Human Action Recognition Project (Group-2)

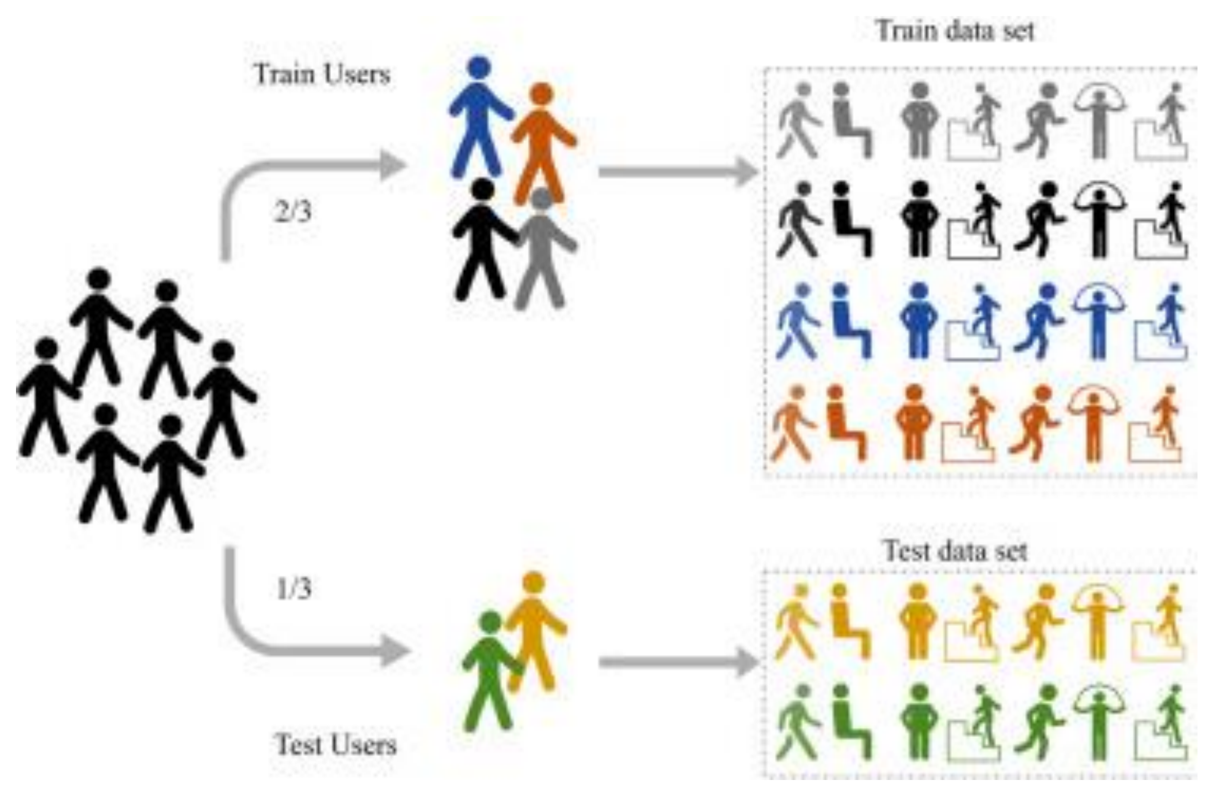
- Human Action Recognition using Deep Learning
- Group-2
- Course: DATS 6303 – Deep Learning
- Instructor: Dr. Amir Jafari

Motivation & Applications

- What is Human Action Recognition?
- Why recognition from **single RGB images** is difficult
- Key challenges (Pose ambiguity, background clutter, similarity across actions)

Applications

- Smart surveillance
- Assistive healthcare monitoring
- Sports analytics
- Human-Computer Interaction (HCI)



Dataset Overview

Dataset: Human Action Recognition (HAR)

• 12,600 RGB images with 15 everyday actions (balanced: 840 images per class)

• Single-frame images only → no motion / temporal information

• Real-world variation: indoor/outdoor scenes, cluttered backgrounds, different body types

Preprocessing & Splits

• Resized to 240×240 (and 299×299 for Inception-based models)

• Normalized using ImageNet mean & std

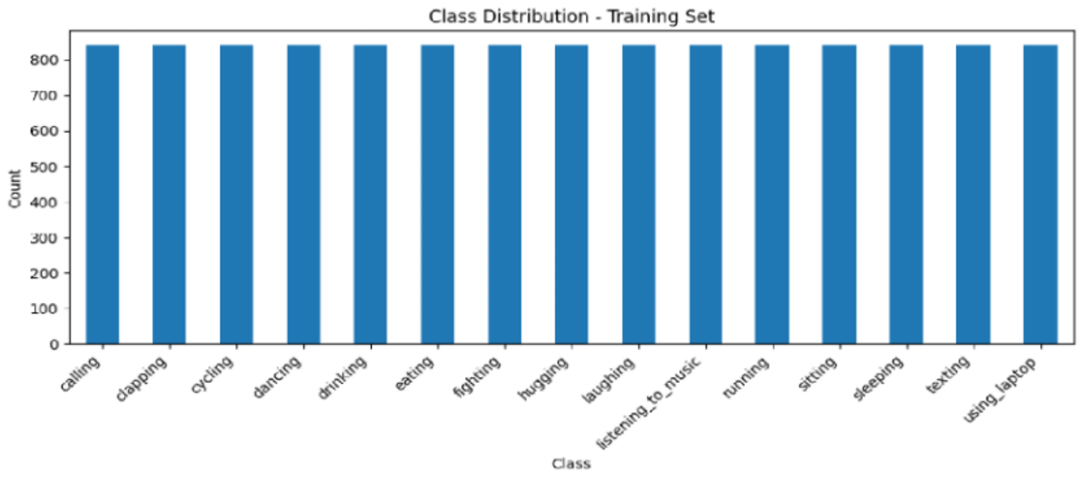
• Stratified split: 80% train, 10% validation, 10% test

EDA Highlights

• Clear separation for dynamic actions (cycling, running, sleeping)

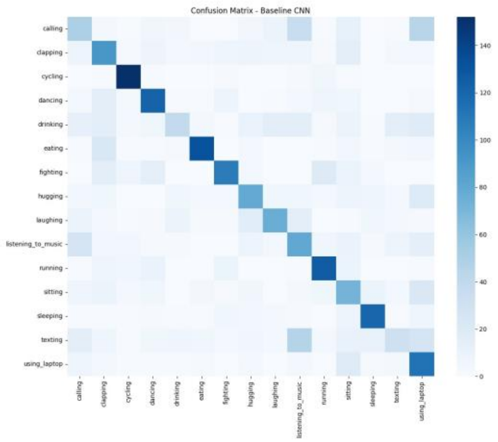
• Significant overlap for seated actions (texting, using laptop, listening to music)

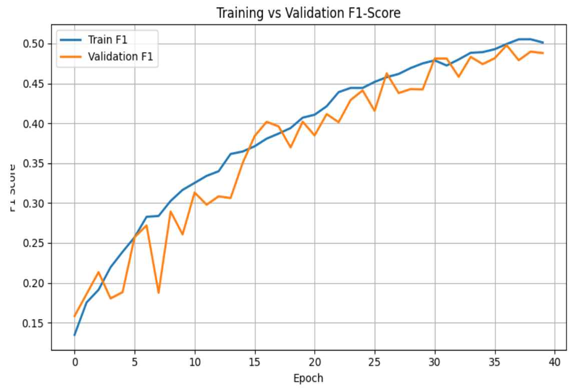
• Lighting and background variation make generalization more challenging

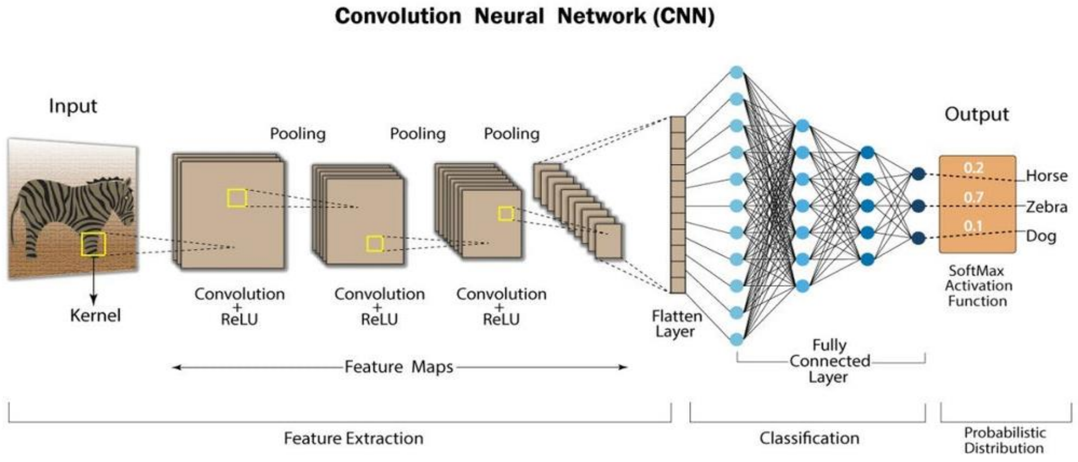


Architecture + Training Techniques- CNN Baseline

- The network uses 5 convolutional blocks — **Conv → BatchNorm → ReLU → MaxPool** — which help the model learn patterns step-by-step:
  edges → limb shapes → full body postures
- After feature extraction, Adaptive Average Pooling compresses spatial info, and then 2 dense layers — with dropout — classify all 15 actions.







Results & Insights

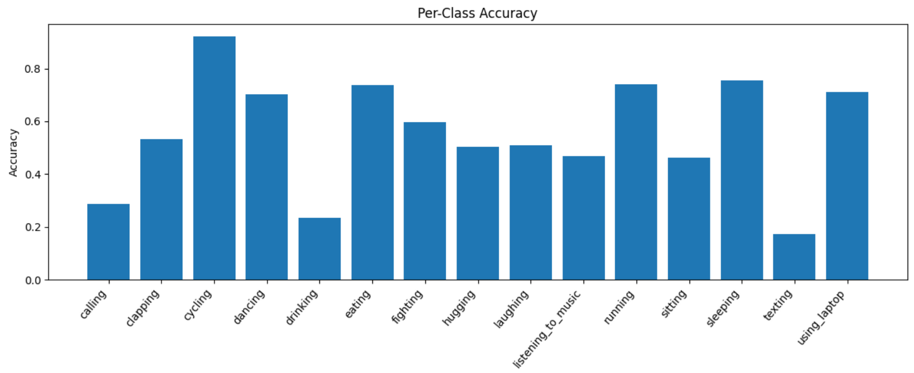
Strengths

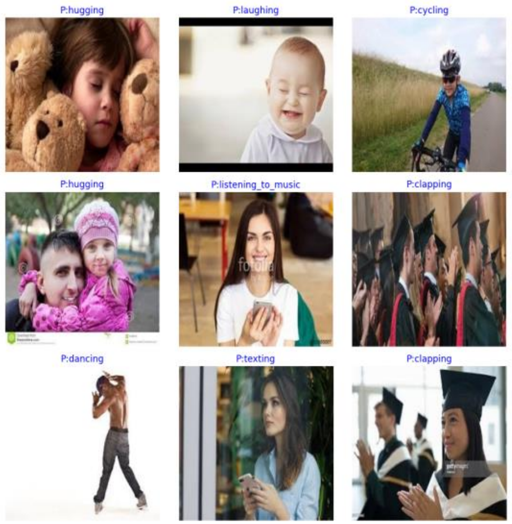
- Light architecture (1.2M params)
- Fast inference — deployable on edge devices

Challenges

- Lacks deep feature richness
- Struggles to differentiate fine hand-pose-based actions

Metric	Value
Accuracy	**54.50%**
Macro F1-Score	**54.00%**





EfficientNet-B1-Method & Improvements

Why EfficientNet-B1?

- Compound scaling → optimized depth, width & resolution
- High accuracy with low computational cost

Training Strategy

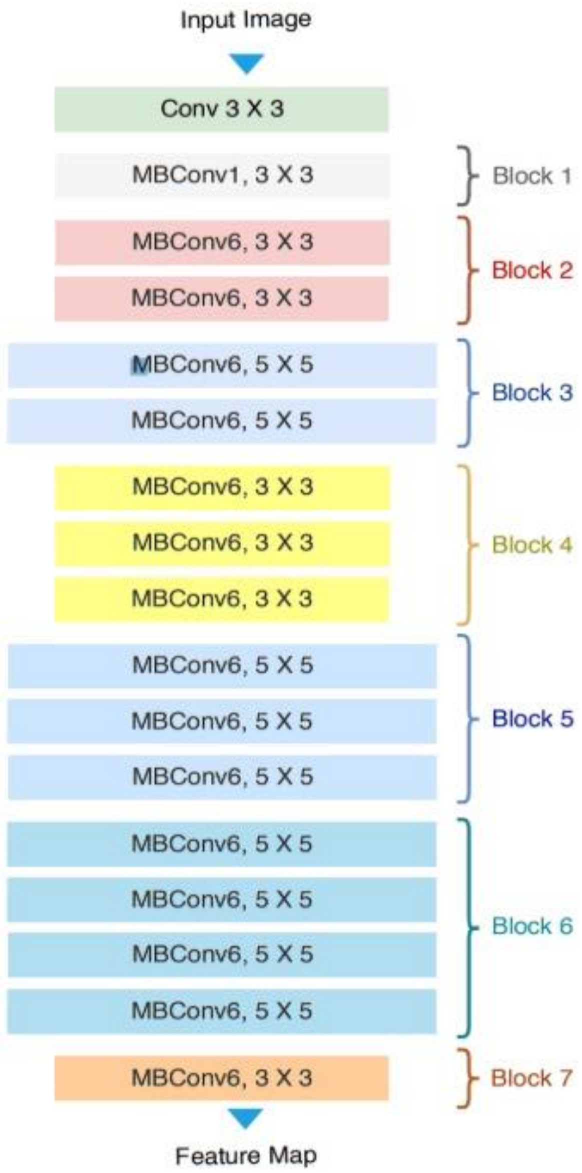
- Transfer learning with ImageNet weights

Gradual Unfreezing (Classifier → Full Network)

- Strong augmentation pipeline
- Label Smoothing for visually similar actions
- MixUp Regularization → reduces overfitting
- Adaptive learning rate (ReduceLROnPlateau)

Architectural Strengths

- MBConv blocks + Squeeze-and-Excitation attention
- Learns subtle posture cues & contextual clues (phones, laptops, bottles)



Results & Conclusions

- Strong generalization across all 15 classes

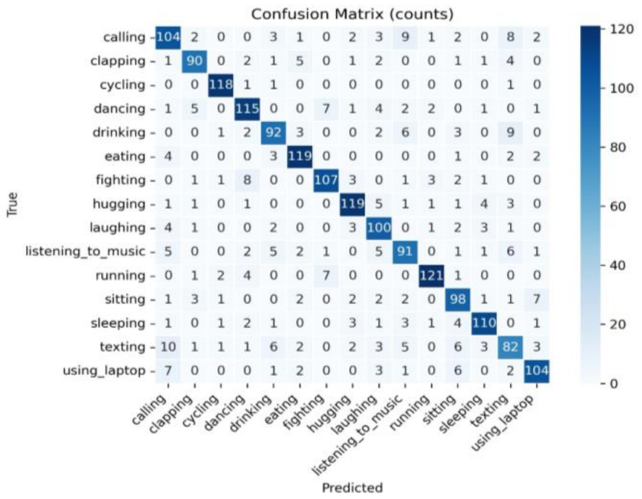
Per-Class Performance Highlights

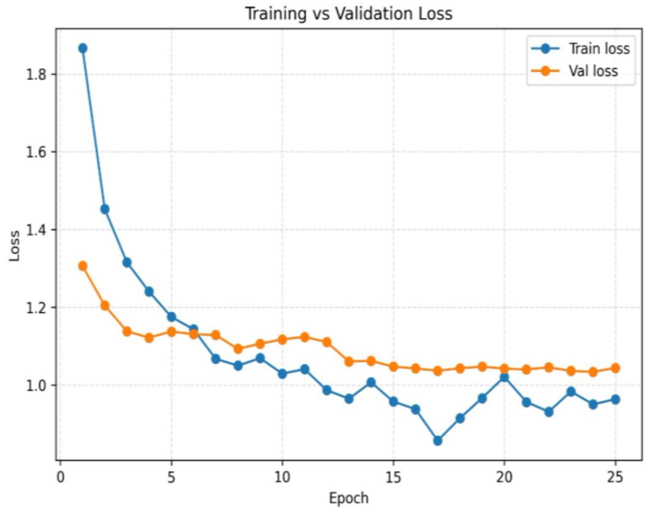
High-performing classes:

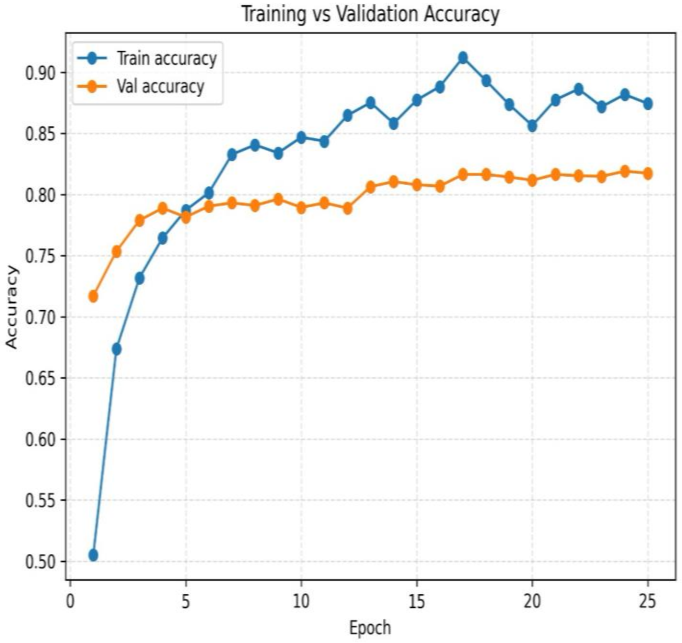
- **Cycling** → F1 ≈ **0.95**
- **hugging**→ F1 ≈ **0.87**
- **Eating** → F1 ≈ **0.89**
- **Running** → F1 ≈ **0.90**

Challenging classes (pose ambiguity):

- **Texting vs. Using Laptop**
- **Sitting vs. Listening_to_music**







Metric	Value
F1 Accuracy	**83.3%**
F1 Macro	**83%**
Macro Recall	**83.02%**
Precision Macro	**83.03%**
Top-3 Accuracy	**97.5%**

ResNet-50

Why ResNet-50 for HAR?

- Deeper networks typically extract richer pose + contextual features
- But training deep CNNs can cause vanishing gradients
- Residual skip connections solve this by enabling smoother gradient flow

Component	Role in HAR
Initial Convolution + MaxPool	Captures global body shape
4 Residual Stages (Conv2_x → Conv5_x)	Learn high-level posture + context
Identity & Projection Shortcuts	Prevent feature degeneration in deep layers
Global Average Pooling	Robust to pose shifts & spatial changes
Fully Connected Layer (15-class output)	Final action prediction

Results

- **Per-Class Observations**

High performing:

- **Cycling, Running, Sleeping** (distinct poses)

Challenging:

- **Texting vs Using laptop**
- **Sitting vs Listening_to_music**

Technical Strength

- Residual Skip Connections → Better gradient flow → Stable convergence

Metric	Value
Accuracy	83.21%
Macro Precision	83.40%
Macro Recall	83.21%
Macro F1-Score	83.10%

VGG-16

- Pretrained on ImageNet → helps recognize human structure and objects
- Uses stacked 3×3 convolutions → captures fine-grain pose features
- Strong baseline due to simplicity and stable feature extraction
- Resized to 224x224 --->RGB ---->Pytorch tensors
- Final fully-connected head modified to 15-class HAR
- Dropout applied to reduce overfitting risk

Best at actions with clear posture differences
(e.g., running vs. clapping, cycling vs. dancing)

VGG-16 Performance & Insights

- **Strengths**
  - Captures strong low-level spatial features
  - Good on clear and simple actions
- **Weakness**
  - Heavy model (138M params)
  - Overfits quickly without aggressive augmentation
  - Limited multi-scale perception

Metric	Value
Accuracy	**75.52%**
Macro Precision	**75.40%**
Macro Recall	**75.52%**
Macro F1-Score	**75.20%**

InceptionV3- Architecture Summary

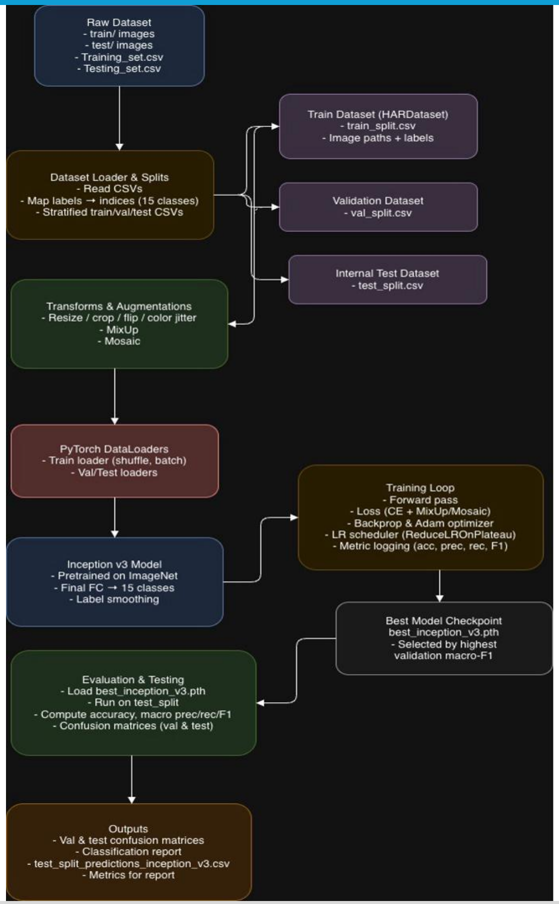
- Designed to capture multi-scale spatial features using parallel convolutional branches (1×1, 3×3, 5×5)
- Uses factorized convolutions to reduce computational cost while maintaining accuracy
- Pretrained on ImageNet → strong generalization for pose and context cues
- Modified final classification head → 15 action classes
- Includes label smoothing and dropout for better generalization
- Efficient on GPU memory → handles larger resolution inputs

Strengths

- Good at recognizing actions involving distinct object cues
- Multi-scale filters help detect fine + global pose features

Challenges Noted

- Confusion in very similar seated actions
  (e.g., sitting vs using laptop vs listening to music)



Performance

Strengths

- Multi-scale convolution → good for identifying **object cues** (phone, bottle)

Weakness

- Slight confusion in **seated actions** due to background similarity

Metric	Value
Accuracy	**82.62%**
Macro Precision	**82.70%**
Macro Recall	**82.30%**
Macro F1-Score	**82.66%**

Model Performance Comparison

Why EfficientNet-B1 Wins?

- Best generalization → less confusion in look-alike actions
- High accuracy without huge computational cost
- Works well with strong augmentations + label smoothing
- Suitable for deployment (mobile-friendly)

Key Insight

- More parameters ≠ better performance

Balanced scaling (EfficientNet) > deeper backbone (VGG/ResNet)

Model	Accuracy	Macro F1	Params	Performance Level
EfficientNet-B1	**83.2%**	**83.06%**	~7.8M	Very Strong
ResNet-50	**83.21%**	**83.10%**	~25.6M	Very Strong
InceptionV3	82.30%	82.10%	~23.9M	Strong
VGG-16	75.52%	75.20%	~138M	Good but heavy
Custom CNN	54.50%	54.00%	~1.2M	Baseline

Key Observations

- Pretrained deep models outperform the custom CNN significantly on HAR.
- Actions involving distinct body motion (e.g., *cycling, running*) achieve highest F1-scores.
- Fine-grained seated actions (e.g., *texting, using laptop, sitting*) remain the hardest to distinguish.
- Performance strongly correlates with model depth and feature extraction capability.
- Data augmentation + label smoothing notably improved generalization across all models.

Conclusion + Future Work

- Deep learning can effectively classify single-image human actions despite lack of motion cues.
- Among the tested models, **EfficientNet-B1** and **Resnet 50** offered the best balance of accuracy and computational efficiency.
- Future improvements should focus on **better pose encoding**, **attention modules**, and **temporal cues** where available.
- Our unified pipeline demonstrates a **scalable and practical solution** for real-world HAR systems.