Spatio-Temporal Action Recognition using YOLO-Pose and ST-GCN for Elderly Care and Rehabilitation

By: Fatimah Saadah, Hao Yang Hsu, Piangor Wijitbunditkun, Tzu Chien Tsai

Professor: Shyh-Hau Wang

Date: June 12th, 2026

Outline

- **Introduction**
- Methods and Materials
- Results and Conclusion

Introduction

- Background: Comparison of Traditional Clinical Gait Assessment and Vision-Based Gait Analysis

- **Traditional clinical gait assessment** in elderly care and rehabilitation relies on in-person observation and wearable or lab-based sensors, which can be accurate but are often costly, time-consuming, and uncomfortable for long-term use.
- **Vision-Based Gait Analysis** detection offer a non-contact and highly scalable, low-cost alternative, but is Sensitive to lighting and video quality, Privacy concerns, Dataset limitations, and Performance may vary across different subjects and conditions.

Citation: A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work, by Dimple Sethi, Sourabh Bharti, Chandra Prakash, https://doi.org/10.1016/j.artmed.2022.102314.

Introduction
- Literature Review

Core Framework	Actions	Pros	Cons
YOLOv8 + HRNet/SimCC + STFGCL (Ying et al.)	Walking, standing, turning	High-accuracy, non-invasive gait tracking	Two-stage latency; lacks environmental variation
DenseNet-201 + Parallel Fusion + SVM (Mehmood et al.)	Gait recognition	Robust to perspective, clothing, & baggage changes	Heavy parallel fusion causes high compute overhead; fails on small data
YOLOv8 + CNN-SNN Hybrid + cGAN (Slimi et al.)	Pathological gait (KOA vs. Parkinson's)	Low computation cost s & high energy efficiency	19× inference latency spike on standard hardware

Citation: https://www.mdpi.com/2079-9292/14/7/1437, https://link.springer.com/article/10.1007/s11042-020-08928-0, https://www.mdpi.com/3578712

Introduction
- Motivation & Objectives

- **Motivation**

1. Pipeline Latency: Existing frameworks rely on heavy, multi-stage models, such as Yolov8 and SimCC for human object and keypoints extraction, which can cause severe bottlenecks in real-time screening .

2. Data Hunger: Recent high-accuracy models use complex CNN parallel feature fusion, which completely fails when clinical or training datasets are small.

3. Hardware Bottlenecks: Advanced hybrid networks (like SNNs) reduce mathematical operations but suffer up to a 19× latency penalty on standard edge hardware.

- **Objective**

1. Efficient framework for real-time scenario: Build an ultra-fast, non-invasive gait analysis framework using YOLOv11-pose to extract human objects and keypoints simultaneously in a single step

2. Action recognition for rehabilitation and fall detection: Deploy ST-GCN (Spatial-Temporal Graph Convolutional Network) to map skeletal data efficiently, ensuring high accuracy without massive dataset requirements.

3. Clinical Application: Achieve real-time inference speeds optimized for edge deployment in rehabilitation tracking and elderly fall detection.

Outline

- Introduction
- **Methods and Materials**
- Results and Conclusion

Methods and Materials

- Flowchart

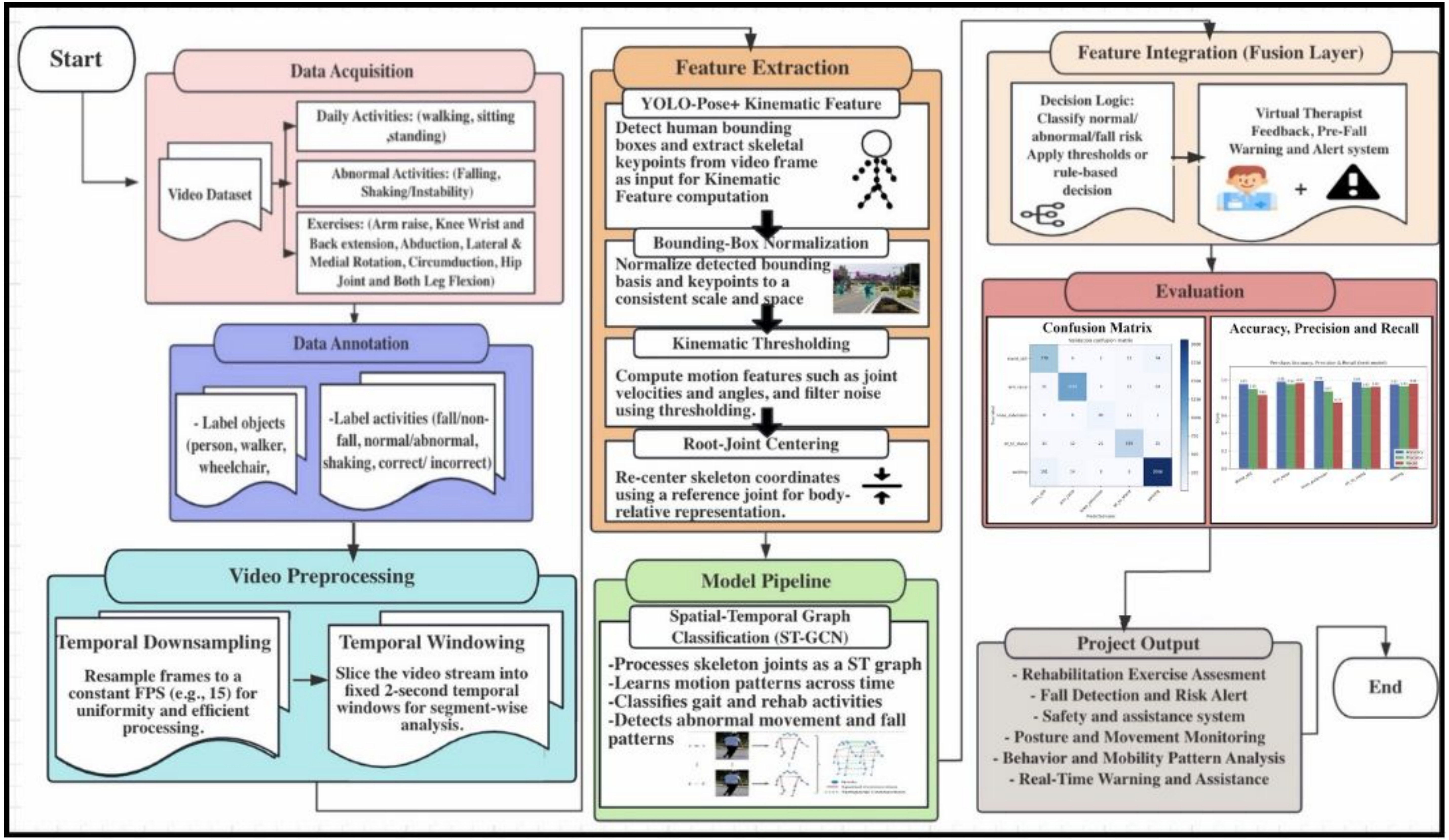


Fig 1. Flowchart.

Methods and Materials

- Dataset

Dataset	Volume & Timeline	Location & System	Actions
WLU Rehabilitation Posture - Kaggle (Amaan Javed et al.)	338 videos taken in 09 Sep, 2024	WLU Wilfrid Lauier University, Canada Multi-angle video recording system	Arm raise, knee extension, sit to stand all actions include correct and incorrect version
Havard Dataverse MobiPhysio (M. T. B. Iqbal et al.)	3688(63 used) taken in 20 July, 2025	University of Asia Pacific, Bangladesh 2D video-based recording system	Abduction, adduction, circumduction
Fall Video Dataset - Kaggle (https://www.kaggle.com/datasets/payutch/fall-video-dataset/data)	6988 videos taken time is not explicit stated	University of Asia Pacific, Bangladesh video cameras, mobile phone, digital camera, etc.	Fall, non-falling
Human Activity Recognition - Kaggle (Sharjeel M. Rajput)	667 videos taken in 2023	Capital University of Science & technology, Pakistan phone camera with a tripod	standing still, walking

Citation: Amaan Javed et al., doi:10.34740/KAGGLE/DSV/10250073, https://doi.org/10.34740/kaggle/dsv/5722068

Iqbal et al., “MobiPhysio,” Harvard Dataverse, 2025, doi:10.7910/DVN/XSI0QN. https://www.kaggle.com/datasets/payutch/fall-video-dataset/data, https://doi.org/10.6084/m9.figshare.26282872.v2

Methods and Materials
- Data Augmentation/Annotation

- **Data augmentation** is applied to smaller dataset to enhance data balance. **Gaussian noise** and **horizontal flip** are utilized for data augmentation
- **Data annotation** is used to labeled each video using the **start frame** and **end frame** to capture the activity duration accurately.

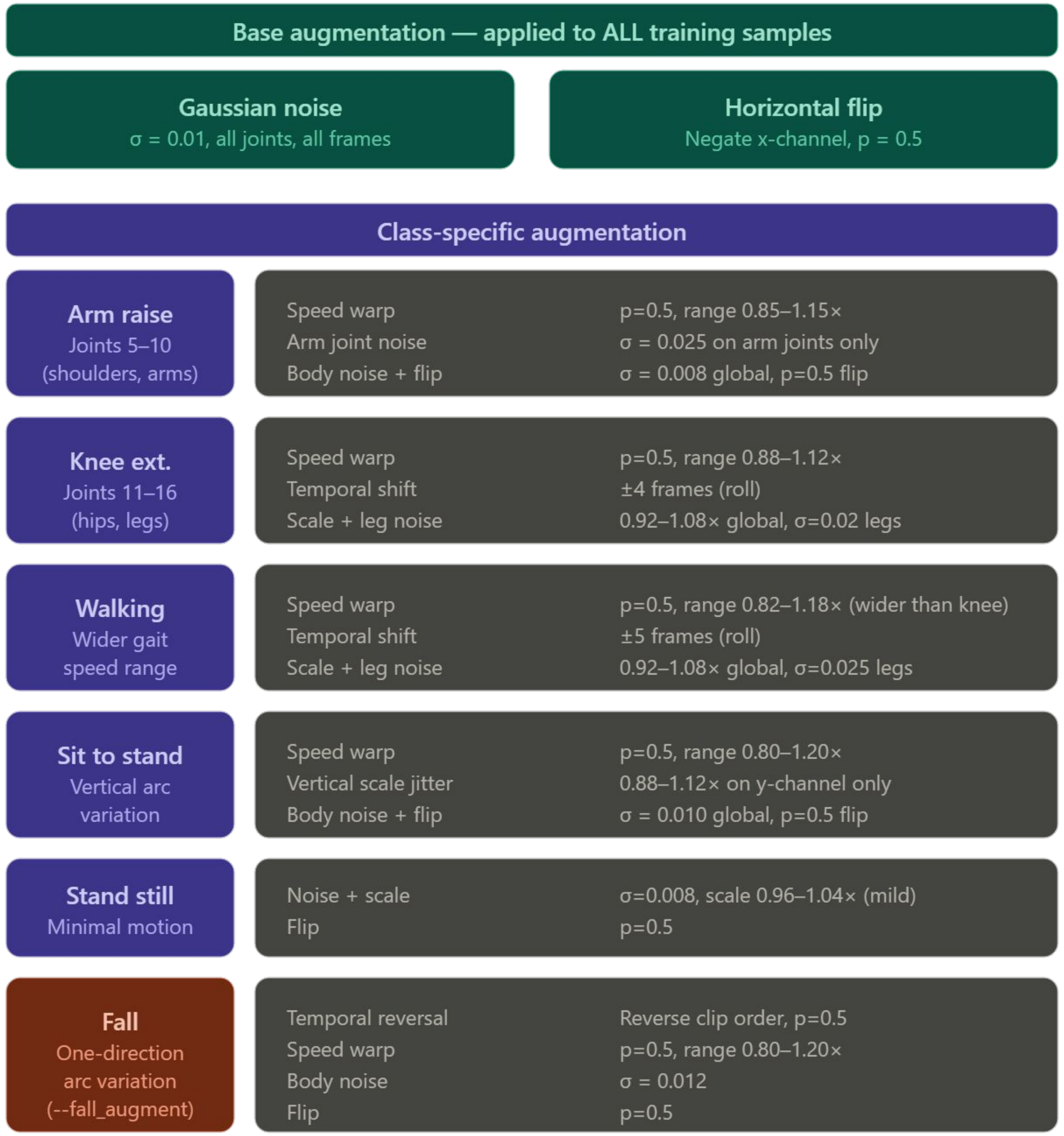


Fig 2. Steps of data augmentation.

Methods and Materials

- VIdeo Preprocessing

- **Temporal Downsampling**: Standardizes frame rates across diverse video sources (e.g., 30/60 FPS) to a **30/15/10 FPS** by skipping unnecessary frames.
- **Temporal Windowing**: Divides continuous video stream into **2-second windows (30 frames)** with a **50% temporal overlap (stride)** and captures the dynamic evolution of an action.

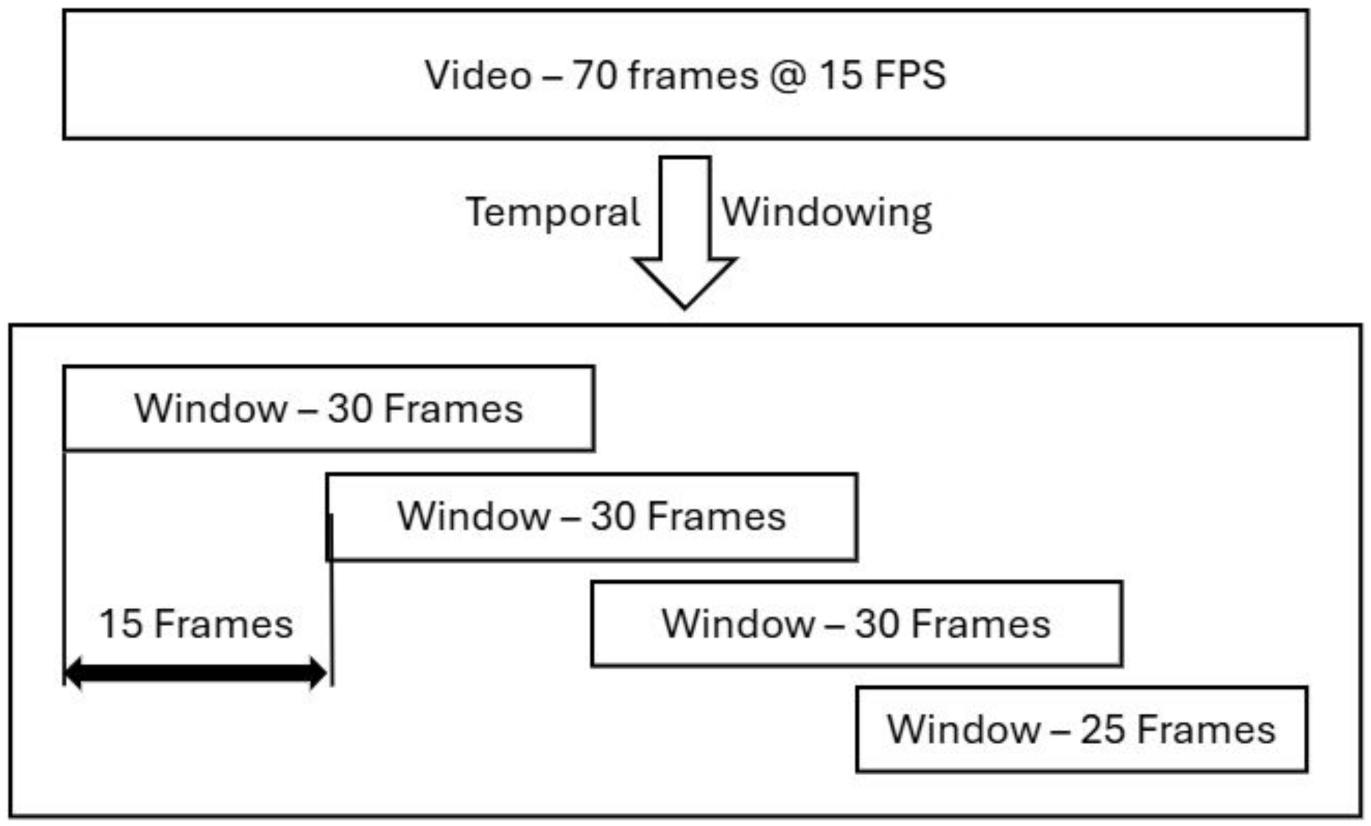


Fig 3. Example of Temporal Windowing slices a video with 70 frames into multiple windows.

Methods and Materials

- Feature Extraction

Kinematic Feature Extraction(Pose Estimation)

- YOLOv11-pose for human pose estimation
- Extracts 17 COCO-format keypoints per frame
- Generate bounding box: (x, y, w, h)
- Frame-wise skeletal representation

Spatial Bounding-Box Normalization

- Coordinate normalization based on bounding box (0.0 – 1.0 scale)
- Reduces scale and distance variation
- Improves model robustness

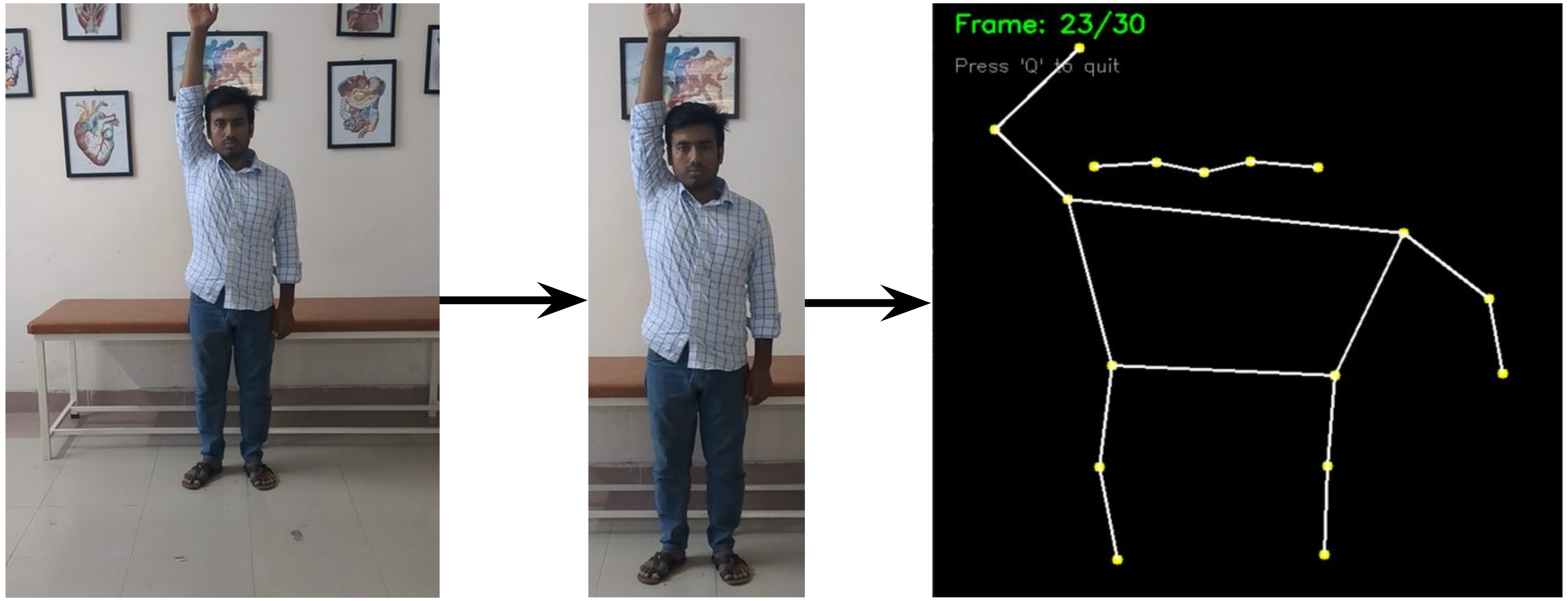


Fig 4. Steps of transforming the image to skeletal keypoints.

Methods and Materials

- Feature Extraction

- **Kinematic Thresholding**: Calculates the spatial variance of joints within each window and discards "Idle" windows (variance < threshold).
- **Root-Joint Centering**: Recalculates all joints relative to the hip center (Root Joint) by setting the coordinates between the of the hip keyoints as (0,0), then update all the keypoints.

$$V_w = \frac{1}{V \cdot T} \sum_{v=1}^{V} \sum_{t=1}^{T} \|x_{v,t} - \bar{x}_v\|^2$$

V_w: The spatial variance of joints

V: The number of joints

T: The number of frames in current window

$x_{v,t}$: Coordinates of joint v at time t

$\bar{x}_v$: Mean spatial position of joint v over window T

Fig 5. Formula used for kinematic thresholding. If V_w exceed the threshold(0.001), the the person is determined not moving.

Methods and Materials

- STGCN Flowchart

Spatial Temporal Graph Convolutional Networks (ST-GCN)

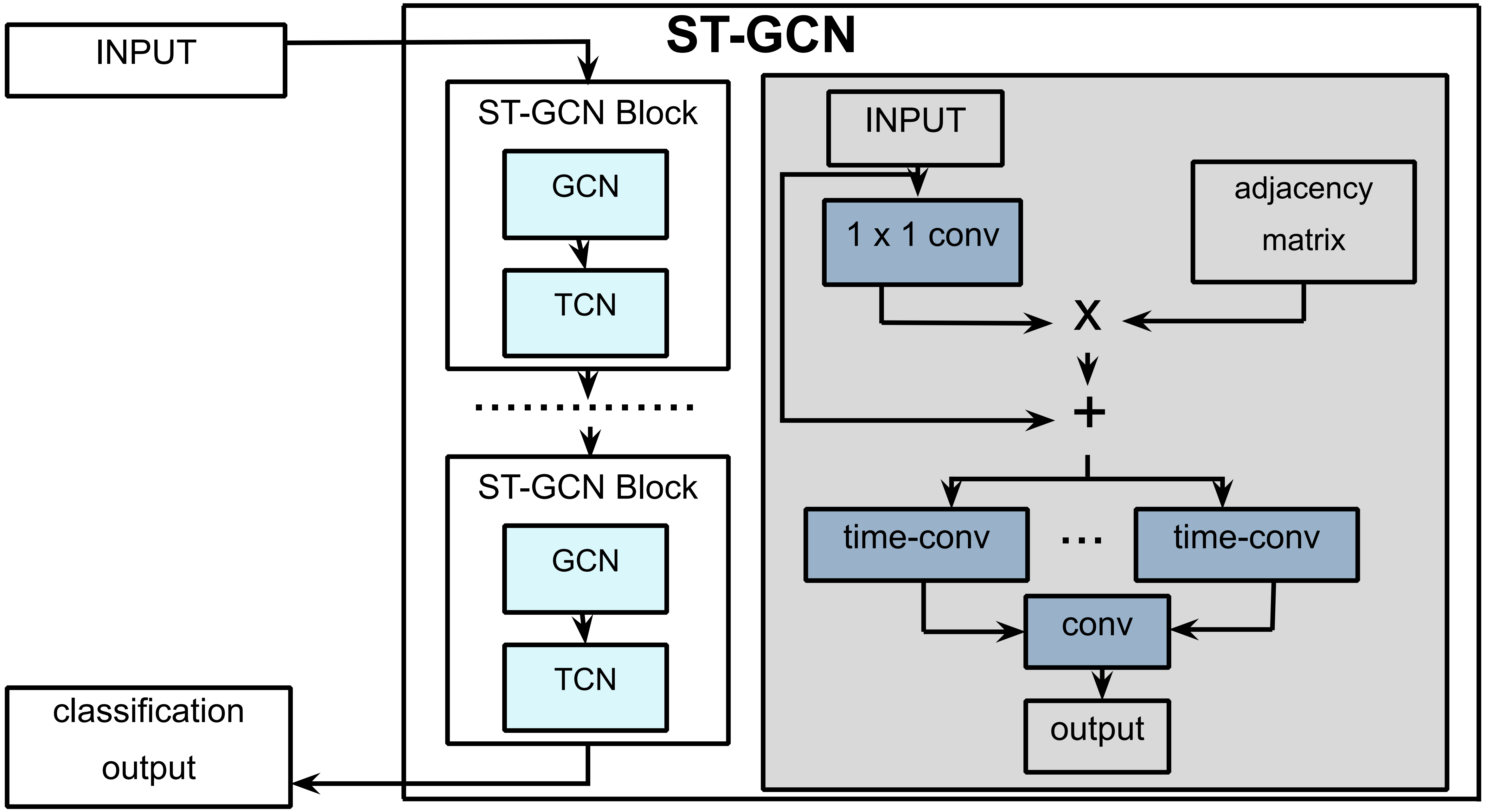


Fig 6. Flowchart of ST-GCN

Citation: https://www.researchgate.net/figure/Multi-scale-time-window-transformation_fig5_390435658

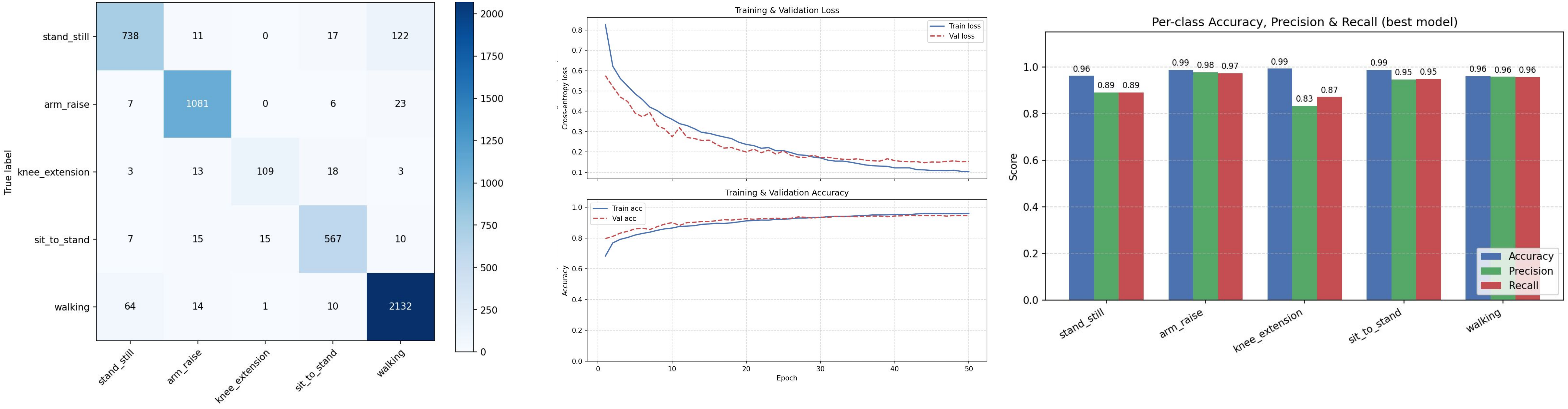
Outline

- Introduction
- Methods and Materials
- **Results and Conclusion**

Results and Conclusion

- Model Performance

ST-GCN achieved **92.7% accuracy** under the condition of **10 FPS, 40 frames per window and 20 frame stride**. The loss function and the accuracy converges steadily.



(a) Confusion matrix of our model.

(b) Training Accuracy & Loss Trends

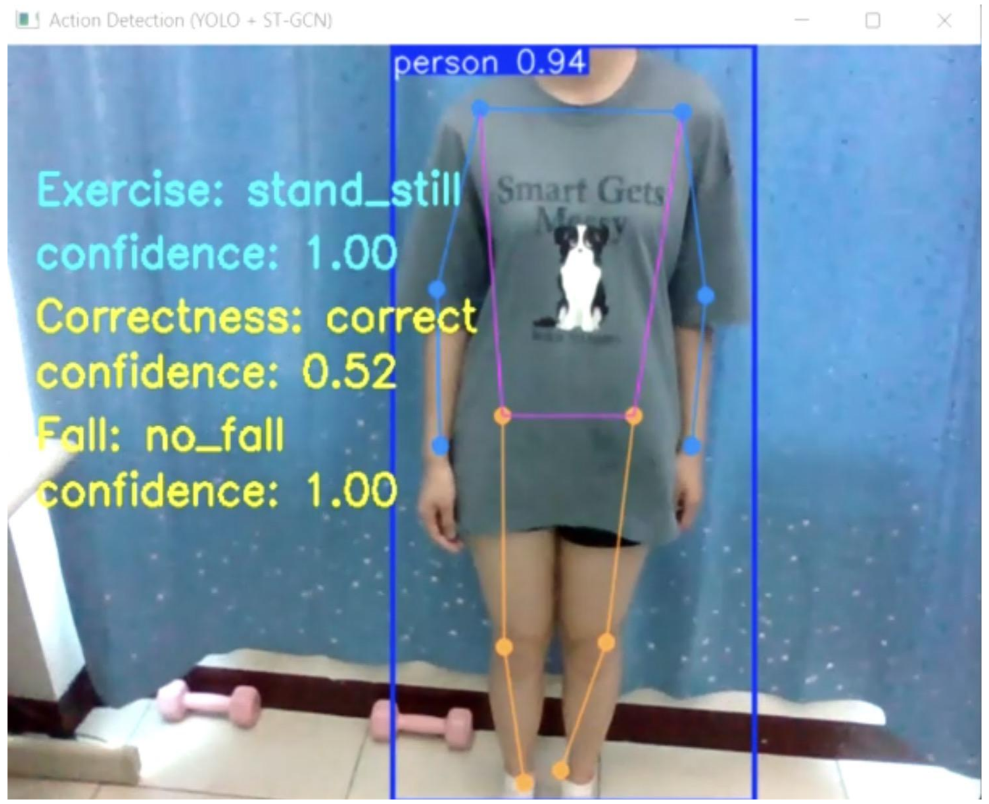
(c) Per-class accuracy, precision and recall

Fig 7. Results of the model's performance

Results and Conclusion

- Real-time Detection

- A **real-time action sensing system** is developed for feature integration.
- The system extracts **sequences of human keypoints** using YOLO-pose, and ST-GCN will detect its action to determine the exercise and the risk of falling
- Integrated the LINE Messaging API , automated alert system that sends real-time notifications to caregivers upon fall detection.



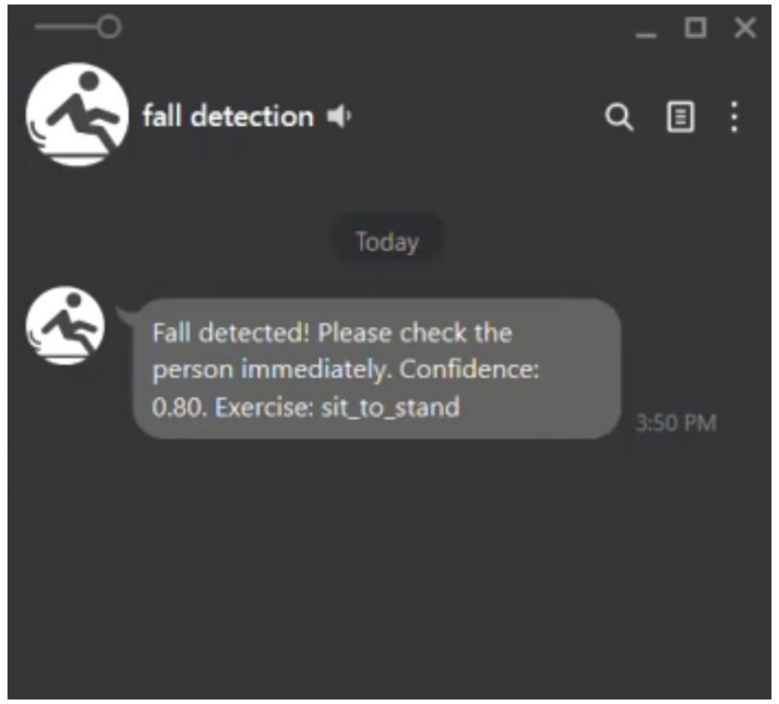


Fig 8. real-time detection camera and line bot alert for fall detection (Change)

Results and Conclusion

- Discussion & Conclusion

- By using YOLOv11-pose and ST-GCN, we are able to extract the keypoints at **high efficiency**, while maintaining **good accuracy**
- Future works would revolve around using other model for action detection to increause its accuracy while maintaining the efficiency, while optimizing user interface to allow better action correction for exercising, making it a real virtual medical assistance.

Thank you for your attention

(Fatimah Saadah, Hao Yang Hsu, Piangor Wijitbunditkun, Tzu Chien Tsai)