenges like data heterogeneity and the need for personalization.

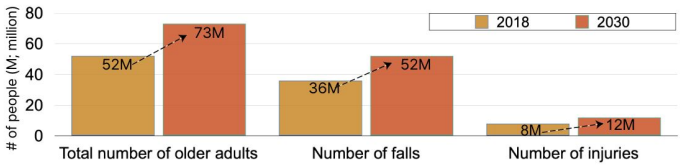

Fig. 1: Elderly Population: Falls and Injury Rates - Expected Increase Between 2018 and 2030 [8]

II. Related Work

A. Existing Methods

The current literature is focused on a variety of methodologies for fall detection, categorized by their primary data source: sensor-based, vision-based, and multimodal. Sensor-based approaches (employed in [1, 3, 6]) rely on data from accelerometers, gyroscopes, and similar sensors to capture physical movements, with [1] incorporating biometric information like age and gender for personalized fall detection. Vision-based approaches (explored in [3, 5]) utilize cameras to capture video data, where [3] leverages human pose estimation to track subjects and extract features relevant to falls, while [5] investigates Convolutional Neural Networks (CNNs) for extracting informative patterns from camera images. Multimodal fall detection combines sensor data with visual data (addressed in [2, 5]), with [2] proposing an input-level fusion approach merging data streams before feature extraction, and [5] exploring separate feature extraction for each data type, followed by later fusion.

Deep learning architectures are a popular choice for fall detection models (used in [1, 4, 5, 6]). For instance, [1] applies a Temporal Fusion Transformer (TFT) for analyzing time series sensor data, while [4] and [6] explore Transformers for classifying human pose key points. [5] investigates a wider range of models including Neural Networks, XGBoost, CatBoost, and CNNs for processing both sensor and camera data. Notably, [2] introduces a federated learning framework that addresses privacy concerns during fall detection using multimodal data, allowing training of a fall detection model while keeping user data on their devices. These papers high-

light the ongoing exploration of diverse methodologies for fall detection, with the selection of the most suitable technique depending on factors such as the type of data available, the need for privacy preservation, and the desired level of accuracy in fall detection.

B. Available Datasets

This section explores the specific data sources leveraged by the reviewed papers, highlighting the fascinating diversity employed in this field.

- Sensor-based Datasets:

Several studies leverage sensor data for fall detection, with varying degrees of complexity [1, 4, 6]. For instance, [1] focuses on publicly available datasets like SmartFall, Notch, DLR, and MobiAct, which primarily contain readings from accelerometers and gyroscopes worn by participants. This approach offers a simple and widely accessible method for fall detection. Building upon sensor data, [4] utilizes the University of Rzeszow Fall Detection (URFD) Dataset, which includes accelerometer data alongside depth and RGB images from Kinect cameras for a more comprehensive view. Similarly, [6] employs the KFall dataset, specifically designed for pre-fall detection using data from a nine-axis inertial sensor worn on the lower back.

- Multimodal Datasets:

Several studies advocate for a multimodal approach to fall detection, incorporating data from various sources beyond sensors. For instance, [2] and [3] leverage the UP-Fall dataset, which encompasses a diverse range of data modalities, including wearable sensors capturing acceleration, angular velocity, and light levels at various body locations, infrared sensors detecting activity disruptions, and cameras providing visual data. This richness of data allows for a more robust and comprehensive understanding of fall events. Pushing the boundaries even further, the study by [5] also utilizes the UP-Fall Detection dataset but takes an innovative approach by integrating data from wearable sensors placed on multiple body parts, an EEG headset for monitoring brainwave activity, infrared sensors, and dual cameras, thereby enabling a holistic and multifaceted analysis of fall events through a comprehensive multimodal fusion strategy.

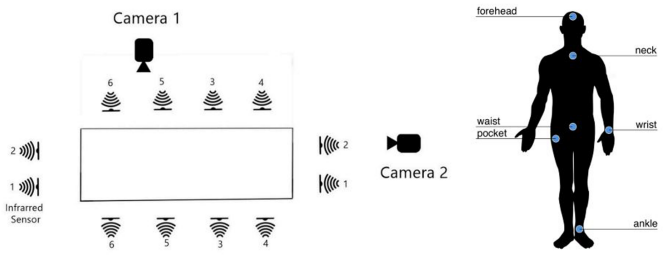

Fig. 2: The monitoring site utilizes a network of 8 sensors: 2 cameras and 6 infrared sensors for perimeter security, along with 6 wearable sensors for human monitoring.

Activity ID	Description	Durations
1	Falling forward using hands	10
2	Falling forward using knees	10
3	Falling backward	10
4	Falling sideways	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 I: Available activities in the UP-Fall dataset with their ID, description, and duration of videos.

- Personalization and Dataset Scope:

Interestingly, some studies tend to incorporate user-specific biometric data (age, gender, height, weight) from certain datasets when available, highlighting a potential avenue for improving fall detection accuracy by tailoring it to individual characteristics [1]. Studies by [3] and [6] try to focus on relatively controlled settings with a limited number of subjects and activities. Conversely, [4] and [5] utilize larger datasets with a wider variety of daily activities, offering a more generalizable view of fall detection across diverse scenarios.

C. Common themes, Contradictions, and Gaps

Several key themes emerge from the literature on fall detection research. This research area is making significant progress towards improved accuracy, user privacy, and real-world applicability. There is a strong consensus on the benefits of data fusion, where combining information from wearable sensors and cameras leads to better fall detection than relying on a single data source ([2, 5]). This paves the way for future systems that leverage multiple sensors for a more comprehensive understanding. Additionally, privacy concerns, especially with camera data, are being addressed by techniques like federated learning (proposed in [2]). This approach allows training models while keeping user data on their devices, offering a promising solution to balance privacy and effectiveness.

However, challenges remain. While many studies achieve high accuracy in controlled environments [3, 4], the limitations of current datasets are highlighted. These datasets often lack the complexities of real-world scenarios, such as imbalanced fall occurrences or varying lighting conditions [3, 4]. This necessitates the development of more diverse and realistic datasets to ensure models can generalize well to real-world situations. Furthermore, ongoing exploration of new deep-learning architectures for fall detection is underway [4, 6]. Studies showcase the potential of transformers and knowledge distillation techniques for achieving high accuracy and efficiency [4, 6]. However, these approaches often rely on accurate human pose estimation, which can be difficult in real-world scenarios with occlusions, as highlighted in [4].

Several interesting contradictions and gaps emerge when considering future directions in fall detection research. One key issue is data imbalance in current datasets. As identified in

[3], these datasets often have far fewer fall examples compared to everyday activities. This can lead to models biased towards non-fall events, potentially missing real falls. Additionally, a trade-off exists between privacy and generalizability. While federated learning offers strong privacy protections, it can also lead to lower accuracy due to the variations in data across different users' devices [2]. Similarly, knowledge distillation, a technique showcased in [6] for reducing the computational load in fall detection systems, might come at the cost of some accuracy. Finding the right balance between these competing factors is essential for developing practical fall detection systems.

III. Methodology

A. Evaluation Matrices

In this Project, we assessed fall detection frameworks' effectiveness as a binary classification problem (fall vs. non-fall) using various metrics. Metrics like accuracy, precision, recall, F1-score, and AUC-ROC gauged the model's ability to identify true falls and avoid false alarms, while also considering class imbalance in the data. Additionally, confusion matrices provided a detailed breakdown of the model's performance across different fall and non-fall scenarios, revealing strengths, weaknesses, and areas for improvement. Finally, testing on diverse datasets with simulated and real-world falls ensured the framework's robustness and ability to generalize to real-life situations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table II presents a comprehensive analysis of the performance of various machine-learning models in detecting human falls across several datasets. While more detailed versions of the analysis were conducted, this table summarizes the best and worst performances for each method and dataset.

The key findings from the table reveal several notable trends. The top-performing models include the Graph Convolutional Network (GCN) + Transformer [15] (rely only on video data), which achieved a perfect score of 1.0 across all evaluation metrics on the NTU dataset, and the Gramian Angular Field (GAF) model [16] with time series and C2 (camera 2) data fusion, which achieved an F1-score of 0.9992, precision of 0.9984, recall of 1.0, and accuracy of 0.9993 on the UP-Fall dataset. Additionally, the CNN-based data fusion (S (sensor) + C1 + C2) model performed exceptionally well on the UP-Fall dataset, with an F1-score of 0.9955, precision of 0.9956, recall of 0.9956, and accuracy of 0.9956.

Transformer-based models, such as the Transformer and Transformer (Temporal Fusion), also demonstrated strong performance across various datasets, including NTU, MobiAct, Kfall, UR, Notch, and DLR. The Transformer model achieved an F1-score of 0.9910 on the NTU dataset, while the Transformer (Temporal Fusion) model achieved F1-scores ranging from 0.8770 to 0.9702 on the MobiAct, Notch, and DLR datasets. The table also highlights the performance of Vision Transformer (ViT) models, with the Vision Transformer (Tiny) model achieving an F1-score of 0.9384 on the Kfall dataset and the CNN + ViT Knowledge Distillation (PreFallKD) model achieving an F1-score of 0.9266 on the same dataset.

In contrast, LSTM-based models exhibited varying performance. The LSTM (5 features-based) model performed well on the UP-Fall dataset, with an F1-score of 0.9256, but the LSTM (Stacked) and LSTM (Single) models struggled on the SmartFall and MobiAct datasets, respectively, with F1-scores of 0.1378 and 0.0040. The table also includes the performance of other models, such as the Logistic Regression model, which achieved an F1-score of 0.6065 on the UP-Fall dataset, and the Temporal Attention Convolutional Neural Networks (TACN) model, which had the lowest performance on the DLR dataset, with an F1-score of 0.0825. When examining the trends and comparisons, the table reveals that the models generally performed better on the NTU, UP-Fall, and Kfall datasets compared to the other datasets, while the SmartFall and DLR datasets posed more challenges for the models, with lower F1-scores across the board.

In terms of model performance comparison, Transformer-based models, such as the GCN + Transformer and Transformer, consistently outperformed other models across multiple datasets. Vision Transformer (ViT) models also showed promising results on the Kfall dataset. LSTM-based models exhibited varying performance, with the LSTM (5 features-based) performing well on the UP-Fall dataset, but the LSTM (Stacked) and LSTM (Single) models struggling on the SmartFall and MobiAct datasets, respectively. Regarding the evaluation metric trends, the top-performing models achieved near-perfect or perfect scores across all evaluation metrics, indicating their strong overall performance. However, some models, like the Logistic Regression and TACN, had significant discrepancies between their F1 scores, precision, recall, and accuracy, suggesting potential imbalances in their performance.

IV. Results and Discussion

Benchmark evaluations demonstrate the superior performance of the GCN+Transformer model compared to other models in fall detection tasks. However, this evaluation has been conducted on specific datasets and under controlled conditions, which may not accurately reflect the challenges and complexities of real-world deployment scenarios. Simply put, the model's performance is highly likely to face significant challenges when deployed in practical, real-world settings. This section delves into an in-depth analysis of

Dataset	Model	F-1 score	Precision	Recall	Accuracy
NTU	Graph Convolutional Network (GCN) + Transformer	1.0000	1.0000	1.0000	1.0000
UP-Fall	Gramian Angular Field (GAF) (Time Series and C2 data fusion)	0.9992	0.9984	1.0000	0.9993
UP-Fall	Gramian Angular Field (GAF) (Time Series and C1 data fusion)	0.9985	0.9984	0.9987	0.9987
UP-Fall	CNN-based data fusion (S + C1 + C2)	0.9955	0.9956	0.9956	0.9956
NTU	Transformer	0.9910	0.9910	0.9910	0.9910
MobiAct	Transformer (Temporal Fusion)	0.9702	0.9702	0.9702	0.9883
Kfall	Vision Transformer (Tiny)	0.9384	0.9202	0.9573	0.9836
Kfall	CNN + ViT Knowledge Distillation (PreFallKD)	0.9266	0.9062	0.9479	0.9805
UP-Fall	LSTM (5 features-based)	0.9256	0.8976	0.9562	0.9822
UR	Graph Convolutional Network (GCN) + Transformer	0.9030	0.9250	0.9000	0.9000
UR	Transformer	0.8930	0.9120	0.9000	0.9000
Notch	Transformer (Temporal Fusion)	0.8770	0.8766	0.8775	0.9798
Kfall	CNN (Baseline)	0.8589	0.9236	0.8027	0.9656
DLR	Transformer (Temporal Fusion)	0.7187	0.8393	0.7017	0.9490
UP-Fall	Logistic Regression	0.6065	0.6606	0.5445	0.9261
SmartFall	Transformer (Temporal Fusion)	0.4617	0.4075	0.5326	0.9314
Notch	CNN (1D)	0.1859	0.1155	0.4754	0.5899
SmartFall	LSTM (Stacked)	0.1378	0.0762	0.7155	0.5057
DLR	Temporal Attention Convolutional Neural Networks (TACN)	0.0825	0.1035	0.1123	0.0042
MobiAct	LSTM (Single)	0.0040	0.0567	0.0588	0.0354

TABLE II: The table presents a comprehensive analysis of the performance of various machine learning models in detecting human falls across several datasets. The evaluation metrics used to assess the models' effectiveness include F1-score, precision, recall, and accuracy.

the GCN+Transformer architecture, exploring the factors contributing to its outstanding performance. Additionally, it examines the model's training process, key strengths and limitations, and potential solutions to mitigate its possible downsides when deployed in complex real-world scenarios. By understanding the intricacies of this model, we can better appreciate its capabilities and identify areas for further improvement, ultimately paving the way for more robust and reliable fall detection solutions applicable to diverse real-world environments.

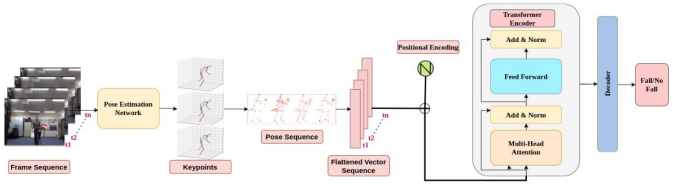

Fig. 3: Fall detection using a transformer-based pipeline

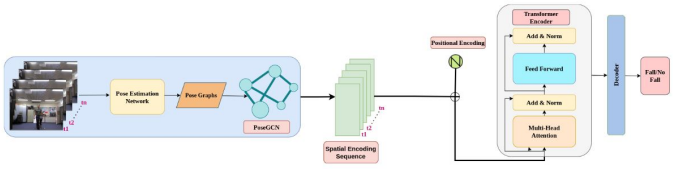

Fig. 4: Fall detection with GCN and Transformer combination

A. Dataset

The training of this model has been based on two datasets: the University of Rzeszow Fall Detection (URFD) dataset and the NTU RGB+D dataset [9]. The URFD dataset was meticulously curated by Kwolek and Kepski in 2014. Comprising a diverse array of 40 daily living activities and 30 explicit fall sequences, this dataset provides a balanced representation of both positive (fall) and negative (non-fall) scenarios. The data acquisition process leveraged the cutting-edge Kinect camera technology, simultaneously capturing depth and RGB images from two distinct viewpoints. Furthermore, the dataset incorporates accelerometer data, enriching the contextual information and enabling more comprehensive analyses.

On the other hand, the NTU RGB+D dataset, a resource encompassing a staggering 56,880 samples, contributes a broader scope by spanning 60 distinct action classes. This extensive collection not only encompasses daily behaviors but also incorporates a wide range of health-related actions, ensuring a holistic representation of human movement and activity. Notably, the dataset was meticulously constructed through the participation of 40 individuals, further augmenting its diversity and real-world applicability. The combination of these two datasets, each offering unique strengths and characteristics, facilitated a rigorous evaluation of the proposed model's performance, enabling comprehensive analyses across various fall detection scenarios, action classes, and contextual settings.

B. Implementation Process of GCN+Transformer Model

The GCN+Transformer model follows a two-stage pipeline for fall detection using human pose keypoint data. [12] The first stage involves spatial encoding using a Graph Convolutional Network (GCN), and the second stage performs temporal encoding using a Transformer architecture.

In the initial stage, the model takes a sequence of human pose key points extracted from video frames as input. [13] These key points are processed through a GCN layer, which treats the key points as nodes in a graph and applies graph convolutions to learn meaningful spatial representations. The

graph convolution operation can be mathematically expressed as:

$$X' = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}XW)$$

Here, X is the input feature matrix (key points), $\hat{A}$ is the adjacency matrix representing the graph structure, $\hat{D}$ is the degree matrix (diagonal matrix with node degrees), W is the learnable weight matrix, and σ is a non-linear activation function. The GCN layer effectively encodes the spatial context and dependencies between key points, capturing the structural and positional information crucial for understanding human poses and actions. The output of this layer is a sequence of spatial embeddings, one for each frame in the input sequence.

In the second stage, these spatial embeddings are fed into a Transformer architecture for temporal encoding. The core component of the Transformer is the Multi-Head Attention mechanism, which can be expressed mathematically as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Where Q, K, V are the Query, Key, and Value matrices, W_i^Q, W_i^K, W_i^V are learnable weight matrices, and d_k is the dimension of the Key vectors.

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

$$X_t = \text{GCN}(X_t)$$

$$Y_t = \text{Transformer}([X_1, X_2, \ldots, X_t])$$

The Multi-Head Attention mechanism allows the Transformer to capture long-range temporal dependencies and patterns in the sequence of spatial embeddings, effectively encoding the temporal information crucial for fall detection. Across both datasets, training parameters included 30 epochs, a batch size of 32, and an initial learning rate of 0.0001. The Reduce-On-Plateau scheduler facilitated adaptive learning rate adjustments during training. [10] The robustness of the GCN+Transformer model was evaluated across various scenarios, including varying camera angles, partial occlusions, and frame quantities. Assessments showed that the model outperformed previous Transformer-only models, demonstrating superior precision, recall, F1 score, and accuracy, particularly in challenging conditions like occlusions and diverse viewing angles. The fusion of spatial and temporal information through GCN and Transformer layers allowed the model to capture intricate dependencies and patterns in human pose keypoint data, leading to enhanced performance in fall detection and other human action recognition tasks.

C. Training process in more detail

In the experimental setup, we conducted extensive explorations to determine the optimal hyperparameter configuration for the GCN+Transformer model. A conventional 60-25-15 training-validation-testing split was adopted for both datasets, ensuring a robust and unbiased evaluation of the model's capabilities.

The training phase involved a meticulous examination of various architectural configurations for the Transformer component. We systematically adjusted hyperparameters such as the number of layers, attention heads, and feed-forward dimensions, thoroughly exploring their impact on the model's performance. After a comprehensive analysis, the optimal configurations that emerged as the best-performing architectures were a Transformer with 2 layers, 8 attention heads, and a feed-forward dimension of 128 as well as a Transformer with 2 layers, 4 attention heads, and a feed-forward dimension of 256. This configuration not only exhibited superior performance on the training and validation sets but also demonstrated remarkable generalization capabilities on the held-out test sets. Consequently, this optimal Transformer architecture was integrated as the backbone for the temporal encoding stage within the GCN+Transformer model, ensuring a harmonious fusion with the spatial encoding capabilities provided by the GCN layer.

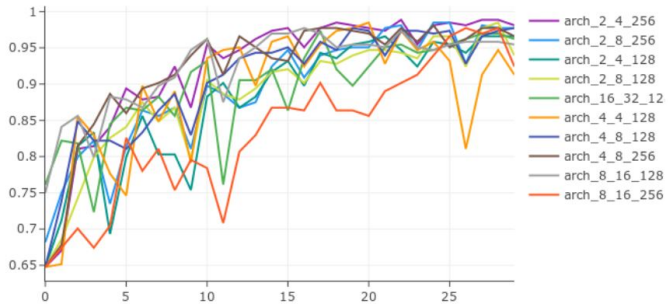

Fig. 5: Validation accuracy per epoch

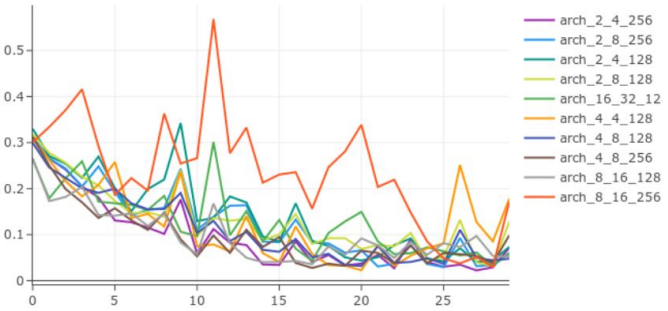

Fig. 6: Validation loss per epoch

D. Additional Considerations

We also conducted an extensive ablation study to evaluate the GCN+Transformer model's performance across varying

camera angles and its generalization capabilities. When trained and tested on NTU View 1, representing a frontal perspective similar to a social robot's vantage point, the model achieved perfect precision, recall, F1 score, accuracy, and geometric mean scores of 1.0. However, when tested on NTU View 2 and View 3, which introduced angular deviations from the training viewpoint, the model's performance metrics experienced slight declines, albeit maintaining a high degree of accuracy and effectiveness. This robustness underscores the model's strong generalization capabilities and adaptability to diverse perspectives, a critical trait for practical applications where camera placement may be constrained or dynamic.

We also assessed the model's resilience under various occlusion patterns, simulating three distinct scenarios: Type 1 (lower-body occlusion), Type 2 (torso occlusion), and Type 3 (upper limb occlusion). In Type 1 scenarios, where crucial joints for fall detection remained visible, the model exhibited remarkable resilience, achieving high-performance metrics. However, in Type 2 cases, where torso-related key points were obstructed, the model's performance experienced a notable decline, highlighting its sensitivity to obstructions in this region. Intriguingly, in Type 3 scenarios, the model demonstrated commendable performance, facilitated by the preserved visibility of lower limb and torso key points. This study emphasized the importance of unobstructed visibility of specific body regions, particularly the torso and lower limbs, for optimal fall detection performance.

To investigate the optimal frame quantity for achieving peak performance, we evaluated the model's performance across skip rates of 1, 7, and 11. At a skip rate of 1, where no frames were skipped, the model exhibited robust performance, attributed to its ability to leverage the full temporal resolution and capture subtle movements. However, as the skip rate increased to 7, the model's performance experienced a noticeable decline, and at a skip rate of 11, the geometric mean plummeted to 0, indicating a complete failure to accurately detect falls. This study provided valuable insights into the trade-off between computational efficiency and model performance, emphasizing the need for judicious selection of the skip rate to maintain optimal anomaly detection capabilities. Processing consecutive frames without skipping yielded the most reliable and accurate fall detection performance.

E. Strength of GCN+Transformer model

The model's strength lies in its two-fold encoding process: spatial encoding using GCNs and temporal encoding using Transformers. The GCN layer is designed to capture the spatial relationships and dependencies between the key points in each individual frame. It treats the key points as nodes in a graph and applies graph convolutions to learn meaningful spatial representations, effectively encoding the spatial context and dependencies between key points, and capturing the structural and positional information crucial for understanding human poses and actions. On the other hand, the Transformer architecture is well-suited for capturing long-range temporal dependencies and patterns in sequential data, such as the sequence of spatial embeddings obtained from the GCN layer. The Multi-Head Attention mechanism allows the Transformer to attend to different parts of the input sequence in parallel, effectively capturing long-range dependencies and temporal patterns. This fusion of spatial and temporal information [14] enables the model to capture the intricate dependencies and patterns present in the human pose keypoint data, leading to superior performance in fall detection and other human action recognition tasks."

F. Limitations

Based on our interpretability analysis, domain expert evaluation, sensitivity analysis, and stress testing, we identified several potential issues in the GCN+Transformer model that could impede its performance when integrated into real-world scenarios. These issues are detailed below.

1) Reliance on Human Pose Estimation (HPE) Key Points: **Limitation:** The model's performance heavily relies on the quality and availability of HPE key points, which means valuable contextual information captured by RGB data is not utilized.

Potential Mitigation: Explore multi-stream architectures that can effectively fuse RGB data with pose key points, leveraging both sources of information for more fine-grained decision-making.

Alternative Approach: Investigate end-to-end models that can directly process raw video data, eliminating the need for explicit pose estimation as a preprocessing step.

2) Additional Preprocessing Step and Latency Concerns: **Limitation:** The dependence on HPE introduces an additional preprocessing step before inference, potentially affecting real-time implementation and suitability for low-latency applications.

Potential Mitigation: Optimize the pose estimation pipeline and leverage hardware acceleration (e.g., GPUs) for efficient preprocessing and inference.

Alternative Approach: Explore lightweight architectures or model compression techniques to reduce computational overhead and enable real-time performance on resource-constrained devices.

3) Dependence on Training Data Diversity and Representativeness: **Limitation:** The model's effectiveness is contingent on the diversity and representativeness of the training data, struggling to generalize to unseen situations when the dataset lacks comprehensive coverage.

Potential Mitigation: Employ advanced data augmentation techniques, such as synthetic data generation and domain randomization, to artificially expand the diversity of the training data.

Alternative Approach: Investigate few-shot or meta-learning approaches that can adapt to new scenarios with limited data, reducing the reliance on large, diverse datasets.

4) Sensitivity to Occlusions: **Limitation:** The model's performance may be significantly impacted by occlusions, highlighting the need for enhanced robustness to various occlusion scenarios.

Potential Mitigation: Explore attention mechanisms and self-supervised learning techniques that can learn to focus on relevant body parts and handle partial occlusions effectively.
Alternative Approach: Investigate hybrid approaches that combine the GCN+Transformer model with depth or infrared sensors, providing additional modalities less susceptible to occlusions.

5) Transformer Complexity and Training Challenges: **Limitation:** Transformers are inherently complex, making them challenging to train and necessitating careful consideration of computational resources and training strategies.
Potential Mitigation: Leverage techniques like knowledge distillation, where a larger model transfers knowledge to a smaller model, reducing computational requirements while preserving performance.
Alternative Approach: Explore efficient transformer variants or alternative architectures, such as convolution-augmented transformers or sparse transformers, that can achieve competitive performance with reduced complexity.

V. Conclusion

In this project, we comprehensively reviewed multimodal human fall detection systems. By critically analyzing existing literature, we identified best practices, potential improvements, and key findings. Our analysis underscored the effectiveness of integrating multiple data sources (vision, wearables, ambient sensors) for robust fall detection. Amongst the explored techniques, the GCN+Transformer model emerged as a leader, demonstrating superior accuracy and resilience to challenging scenarios. We also identified limitations, such as potential sensitivity to occlusions. While the GCN+Transformer presents a promising solution for accurate fall detection, particularly in privacy-conscious settings, its dependence on human pose estimation and training data diversity requires further exploration.

VI. Future Work

The combination of Graph Convolutional Networks (GCNs) and Transformer architectures demonstrates the potential for fall detection tasks; however, additional investigations and advancements are necessary to fully leverage this approach.

A. Enhancing Resilience of the Model to Occlusions and Complex Scenarios through:

- Exploration of advanced data augmentation techniques
- Incorporation of additional data modalities

B. Explore fall detection methods that bypass pose estimation for real-time applications through:

- Investigate methods to reduce the model's reliance on human pose key points
- Eliminate the additional preprocessing step required for real-time inference
- Enable seamless integration into practical applications

C. Expanding Dataset Diversity and Generalization through:

- Capture data in real-world environments (indoor, outdoor, varying lighting)
- Diversify participant demographics (age, gender, body type, abilities)
- Incorporate diverse fall scenarios (contexts, positions, intensities)
- Include a wide range of non-fall activities of daily living
- Utilize multiple data sources (video, wearable sensors, ambient sensors)
- Thoroughly annotate data (labels, contextual information)
- Expand dataset size and continuously update with new data
- Ensure privacy and ethical considerations (consent, anonymization)
- Validate and test the dataset (splits, cross-validation)
- Include a wider range of fall and non-fall scenarios in the training dataset
- Improve the model's generalization capabilities across diverse scenarios

D. Exploring Federated Learning

Investigate the feasibility of a federated learning approach to enable collaborative training of a global model across multiple clients and preserve the privacy of user data on individual devices.

E. Incorporating Additional Modalities

Enhance fall detection accuracy and robustness through the exploration of the integration of environmental sensors and other data sources.

F. Real-World Evaluation and Field Trials

Conduct field trials and real-world evaluations to assess practical viability and identify potential challenges. Refine the system based on real-world performance.

VII. Individual Contributions

Morteza Mogharrab:

- Re-implemented the proposed GCN + Transformer model architectureand Designed the overall framework
- Incorporated the Graph Convolutional Network (GCN) layer for spatial encoding
- Integrated the Transformer architecture for temporal encoding Spearheaded data fusion and multimodal integration efforts
- Ensured an effective combination of data sources (video cameras, wearable sensors, ambient sensors)
- Led the writing and documentation of the research Documented the methodology, implementation details, and analytical discussions

Ritika:

- Focused on evaluation and analysis aspects

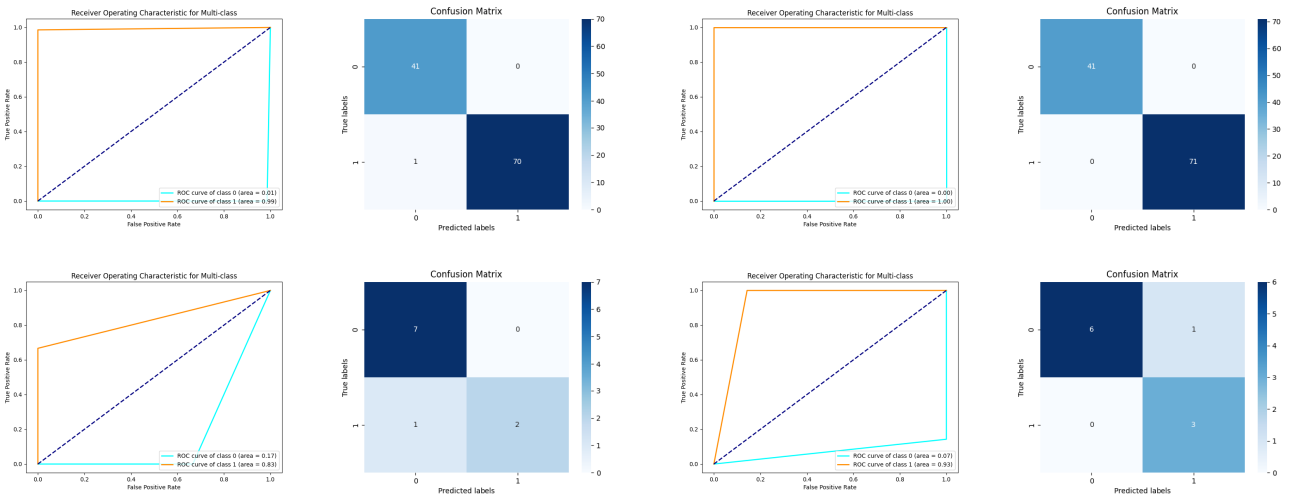

Fig. 7: Performance comparison on NTU (top row) and UR (bottom row) datasets. The left column shows ROC curves and confusion matrices for the Transformer-only architecture. The right column presents the results for the GCN+Transformer architecture.

- Conducted extensive experiments to assess the performance of the proposed model and baseline approaches across multiple datasets
- Provided insights into strengths, weaknesses, and generalization capabilities of different models
- Played a crucial role in dataset preparation Ensured proper formatting, preprocessing, and splitting of data into training, validation, and testing sets
- Contributed to data fusion and multimodal integration
- Collaborated with Morteza to effectively combine data from various sources

Sai Sarath Movva:

- Conducted an extensive literature review on fall detection systems
- Identified common themes, contradictions, and gaps in the current state of knowledge
- Addressed privacy and ethical considerations
- Ensured compliance with relevant regulations and best practices
- Responsible for the presentation and dissemination of research findings
- Prepared presentation materials

REFERENCES

[1] Kim, I., Kim, D., Kwon, S., Lee, S., & Lee, J. (2022). Fall Detection using Biometric Information Based on Multi-Horizon Forecasting. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR). IEEE. DOI: 10.1109/ICPR56361.2022.9956568

[2] Qi, P., Chiaro, D., & Piccialli, F. (2023). FL-FD: Federated learning-based fall detection with multimodal data fusion. Journal of Sensors and Actuators A: Physical. Advanced online publication.

[3] Taufeequea, M., Koitaa, S., Spicherb, N., & Deserno, T. M. (2021). Multi-camera, multi-person, and real-time fall detection using long short-term memory. In T. M. Deserno & B. J. Park (Eds.), Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications (Vol. 11601, p. 1160109). SPIE. DOI: 10.1117/12.2580700

[4] Balram, A., Puthanveettil, T., Singh, A., & Hundekari, K. (Year). Striking the Balance: Human Pose Estimation based Optimal Fall Recognition.

[5] Ha, T. V., Nguyen, H., Huynh, S. T., Nguyen, T. T., & Nguyen, B. T. (Year). Fall detection using multimodal data. arXiv, arXiv:2205.05918 [cs.CV]. Retrieved from https://doi.org/10.48550/arXiv.2205.05918

[6] Chi, T.-H., Liu, K.-C., Hsieh, C.-Y., Tsao, Y., & Chan, C.-T. (Year). PreFallKD: Pre-Impact Fall Detection via CNN-ViT Knowledge Distillation. arXiv, arXiv:2303.03634 [eess.SP]. Retrieved from https://doi.org/10.48550/arXiv.2303.03634

[7] S. Deandrea, E. Lucenteforte, F. Bravi, R. Foschi, C. La Vecchia, and E. Negri, "Risk factors for falls in community-dwelling older people:" A systematic review and meta-analysis"," Epidemiology, pp. 658–668, 2010.

[8] Centers for Disease Control and Prevention, "Expected num of elderly, injuries and falls," 2020, https://www.cdc.gov/steadi/pdf/STEADI ClinicianFactSheet-a.pdf, Accessed: 2021-05-14.

[9] Bogdan Kwolek and Michal Kepski. Human fall detection on an embedded platform using depth maps and wireless accelerometer. Computer Methods and Programs in Biomedicine, 117(3):489–501, 2014.

[10] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. Neurocomputing, 2023.

[11] 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, 146: 105626, 2022.

[12] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. Neurocomputing, 2023.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. CoRR, abs/2010.11929, 2020.

[14] Haoran Wei and Nasser Kehtarnavaz. Simultaneous utilization of inertial and video sensing for action detection and recognition in continuous action streams. IEEE Sensors Journal, 20(11):6055–6063, 2020.

[15] Qipeng Zhang, Tian Wang, Mengyi Zhang, Kexin Liu, Peng Shi, and Hichem Snoussi. Spatial-temporal transformer for skeleton-based action recognition. In 2021 China Automation Congress (CAC), pages 7029–7034, 2021.

[16] 7] H. Han, C. Lian, Z. Zeng, B. Xu, J. Zang, C. Xue, Multimodal multi-instance learning for long-term ECG classification, Knowl.-Based Syst. (2023) 110555.