# Vision-Based Human Activity Recognition using Transformer Architectures


**Advisor: Patrick Denny**

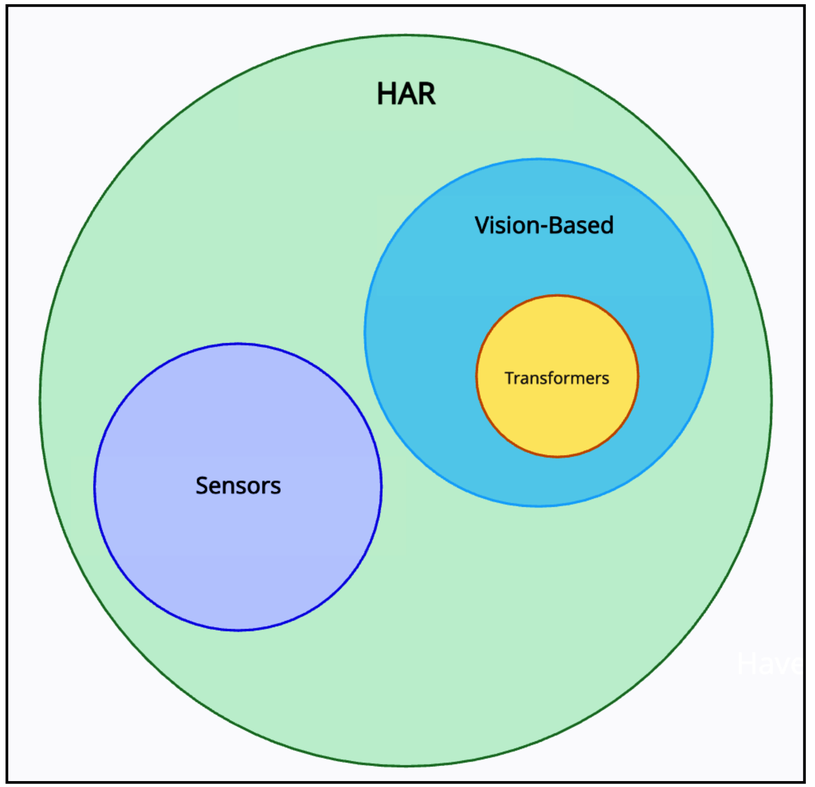
Author: Siddartha Sandeep Peddada

25 August 2025

# Contents

# Introduction & Motivation

- Human Activity Recognition (HAR) → healthcare, fitness, surveillance.
- **Sensors**: robust, low-cost, but lack context.
- **Video**: rich detail, but sensitive to the environment.
- Challenge → domain shift & generalization.



# Problem Statement & Objectives

➤ Problem:

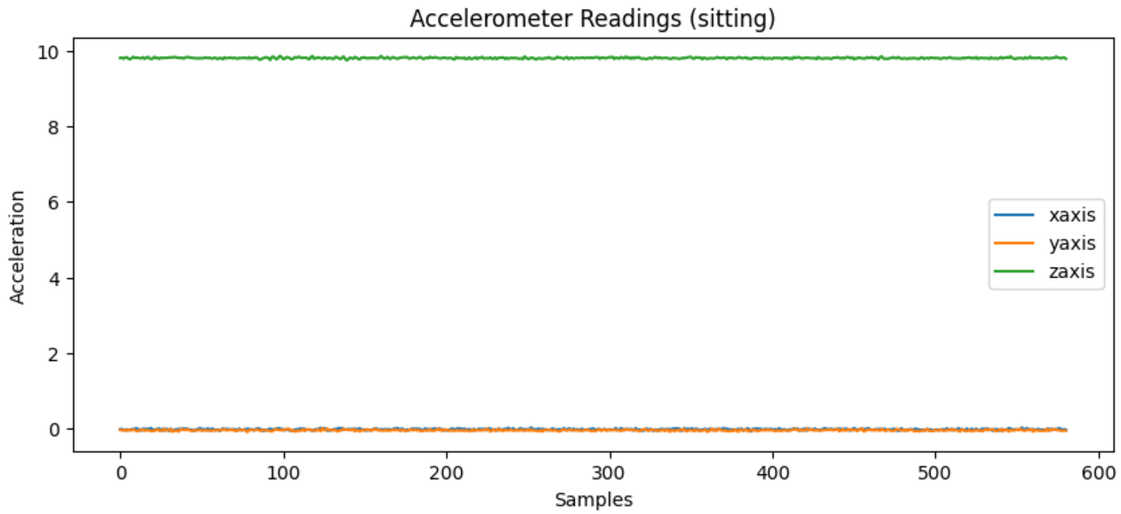
- Deep models → high accuracy in-domain but fail under domain shift.
- Sensors → robust, but miss posture/context

➤ Objectives:

- Train Vision Transformer (ViT) on video
- Train Random Forest (RF) on accelerometer signals
- Fuse predictions with weighted averaging
- Apply explainability: Grad-CAM, Attention Rollout, LRP

# Datasets

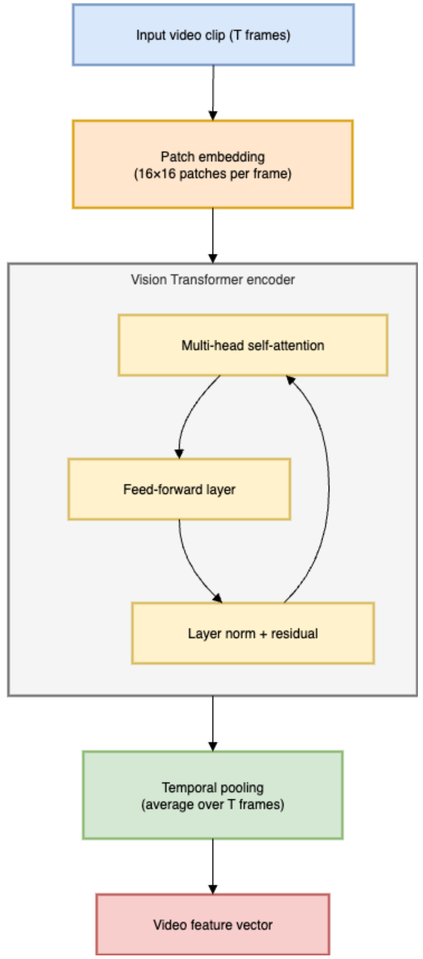
- Video Dataset (Kaggle HAR) [1]
  - 7 activities → used Sitting, Standing, Walking.
- MMAct Dataset (Video + Accelerometer) [2]
  - 20 subjects → 18–20 held out for unseen testing.
- Both datasets aligned on 3 activities for fair multimodal fusion.

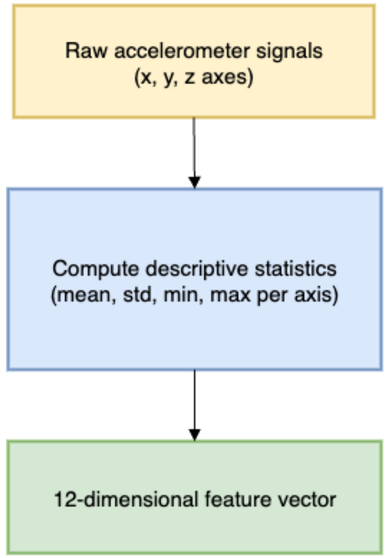



# Datasets


- Video Dataset (Kaggle HAR)
  - 7 activities → used Sitting, Standing, Walking.
- MMAct Dataset (Video + Accelerometer)
  - 20 subjects → 18–20 held out for unseen testing.
- Both datasets were aligned on 3 activities for fair multimodal fusion.

# Models: Vision Transformer & Random Forest




- Vision Transformer (ViT):
  - Input: 8 frames → patch embeddings → transformer encoder.
  - Backbone: vit_base_patch16_224.
  - Class-weighted loss for imbalance.
  - Learns posture & motion cues.
- Random Forest (RF):
  - Input: 12D statistical features (mean, std, min, max per axis X, Y, Z).
  - 100 trees, balanced class weights.
  - Lightweight & interpretable.
  - Captures motion dynamics.

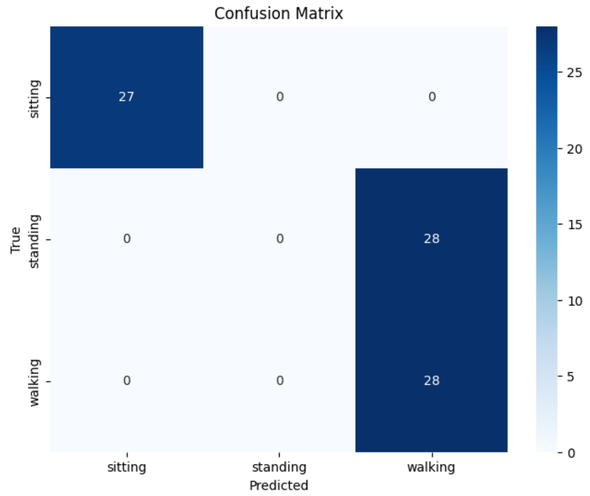
# Results: Vision Transformer (Video Model)

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

```
Classification Report:
              precision    recall  f1-score   support

     sitting       0.97      0.97      0.97        40
    standing       0.97      0.97      0.97        35
     walking       1.00      1.00      1.00        97

    accuracy                           0.99       172
   macro avg       0.98      0.98      0.98       172
weighted avg       0.99      0.99      0.99       172
```




- In-domain (Video dataset):
  - Accuracy: **99%**
  - Almost perfect Sitting/Standing/Walking recognition.
  - Class-weighted loss for imbalance.
  - Learns posture & motion cues.
- Unseen MMAct subjects (18–20):
  - Accuracy drops to **66%**
  - Standing misclassified entirely.
- Strong posture recognition but poor domain generalization.

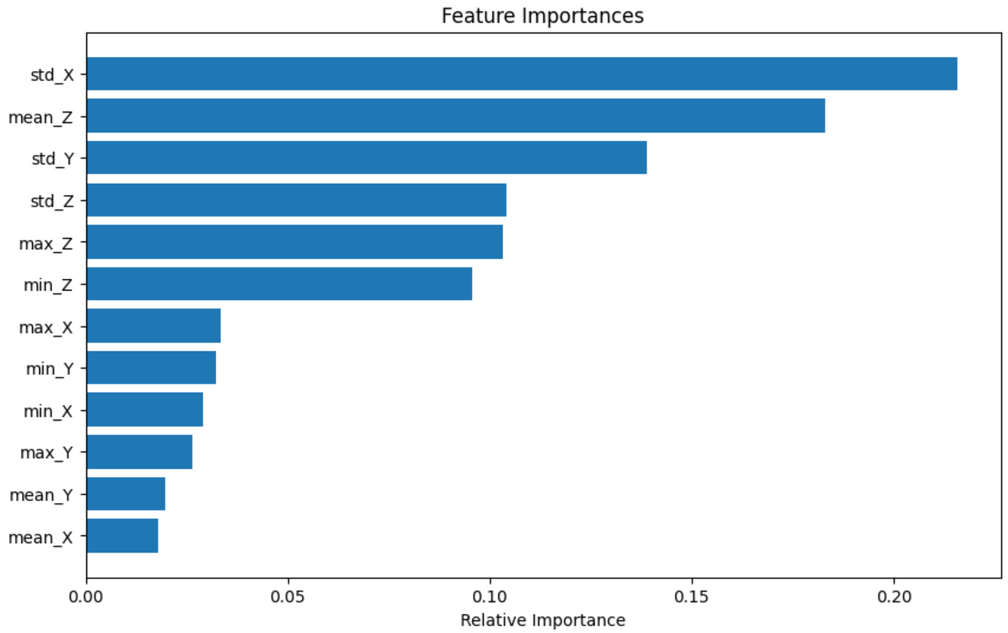
# Results: Random Forest (Accelerometer Model)

- In-domain (Accelerometer dataset):
  - Accuracy: 98.96%
  - Stable across cross-validation folds (~95%).
- Unseen MMAct subjects (18–20):
  - Accuracy: 99%
  - Strong generalization across users.
- Feature importance: Variance & max values most discriminative.

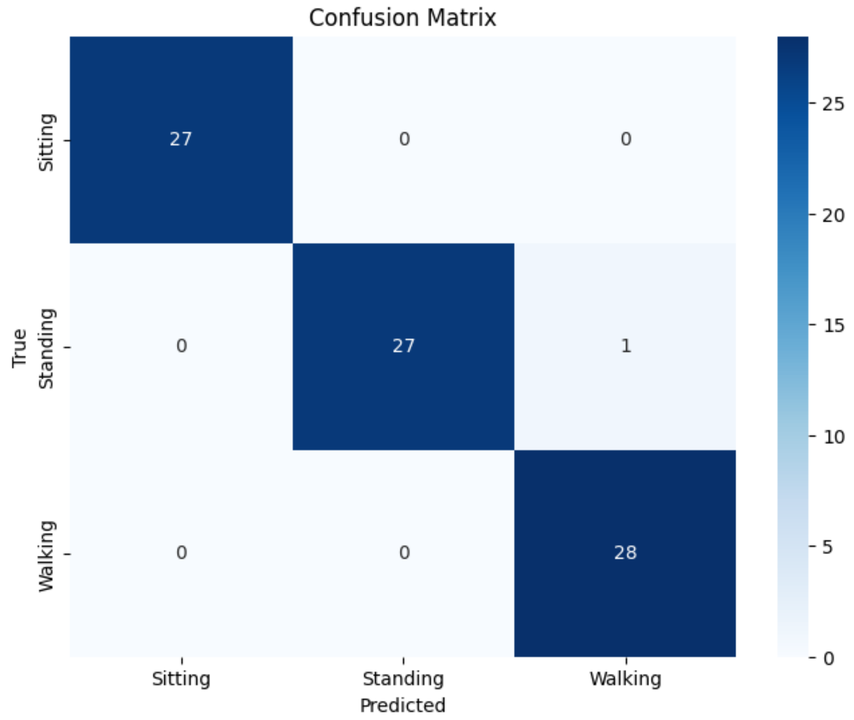
```
Accuracy: 0.9896
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.98        30
           1       1.00      1.00      1.00        36
           2       0.97      1.00      0.98        30

    accuracy                           0.99        96
   macro avg       0.99      0.99      0.99        96
weighted avg       0.99      0.99      0.99        96
```



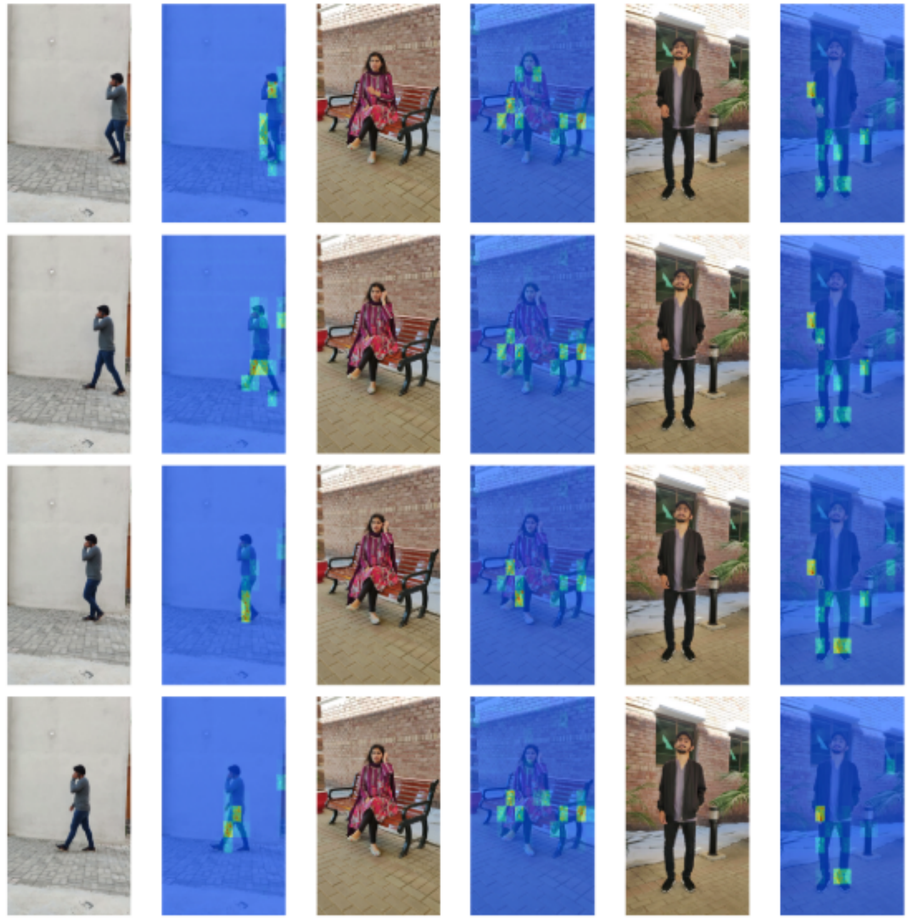
# Results: Fusion Model (RF + ViT)




- **Fusion strategy:** Probability-level weighted averaging (weights ∝ F1).
- Improvement:
  - Corrected Standing misclassifications from ViT.
  - Maintained high Sitting & Walking recognition.
- **Outcome:** More balanced recognition across all 3 activities.

# Explainability Results (ViT)

- Grad-CAM:
  - Walking → legs highlighted.
  - Sitting → torso + chair.
  - Standing → inconsistent focus.
- Attention Rollout:
  - Sharper, localized focus → especially legs for Walking/Standing.
- LRP:
  - Torso for Sitting, legs for Walking.
  - Weak & diffuse focus for Standing.



Grad-Cam

# Explainability Results (ViT)

Attention Rollout

LRP (A x G)

# Conclusion & Future Work

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

**Conclusion:**

- ViT → powerful, but brittle under domain shift.
- RF → robust, interpretable, but limited detail.
- Fusion → best balance across activities.
- Explainability → validated model decisions.

**Future Work:**

- Train on larger, more diverse datasets.
- Use advanced fusion (attention-based, joint embeddings).
- Add more modalities (gyroscope, depth, physiological signals).
- Optimize for real-time, lightweight deployment (TinyML).

# References

[1] Sharjeel M. Rajput, Muhammad Bilal, and Areesha Habib. (2023). Human Activity Recognition (HAR - Video Dataset) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/5722068

[2] Chen, S., He, H., & Wang, K. (2020). *MMAct: A large-scale dataset for cross-modal human action understanding*. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20) (pp. 4390–4398). https://doi.org/10.1145/3394171.3413793

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

# Thank you!

# Questionnaire?