

Lightweight Modality-Agnostic Model for Multimodal Human Activity Recognition

by

Sanzhar Abdulin

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

April 2026

© Nazarbayev University 2026. All rights reserved.

Author
Department of Data Science
2026/04/27

Certified by
Adnan Yazici
Chair, Professor
Thesis Supervisor

Accepted by
Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

Lightweight Modality-Agnostic Model for Multimodal Human Activity Recognition

by

Sanzhar Abdulin

Submitted to the Department of Data Science
on 2026/04/27, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

This paper presents Mamba-based lightweight and modality-agnostic framework for human activity recognition (HAR) that accommodates any combinations of input modalities, including RGB video, depth maps, skeleton joints, inertial and ambient sensors without requiring architectural modifications. The system employs modality-specific backbones each projecting into a shared embedding space. Evaluation was done on UTD-MHAD, HWU-USP, MSR-Action3D, and MM-Fit datasets using leave-one-subject-out cross-validation. Results show 93.8% and 93.1% test accuracy on Enhanced and Lightweight versions (UTD-MHAD). Modality performance is also strong at between 80% to 92%, depending on modality. Low parameter count and test inference times show that the proposed approach is suitable for edge devices. This claim is further supported by an actual testing on the Raspberry Pi 4 Model B.

Thesis Supervisor: Adnan Yazici
Title: Chair, Professor

Acknowledgments

I want to sincerely thank Professor Adnan Yazici, my thesis supervisor for his guidance and support throughout whole project. His domain expertise and advices helped me to refine and finish thesis work. Also, I am very grateful to him and Nazarbayev University for the provided research environment and necessary devices. Finally, this work will not be possible without support of my wife Aruay, my family and friends, and their constant support during my study. Their warming and encouragement words inspired me on the finish stage of this project.

Contents

1	Introduction	13
1.1	Motivation	14
2	Related works	17
2.1	Multimodal HAR	17
2.1.1	Multimodal HAR on edge devices	18
2.2	Modality-Agnostic models	19
3	Methodology	23
3.1	Datasets	23
3.1.1	UTD-MHAD	23
3.1.2	HWU-USP	24
3.1.3	MSR Action3D	26
3.1.4	MM-Fit	26
3.1.5	Testing strategy	27
3.2	Model pipeline	28
3.3	Data preprocessing	28
3.3.1	RGB data preprocessing	29
3.3.2	Skeleton Data preprocessing	31
3.3.3	Inertial data preprocessing	31
3.3.4	Ambient data preprocessing	31
3.4	Proposed Pipeline	33
3.4.1	Feature Selection	34

4	Results	41
4.1	Results on UTD-MHAD	41
4.2	Results on HWU-USP	48
4.3	Results on MSR Action 3D	51
4.4	Results on MM-Fit	53
4.5	Results without RFE	55
4.6	Inference Efficiency Analysis	55
4.7	Summary	57
4.8	Statistical significance test	57
4.9	Summary	58
5	Conclusion	61

List of Figures

3-1	UTD-MHAD data example	24
3-2	Data samples of HWU-USP Dataset	25
3-3	Examples of the sequences of depth maps for actions in MSR Action 3D dataset [16] (a) Draw tick and (b) Tennis serve	27
3-4	Overall high-level proposed model pipeline	28
3-5	Feature extraction backbone architecture.	32
3-6	Proposed model full architecture.	33
3-7	Mamba Fusion block architecture.	38
3-8	Improved Fusion Classifier block architecture.	39
4-1	Proposed Enhanced performance on all combined modalities UTD- MHAD (LOSO)	42
4-2	UTD-MHAD Proposed Efficient Model Per Fold Accuracy results . .	43
4-3	Modality Ablation Results on UTD-MHAD	47
4-4	UTD-MHAD Single Modality results per fold of Efficient Model . . .	47
4-5	UTD-MHAD Single Modality results per fold of Lightweight Model .	48
4-6	Per fold accuracy on MSR Action 3D dataset under LOSO CV	52

List of Tables

2.1	Comparison of recent human activity recognition studies.	17
3.1	Major changes between Proposed Efficient and Proposed Lightweight models	40
4.1	Combined multimodal classification results on UTD-MHAD (LOSO) .	42
4.2	Single-modality results on UTD-MHAD (LOSO)	43
4.3	Combined multimodal classification results on UTD-MHAD (Random-Split CV)	44
4.4	Single-modality results on UTD-MHAD (Random-Split CV)	44
4.5	Per-modality feature retention after RFE feature selection (50% retention) across two model configurations.	46
4.6	Prior work results on UTD-MHAD comparasion	48
4.7	Prior works with modality ablation results on UTD-MHAD	48
4.8	Overall multimodal classification results on HWU-USP (LOSO) . . .	49
4.9	Single-modality ablation results on HWU-USP (LOSO)	49
4.10	Overall multimodal classification results on HWU-USP (Random-Split CV)	49
4.11	Single-modality ablation results on HWU-USP (Random-Split CV) .	49
4.12	HWU-USP single modality testing results	50
4.13	HWU USP: Comparison with state-of-the-art.	50
4.14	Overall multimodal classification results on MSR Action 3D (LOSO)	51
4.15	Single-modality ablation results on MSR Action 3D (LOSO)	51

4.16 Overall multimodal classification results on MSR Action 3D (Random Split CV)	51
4.17 Single-modality ablation results on MSR Action 3D (Random Split CV)	51
4.18 MSR Action 3D: Comparison with prior work	52
4.19 Overall multimodal classification results on MM-Fit (LOSO)	53
4.20 Single-modality ablation results on MM-Fit (LOSO)	53
4.21 Overall multimodal classification results on MM-Fit (Random Split CV)	53
4.22 Single-modality ablation results on MM-Fit (Random Split CV) . . .	53
4.23 MSR Action 3D: Comparison with prior work	54
4.24 Summary of multimodal combined results across all datasets.	54
4.25 Model performance comparison with and without RFE	55
4.26 Average per-sample inference time (in ms) on UTD-MHAD dataset .	56
4.27 Average per-sample inference time (in ms) on HWU-USP dataset . .	57
4.28 Average per-sample inference time (in ms) on MM-Fit dataset	57
4.29 Average per-sample inference time (in ms) on MSR Action 3D dataset	57
4.30 Statistical significance test	58

Chapter 1

Introduction

Human Activity Recognition (HAR) is the classification of human actions within a specified timeframe. It underpins a wide range of applications: healthcare monitoring, fitness tracking, smart homes, and autonomous surveillance [1]. Accurate recognition of human actions unlocks better decision-making, deeper behavioral insight, and safer automation. Yet, HAR presents significant challenges as it requires handling of both spatial and temporal patterns well, that hold up against noisy or inconsistent sensor data, and that run fast enough on devices with limited processing power [17].

Most early HAR methods relied on a single type of input (either wearable sensors or camera feeds) and that created a ceiling on how robust and accurate they can be in real-world conditions [26]. Multimodal HAR approaches tackle these by pulling in data from multiple sources at once. Different sensors capture different things, and combining them gives models a much richer picture, which again translates into better recognition performance [25].

Early machine learning methods for HAR, such as SVMs, K-NNs, and Random Forests, depend heavily on hand-crafted features. This process is not only labor-intensive, but also is poorly generalizable to new datasets [10]. The introduction of deep learning (DL), especially CNNs, automated feature extraction, but these architectures still struggled with long-range temporal patterns needed for continuous activity sequences [24]. More recently, Transformer-based architectures have changed the approach. Their self-attention mechanism is naturally suited to modeling de-

dependencies across time and modalities, which is exactly what multimodal HAR demands [25].

Despite notable advancements, there are practical constraints that remain: memory, compute, and high inference times limit what can realistically be deployed on edge devices [15]. Achieving the right balance between accuracy and efficiency is still an open problem.

1.1 Motivation

Current multimodal HAR frameworks face two core issues. First, they are computationally expensive, which makes deployment on edge devices impractical. Second, they tend to be tightly coupled to specific sensor setups, so performance drops sharply whenever the available modalities change [17, 26].

This last point matters: in real-world settings, budget and logistics often make it impossible to deploy a full suite of sensors. A HAR system that only works well when every modality is present becomes not particularly useful.

This thesis sets out to build a lightweight, modality-agnostic multimodal HAR model that performs reliably even under constrained conditions. The goal is twofold: maintain strong recognition accuracy while enabling real-time, energy efficient inference on edge hardware. Central to this is learning robust cross-modal representations and handling missing modalities adaptively, without requiring supervision tailored to any single input type.

To keep the model compact, the proposed approach replaces heavyweight backbones like ResNet with smaller, more efficient alternatives for each modality. Rather than applying post-hoc compression techniques, the architecture is designed from the outset to have a low parameter count. Evaluation goes beyond test accuracy alone: inference time, memory consumption, total parameter count, and model size are all measured to give a complete picture of deployment viability. Performance is assessed both on a standard machine and on a Raspberry Pi 4 Model B, providing a direct measure of real-world edge feasibility.

The system is evaluated on established multimodal HAR benchmarks, UTD-MHAD, HWU-USP, MSR-Action3D, and MM-Fit. The central contribution is a lightweight fusion model built around a unified Mamba and Transformer-inspired architecture. Its design is flexible: it accomodates dynamic sensor configurations and performs adaptive fusion regardless of which or how many modalities are present.

Chapter 2

Related works

2.1 Multimodal HAR

Table 2.1 offers a comparative look at recent HAR approaches, their architectures, results, and modalities used.

Table 2.1: Comparison of recent human activity recognition studies.

Study	Year	Architecture	Accuracy (%)	Lightweight	Multimodal	Video	Depth	Skeleton	Inertial	Modality-agnostic
[32]	2020	DSIEM (MLML)	MSR Action3D: 93.55%; UTD-MHAD: 88.37%	Yes	Yes	No	Yes	Yes	No	No
[22]	2021	Dynamic Images + CCA + LSTM	UTD-MHAD: 98.88, NTU RGB+D: 75.5	No	Yes	Yes	Yes	Yes	No	No
[12]	2022	Mumu	UTD-MHAD: 97.6	No	Yes	Yes	Yes	Yes	Yes	No
[18]	2022	GCN-based Dynamic WHAR	Oppo: 74.27, RealWorld: 76.63, RealDisp: 92.58, Skoda: 93.80	Yes	Yes	No	No	No	Yes	No
[4]	2024	CNN-GRU classifier + novel feature selection	UTD-MHAD: 97.9; HWU_USP: 97.99 w/o RGB: 92.27, w/o inertial: 93.09	No	Yes	Yes	Yes	No	Yes	No
[5]	2025	TinierHAR	PAMAP2: 77.4; MHEALTH: 94.4; USCHAD: 73.4	Yes	No	No	No	No	Yes	No
[36]	2025	Multi-Stream Multimodal Factorized Transformer	UTD-MHAD: 99.3; UP-Fall: 96.89; MMAct: 81.07	No	Yes	Yes	No	No	Yes	No
[13]	2025	Multimodal feature fusion with HCT	UCF101: 97.99, UCF50: 96.69, HMDB51: 75.82	No	Yes	Yes	No	Yes	No	No
[3]	2026	MultimodalMamba (DHCNN-FAM-Mamba)	HWU-USP: 99.1, Berkeley-MHAD: 98.4, UTD-MHAD: 97.7	Yes	Yes	Yes	Yes	No	Yes	No
	2026	Proposed		Yes	Yes	Yes	Yes	Yes	Yes	Yes

Yang et al. [32] introduced Depth Sequential Information Entropy Maps (DSIEM), which extract spatial-temporal features from depth sequences by computing entropy over orthogonal projections. HOG descriptors were layered on top to capture motion patterns and cut down on redundancy. Through their Multi-label Subspace Learning (MLSL) method, these features were fused with skeleton data via projection matrices. The results were solid: 93.55% on MSR-Action3D, 88.37% on UTD-MHAD, and

90.66% on UTD-MHAD-Kinect V2.

Batool et al. [4] went in a different direction, constructing a multimodal framework around RGB, depth, and inertial sensor data. They put forward several novel descriptors: Dynamic Likelihood Random Field (DLRF), Angle along Sagittal Plane (ASP), Lag-Regression (LR), and Gammatone Cepstral Coefficients (GCC), and fused them using a Genetic Algorithm. For classification, a CNN-GRU pipeline was used, substituting Kalman Gain for ReLU to mitigate vanishing gradients. Under LOSO k-fold evaluation, this setup reached 97.99% on HWU-USP and 97.90% on UTD-MHAD.

More recent work has turned to Transformers. Khan et al. (2025) [13] pair a Spatial Vision Transformer Network (SVTN) with YOLOv8-derived skeleton keypoints, producing accuracies of 75.82% on HMDB51, 97.98% on UCF101, 96.68% on UCF50, and 99.12% on YouTube Actions. In parallel, Zhou et al. (2025) [36] introduced MSMFT: a multi-stream factorized Transformer that tokenizes video and sensor input via Statistical Feature Projection (SFP) and Temporal Embedding (TE). F1-scores of 93.06% on MMAAct (cross-session), 99.30% on UTD-MHAD, and 99.01% on UP-Fall were achieved. Together, these studies mark a clear shift toward Transformer-driven integration of vision, skeleton, and sensor modalities.

2.1.1 Multimodal HAR on edge devices

Multimodal approaches matter for HAR on edge devices because no single sensor does everything well; cameras blur during movement; IMUs miss actions with little body motion [19].

Nakabayashi et al. [19] deployed an attention-based multimodal model on the Sony SPRESENSE board (2MB RAM) using CNN encoders paired with a 3-layer MLP classifier. On a custom dataset of just 5 activity classes, the system reached 97% precision. That said, unimodal evaluation showed substantially lower scores, suggesting that the model leans heavily on IMU features and lacks robustness when modalities go missing.

Bian et al. [5] went after efficiency directly with TinierHAR, an ultra-lightweight design that stacks residual depthwise separable convolutions, bidirectional GRUs,

and attention-based temporal aggregation. Across 14 public HAR benchmarks, it cut parameters by 2.7x and MACs by 6.4x compared to prior TinyHAR without sacrificing macro-F1.

Xie et al. [30] focused on early fusion, using skeleton-guided attention to align RGB frames with skeletal data and suppress background noise. On UTD-MHAD, their BiLSTM variant hit 91.5% accuracy (29.4M parameters, 17.3G FLOPs), while the ST-GCN version did slightly better at 92% accuracy (31.1M parameters, 26.5G FLOPs).

Avinash et al. [3] proposed MultimodalMamba, which brought together Dilated Hierarchical CNNs, Frequency Attention, and Mamba state-space modeling. It was tested on HWU-USP (99.1%), Berkeley-MHAD (98.4%), and UTD-MHAD (97.7%), outperforming CNN, Transformer, and prior Mamba baselines.

2.2 Modality-Agnostic models

Modality-agnostic models are built to handle any combination of inputs they receive (images, sensor streams, text) without needing to know in advance which ones will be available. The idea is straightforward: project everything into a latent space so that a single encoder can process all modalities through one unified pipeline [9].

For HAR, this flexibility is essentially valuable. Activity recognitions learned from RGB, skeleton or IMU data simultaneously tend to hold up better when one or more modalities drop out [34]. For instance, Romaissa et al. [22] convert all sensor inputs into dynamic images via rank pooling, while skeleton coordinates are rendered as RGB images. Once everything looks the same, pretrained ResNet50 or AlexNet can be applied uniformly: no modality-specific engineering is required.

Transformer-based intermediate fusion has pushed things further. Mumu [12] used cooperative multi-task learning with adaptive attention, letting the model dynamically weight whichever sensors are most reliable when others are noisy or failing. HAMLET [11] took a different angle, applying hierarchical self-attention to disentangle and fuse spatiotemporal features from RGB and depth streams.

UCFFormer [31] is perhaps the most direct attempt at true modality-agnostic design. It built a shared embedding space for joint temporal-modality modeling using a Factorized Time-Modality Transformer (FTMT), which decomposed the attention computation to reduce complexity without losing cross-modal interactions. On UTD-MHAD and NTU RGB+D, it reached 99.04% and 97.5% accuracy respectively.

There is a common limitation across all of these methods, though. They all rely on ResNet-50 as the backbone for the RGB modality. That alone introduces a large number of parameters, making edge deployment effectively out of reach.

91.

Chapter 3

Methodology

3.1 Datasets

Experiments were mainly conducted on rich multimodal datasets. UTD-MHAD and HWU-USP were selected as core, because of their accessibility, feasible size, and inclusion of modalities of different types: inertial, skeleton, depth camera, and RGB videos. UTD-MHAD was selected as main dataset on which proposed approach was tested and a similar pipeline will be tested on the remaining ones.

3.1.1 UTD-MHAD

UTD-MHAD [6] is a comprehensive and extensively annotated dataset for human action recognition designed specifically to study sensor fusion between a depth camera and a wearable inertial sensor. It provides four temporally synchronized modalities per trial: RGB video, depth video, 3D skeleton joint positions (from Kinect), and inertial signals captured by a wearable IMU sensors. The samples of the data could be seen in Figure 3-1 in coressponding order (from top to bottom).

The dataset contains 27 action classes performed by 8 subjects (4 female, 4 male), with each subject repeating each action 4 times; after removing three corrupted sequences, the released dataset includes 861 total sequences. Data were collected indoors. The actions list is as follows: Swipe left, Swipe right, Wave, Clap, Throw,

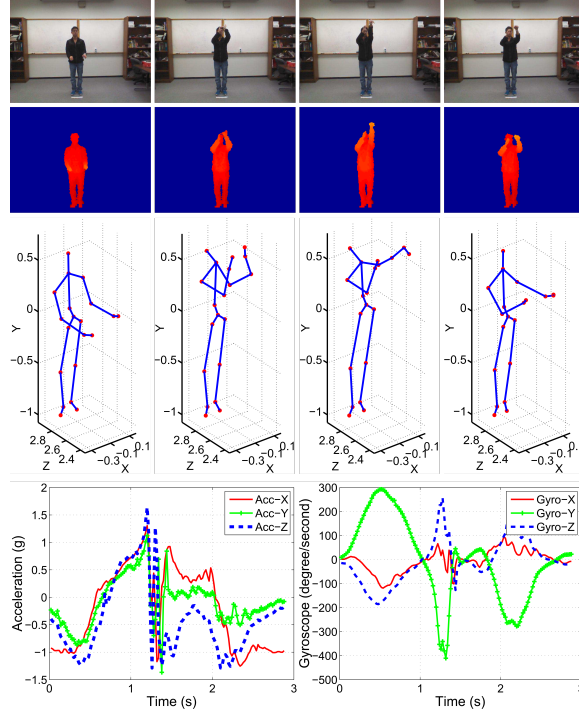


Figure 3-1: UTD-MHAD data example

Arm cross, Basketball shoot, Draw X, Draw circle (clockwise), Draw circle (counter clockwise), Draw triangle, Bowling, Boxing, Baseball swing, Tennis swing, Arm curl, Tennis serve, Push, Knock, Catch, Pickup and throw, Jog, Walk, Sit to stand, Stand to sit, Lunge, Squat.

3.1.2 HWU-USP

Heriot-Watt University/University of Sao Paulo (HWU-USP) dataset was collected by the Robotic Assisted Living Testbed at Heriot-Watt University [20] There are 16 healthy adult subjects performing the set of 144 activities for a long period of time. The activities list includes: making a cup of tea, making a sandwich, making a bowl of cereals, setting the table, using a laptop, using a phone, reading a newspaper, cleaning the dishes, tidying the kitchen. RGB videos and Depth images were collected using TIAGo robot, placed in the corner of the kitchen. Inertial data was collected via wrist-smartwatches and waist-band sensors on participants.

This is a second complete multimodal dataset, which was used for verification of

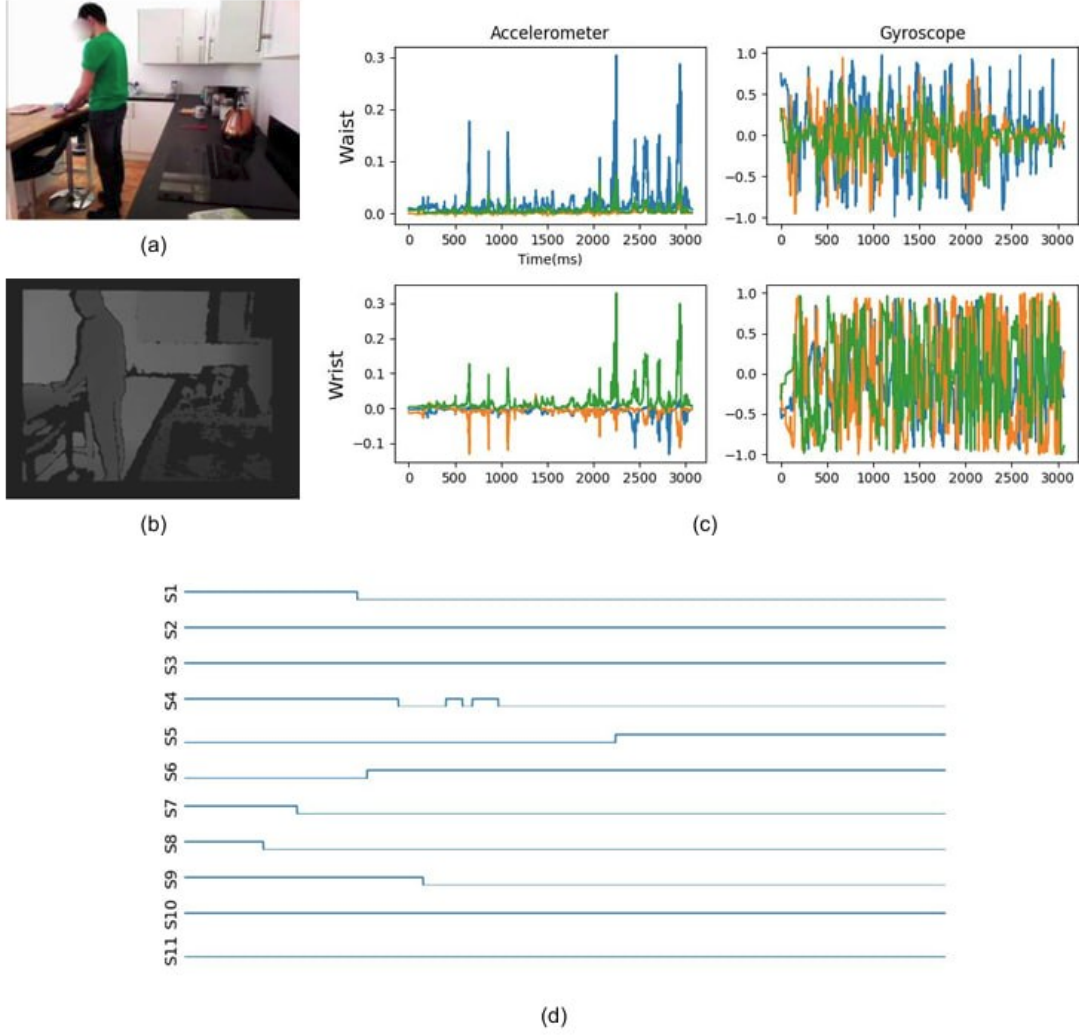


Figure 3-2: Data samples of HWU-USP Dataset

our methods for two reasons: 1) Different domain: it contains a very different set of heterogeneous activities of daily living. 2) Matching modalities with UTD-MHAD makes the dataset evaluation more consistent.

The HWU-USP [20] data examples from one timestep are shown on Figure 3-2 with a): RGB video frames; b) a depth image; c) raw data from the inertial sensors, related to whole sequence; d) raw data from the ambient sensors in binary form.

3.1.3 MSR Action3D

To test the model’s modality-agnostic feasibility, its reliability was evaluated across different combinations of modalities in the datasets. One of the multimodal datasets that suited my project was MSR Action3D, collected by Wanqing Li [16]. It consists only of depth and skeleton modalities. The data was captured using a depth sensor similar to the Microsoft Kinect device, recording at a resolution of 640×240 pixels. The skeleton representation consists of 20 body joints, represented by 3D coordinates (X, Y, Z). In total, the dataset contains 567 depth map sequences. There are 10 subjects, performing 20 actions for 2-3 times. The action list is standard: Arm/hand gestures, body movements, sports actions. The sample depth sequences are displayed on Figure 3-3. Despite its small size, MSR Action3D serves as a good benchmark for agnostic multimodal HAR research, as it is characterized by a combination of depth and skeleton modalities.

3.1.4 MM-Fit

The fourth dataset used in the study serves as unique combination of skeleton and inertial data. MM-Fit dataset was created by Strömbäck D., Huang S., and Radu V. in 2020 and it contains over 800 minutes of multimodal data of physical exercises recordings collected from participants. [27]. Data is captured from 5 synchronized devices: Orbbec Astra Pro rgb and depth camera, eSense earbud, smartwatch Mobvoi TicWatch Pro, Samsung S7 and Huawei P20 smartphones for accelerometer, gyroscope modalities collection. Skeleton data (2D and 3D skeleton keypoints) is derived from camera data. There are 11 activity classes, with one empty class and 10 others are composed of full-body workout exercises, such as: squats, lunges with dumbbells, sit-ups, push-ups, triceps extensions sitting overhead dumbbell, bicep curls, jumping jacks, sitting dumbbell shoulder press, standing dumbbell rows, and dumbbell lateral shoulder raises. The main reason of selecting this dataset is different nature of actions, multimodality, and public availability. Since the goal of the model is to be robust to any missing modality, it should also take in account the different nature of

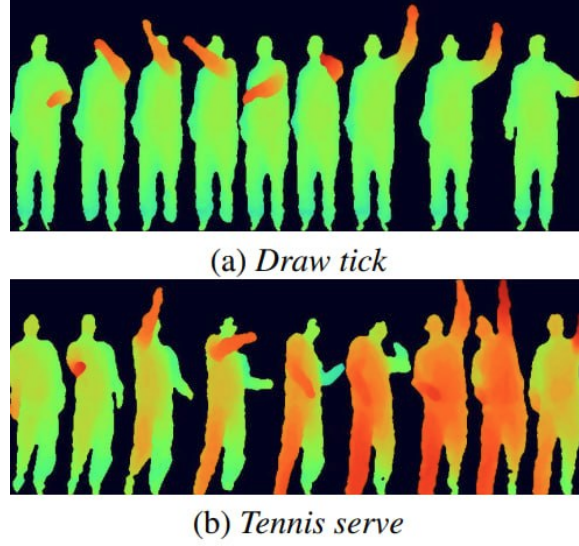


Figure 3-3: Examples of the sequences of depth maps for actions in MSR Action 3D dataset [16] (a) Draw tick and (b) Tennis serve

actions, to obtain its adapting performance.

3.1.5 Testing strategy

The Leave-One-Subject-Out (LOSO) approach was utilized to ensure the generalization of the model on the unseen subjects. Additionally, k-fold cross-validation technique was adopted to ensure a more real, aggregated representation of data. For the UTD-MHAD, that meant that a single subject was dedicated to the test set at each fold, another subject for validation, with the rest being assigned for training, thus covering every person in the dataset. In the UTD-MHAD case, 6 people were selected as training set, 1 for validation, 1 for testing. For the HWU-USP dataset, 12 people were selected as training set, 2 for validation and 2 for testing. For the MSR Action3D Dataset, 6 people were selected as training set, 2 for validation, 2 for testing. When it comes to MM-Fit, there was no explicit division by subjects in that dataset; instead it was compiled into 21 workout sessions. Therefore, I applied the same k-fold logic on the workouts: 17 for train, 2 for validation, 2 for test. For evaluating the model, the average accuracy across all folds was selected. This approach makes the model more strict, reliable and generalizable to unseen subjects.

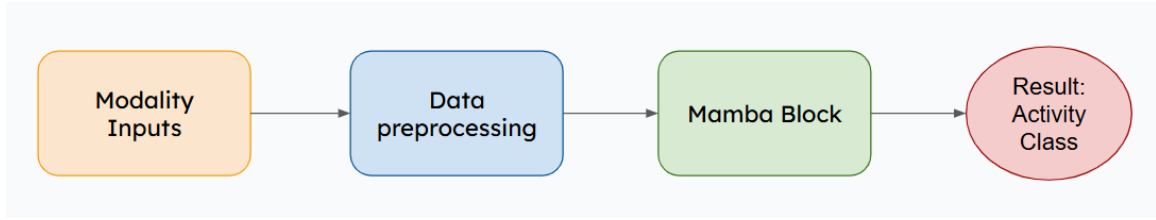


Figure 3-4: Overall high-level proposed model pipeline

Note that due to memory constraints on the laboratory PC, the large-scale multimodal datasets, such as NTU RGB+D and UR Fall detection which weights over 1.2TB, 850 GB respectively had to be excluded.

3.2 Model pipeline

The figure 3-4 shows the high-level pipeline of the whole process of my proposed models. At the start, modality inputs are provided, then for each corresponding modality, their respective features and data preprocessing procedures occur, after this stage the obtain features per modality are each processed through the mamba block (the whole mamba-based model) and then as a result of this model, a class probability matrix for each sample, i.e. action class label.

3.3 Data preprocessing

YOLO

The very first data handling step of my proposed method is to isolate the human subject from background clutter. For this, an offline person-detection stage is performed using YOLOv8n, a lightweight object detector configured to detect only the person class. The reason behind selection of YOLOv8n [35] is it being the most smallest, lightweight version of the object detection modules provided by Ultralytics. YOLOv8n runs on every frame of every RGB video, detects human silhouette and labels it with bounding boxes, selects the largest one per frame, pads it by 25% on all sides, and saves the per-frame box coordinates to external storage. This only runs

once and is cached as NumPy arrays and being used for later stages, so no extra computational cost is present in the training stage. There are two usage modes of them: 1) Per-frame boxes used by RGB modality loader. Each frame is cropped individually to its own detected box; 2) Union box used by Depth Motion Maps (DMM) computation. Since DMMs aggregate motion across multiple frames, so a consistent spatial region is required, the DMM module takes the smallest left corner and largest right corner from YOLO crops and produces one big rectangle that encompasses every position the person occupied throughout the entire video.

3.3.1 RGB data preprocessing

RGB produces two separate tensors: a simple RGB frame stream and an optical flow stream.

After obtaining YOLO-based bounding boxes, Sobel filtering was added to ensure privacy of the subject. A minimum size check of 10×10 pixels prevents degenerate crops. During loading, T frames are uniformly sampled from the video using linear temporal indexing. Each cropped frame is resized to the target spatial resolution (196x196 for Efficient version and 128x128 for Lightweight) and normalized using the ImageNet channel-wise mean weights and standard deviation. The normalization of data happens for train and test for each fold separately, therefore discarding data leakage.

Optical Flow

The RGB images solely are vulnerable to lighting conditions, clothing color, skin tone, and background clutter, which do not contribute anything to the action itself. To mitigate this error, the second parallel RGB stream called optical flow was included as well in the final model. As Sevilla-Lara et al. [23] proved, random permutation of colormap in RGB images reduced overall action recognition accuracy by half, while optical flow remain mostly unaffected by this. Since RGB images and optical flow capture fundamentally different and non-redundant aspects of the scene:

appearance context and motion dynamics, they complement each other in activity recognition task. The study of Sun et al. [28] showed that on UCF-101 dataset, RGB baseline achieved 93.3% accuracy recognition, while adding optical flow-guided features increased accuracy to 96.0%. Optical flow, or motion maps is a simple frame-differencing operation applied on each pair of consecutive RGB frames ($|f_{t+1} - f_t|$). Therefore, RGB modality handling was expanded to dual-stream: sample images and optical flow are processed separately, as it pictured on Fig 3-6. The optical flow backbone is MobileNetv3-Small pretrained on the ImageNet1K dataset. To ensure consistency with Backbone operations, RGB images are normalized and expanded to 3 channels.

Depth Motion Maps

Rather than feeding raw per-frame depth images, the notebook collapses temporal information into Depth Motion Maps (DMMs). These are spatial images that summarize where movement happened in each time window. They are utilized to capture motion cues in a depth video sequence as well as reduce computational cost [21]. By projecting depth onto three orthogonal views simultaneously, DMMs encode the same action from front, side, and top perspectives. The depth images' total number of frames is then partitioned into $T = 12$ equal segments, and within each segment, absolute inter-frame differences are accumulated to produce a single DMM that encodes the spatial distribution of motion.

Then, each depth frame underwent noise removal procedure with a 3×3 median blur applied on raw depth images to remove salt-and-pepper noise common in depth sensors, and normalised to the unit interval using valid (non-zero) pixel statistics.

To suppress irrelevant background regions, a union bounding box is derived from the YOLO crops computed on the corresponding RGB video and rescaled from the RGB resolution (640×480) to the depth resolution (320×240). The resulted single-channel DMM is stacked 3 times to create a pseudo-RGB image, enabling it to be processed by ImageNet-pretrained EfficientNet-B0, which expects 3-channel input.

3.3.2 Skeleton Data preprocessing

Skeleton data was combined with derived bone features. Normalization was done by using body height normalization, all joints relative to the hip center joint were centered, then divided by the mean head-to-hip distance. The inclusion of additional eigenjoint-based features tends to improve accuracy of action recognition task. The eigenjoints-based approach of Yang et al. [33] on MSR Action 3D dataset achieved cross subject test accuracy of 83.3%, larger than HOJ3D (Histogram of 3D joints) method with 79.9% accuracy. Additional bone features extends each joint from 3 to 0 channel stream by adding: bone direction (normalized vector from parent joint and current joint); bone angles (sin, cos, cross product of magnitude between parent and grandparent bone vectors); velocity (temporal first derivative of $|joint_t - joint_{t-1}|$), it captures instantaneous motion speed and direction per joint; pairwise-context (for each joint, the average difference between that joint and every other joint); displacement pPosition relative to the first frame).

3.3.3 Inertial data preprocessing

Inertial measurement unit (IMU) data, typically comprising triaxial accelerometer and gyroscope readings (6 channels), is loaded and temporally partitioned into T non-overlapping segments of equal duration. Within each segment, all readings are averaged to produce a single representative feature vector, yielding robust temporal downsampling that mitigates the effects of varying sampling rates across recordings. If the number of available channels exceeds the expected count, only the first 6 channels are retained; if fewer channels are present, zero-padding is applied to maintain a consistent dimensionality. Then applied Z-score normalization per channel: $(x - mean) / (std + 10^{-8})$.

3.3.4 Ambient data preprocessing

Ambient sensor data, which stores environmental context such as temperature, humidity, barometric pressure, illuminance, passive infrared (PIR) motion detection,

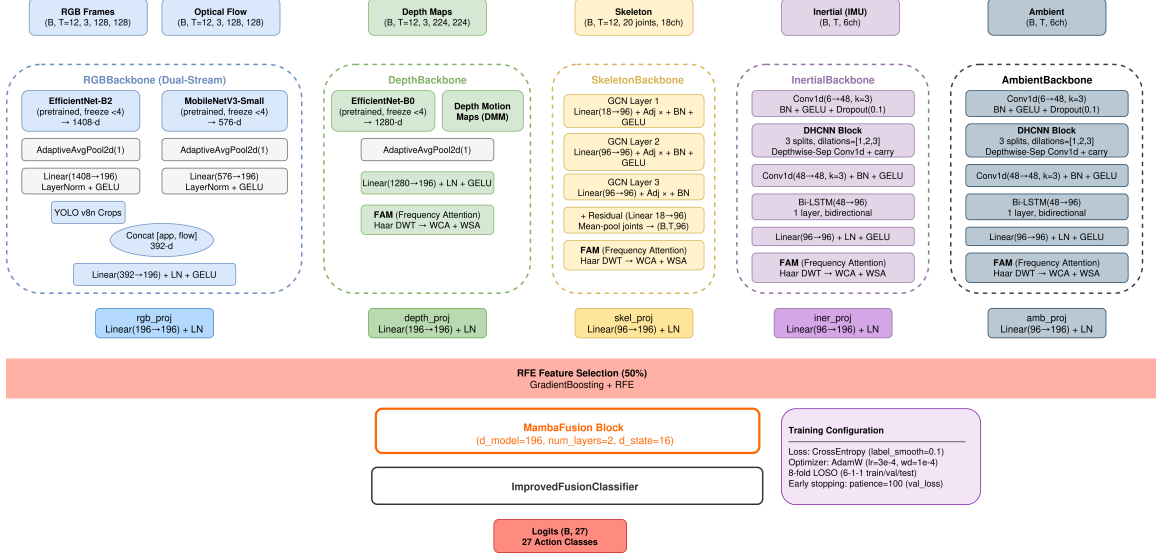


Figure 3-5: Feature extraction backbone architecture.

and door or cabinet contact sensors, is loaded from CSV files with automatic detection of the number of numeric columns. The temporal processing follows the same segment-and-average scheme as the inertial modality: the signal is divided into T equal-length windows, each summarized by its mean. Channel count adjustment (truncation or zero-padding) ensures a fixed-size representation, followed by per-channel z-score normalization. In addition to the scaling and Gaussian noise augmentations used for inertial data, a channel dropout strategy is employed during training, in which a randomly selected sensor channel is zeroed out with 30% probability. This augmentation encourages robustness to individual sensor failures, which are common in smart-home deployments where ambient sensors may intermittently disconnect. The output tensor has shape (T, N_a) , where N_a is the number of ambient sensor channels (auto-detected at runtime, typically 16 for the HWU-USP dataset).

An overview of the per-modality feature extraction backbones can be view in Figure 3-5

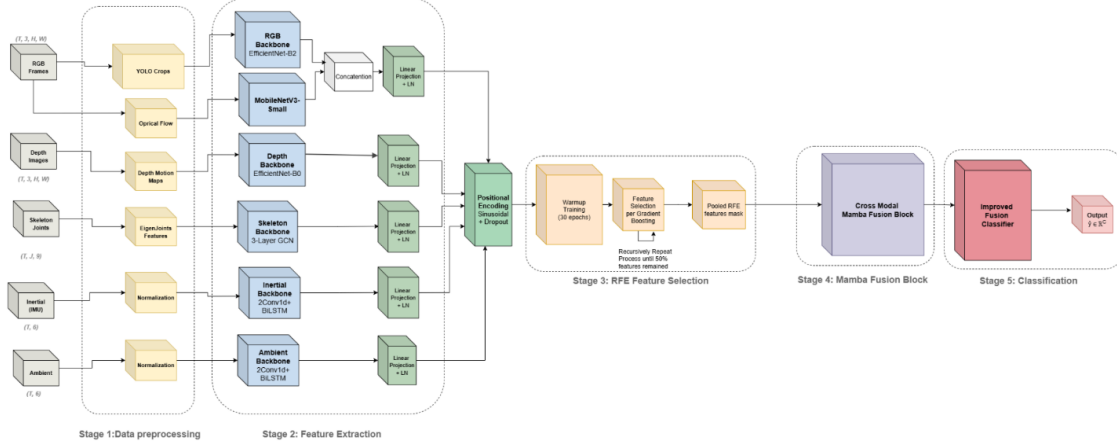


Figure 3-6: Proposed model full architecture.

3.4 Proposed Pipeline

Since my thesis goal was to create a modality agnostic model, robust to any missing modality in action recognition task, the existing solutions were studied. One of them is UCFFormer model (Unified Contrastive Fusion Transformer) [31]. The modality backbone placeholder were inspired by their architecture. Also, since Transformers are computationally heavy, the alternative approach was to search with a goal of creating a more lightweight-oriented model. One of the recent solutions was DHCNN-FAM-Mamba [3]. Motivated by efficiency of Mamba-based approach, I tried to create the final model, which resembles a fusion between two approaches. The fundamental difference between their work and my proposed model is that in [3] all modalities are concatenated into one flat vector first, and then DHCNN, FAM, and Mamba process this already-mixed representation together. However, in my case, to achieve modality-agnostic behavior, each modality has its own specialized backbone, its own independent Mamba temporal processing, and only then communicates with other modalities through a dedicated Cross-Modal Mamba Fusion block. Additionally I extend the framework to five modalities including skeleton and ambient. The full model architecture is in Figure 3-6.

RFE

Before entering the Mamba block, separate per-modality obtained features are sinusoidally positionally encoded and fed to the feature selection method. Recursive Feature Elimination (RFE) is a wrapper-based feature selection method that iteratively identifies and removes the least discriminative features using a model’s internal importance scores. In the proposed model, RFE is applied as a training-time procedure operating on the projected backbone features before they enter the Mamba fusion stage. Specifically, the full model is first trained for 30 warmup epochs to ensure that the per-modality backbone features are sufficiently meaningful for importance evaluation. Following warmup, features are extracted from each modality backbone after the linear projection layer and mean-pooled over the temporal dimension, yielding a fixed-size feature vector per modality per sample. These per-modality vectors are concatenated into a unified representation and passed to a GradientBoosting classifier, which serves as the RFE estimator due to its ability to produce reliable feature importance scores on small datasets. RFE then iteratively ranks all features by their contribution to action classification and eliminates the bottom 10% at each step, retraining the estimator on the reduced feature set after each elimination round. This process repeats recursively until the target amount: 50% of the original features remain. The resulting binary mask — indicating which feature dimensions to retain — is applied element-wise to the projected modality tokens during full model training, effectively zeroing out uninformative dimensions before they enter the Mamba fusion block. At the end of the feature selection stage, the warmup model is discarded after mask computation and the full model is retrained from scratch with the masks applied, meaning RFE adds zero parameters to the final deployed model and introduces no additional computational cost at inference time.

3.4.1 Feature Selection

In multimodal HAR systems, the features extracted from heterogeneous sensors inevitably carry a degree of redundancy. Not to introduce noise, we explored feature

selection as a post-backbone pruning step, applied independently to the whole projected feature vector after fusion.

Our initial experiments considered nature-inspired metaheuristic approaches. Genetic Algorithm (GA) was tested first, as it has shown promise in feature subset optimization for classification tasks [7]. However, the computational overhead was too large: each fitness evaluation required a full forward pass through the pretrained backbone and fusion pipeline, and the population-based search over 196-dimensional feature spaces per modality made the runtime impractical within our training budget. Before entering the Mamba block, separate per-modality obtained features are sinusoidally positionally encoded and fed to the feature selection method.

We therefore turned to Recursive Feature Elimination (RFE) [8], a wrapper-based method that takes a different approach to the same problem. Rather than searching combinatorially, RFE fits a base estimator (in our case, a Gradient Boosting classifier with 50 trees, max depth 3, and a learning rate of 0.1), ranks features by their learned importance scores, and discards the bottom 10% at each iteration. This cycle repeats until a target number of features remains. For all experiments reported here, we retained 50% of the original feature dimensions.

One practical detail worth noting: RFE is not applied from the start of training. The backbones need time to learn meaningful representations before any importance ranking is reliable. We therefore train normally for 30 warmup epochs, then extract backbone features from the training set, run RFE, and store the resulting binary mask. From that point onward, the mask is applied element-wise to the projected features—zeroing out the eliminated dimensions for the remainder of training and at inference.

To validate whether this step actually helps, we ran a controlled A/B comparison on the UTD-MHAD dataset under identical training conditions.

The small accuracy gain came entirely for free in terms of model complexity. The parameter count and storage footprint remain unchanged because RFE does not remove any weights from the network; it simply masks a subset of feature dimensions at runtime. The vast majority of parameters reside in the pretrained back-

bones (EfficientNet-B2, EfficientNet-B0, MobileNetV3-Small), which are unaffected by downstream feature masking. What RFE does provide is a form of implicit regularization: by forcing the fusion and classification layers to rely on a curated subset of features, it reduces the risk of overfitting to noisy or redundant dimensions, which matters especially on a dataset as small as UTD-MHAD (861 sequences across 27 classes).

Frequency Attention Module

The Frequency Attention Module (FAM) helps the network emphasise the most discriminative frequency components in each modality’s features. It works by applying a Haar wavelet transform function to decompose backbone outputs into low-frequency and high-frequency subbands, then learning a set of per-channel attention weights based on the energy distribution across those subbands. Two separate attention paths were employed: one is looking at overall frequency content, the other is contrasting low versus high frequencies. Each of them produce scaling factors through a small bottleneck MLP. Their average is applied element-wise to reweight the original features:

$$\hat{x} = x \odot \frac{\alpha_{\text{ch}} + \alpha_{\text{sp}}}{2}. \quad (3.1)$$

FAM is placed after the backbone projection in the depth, skeleton, inertial, and ambient branches. The RGB branch does not use it, because its dual-stream architecture already captures temporal frequency information through optical flow. The wavelet filters are fixed and the MLPs are small, so the added cost is minimal.

Mamba Block

We adopt Mamba [3], a selective state space model (SSM) that processes sequential data in linear time. Unlike classical SSMs with fixed dynamics, Mamba makes its state transition parameters input-dependent, allowing the model to decide at each time step how much prior context to retain and how much to overwrite with new

information.

We use a bidirectional variant where each input is scanned both forward and backward through the SSM, and the two outputs are combined. This ensures the model has access to both past and future context at every position in the sequence. A gating mechanism controls the flow of information, and residual connections preserve gradient flow across layers.

The fusion module, **MambaFusion**, works in two stages. In the first stage, each modality’s temporal sequence passes through a stack of $L = 2$ bidirectional Mamba blocks with sinusoidal positional encoding, producing modality-specific temporal representations. In the second stage, a cross-modal Mamba block refines each modality: its sequence is concatenated with the representations from all other modalities and processed jointly, but only the tokens belonging to the original modality are kept. A feed-forward network with a residual connection completes the refinement. This way, each modality can draw on complementary cues from the others without losing its own structure.

The SSM state dimension is set to $N = 16$ with an expansion factor of 2, and dropout of 0.2 is applied after each block. The fusion module outputs refined per-modality sequences, enriched with the context about other modalities are then pooled and combined through attention-weighted aggregation in the classifier head. The full architecture of Mamba Fusion block is presented in Figure 3-7.

Improved Fusion Classifier

The last stage of my model is the classifier block. The Improved Fusion Classifier takes the five refined token sequences output by the Cross-Modal Mamba Fusion block and produces a single predicted action class. There are five steps involved. First, each modality’s $T=12$ frame tokens are averaged over time into one summary vector, thus collapsing the temporal dimension so each modality is represented by a single 196-dimensional vector. Second, these five vectors are stacked into a matrix of shape 5×196 . Third, a small attention network scores each modality: each vector is compressed to 49, Tanh is applied, then linearly compressed from 49 to 1, and

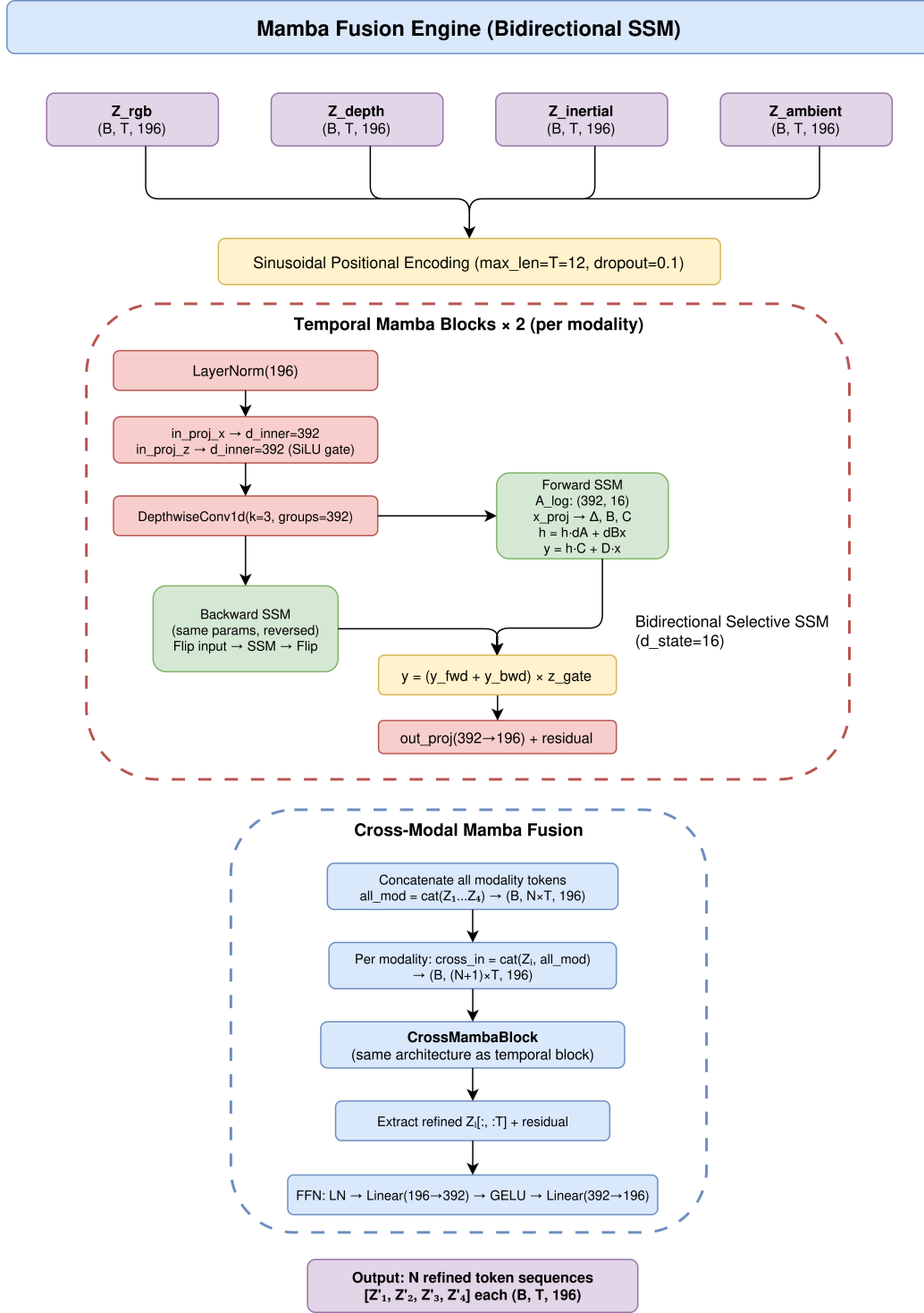


Figure 3-7: Mamba Fusion block architecture.

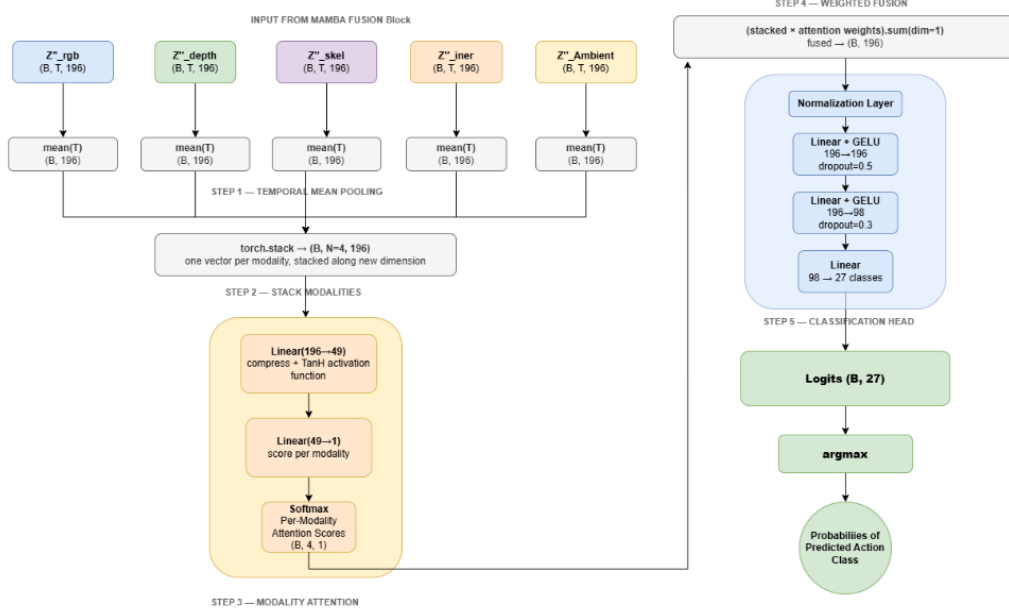


Figure 3-8: Improved Fusion Classifier block architecture.

finally goes through softmax activation function, producing five weights. The model dynamically decides how much to trust each modality for each specific sample, rather than treating all modalities equally. Fourth, the five modality vectors are scaled by their attention weights and summed into a single fused 196-dimensional vector. Lastly, this fused vector passes through a three-layer classification head: normalization layer with GELU activation and aggressive dropout between 0.3 and 0.5, producing logits over action classes. The argmax of these logits gives the predicted activity class.

Loss Function

We use standard cross-entropy with label smoothing ($\epsilon = 0.1$) as the training objective. Learning rate follows a cosine schedule with 5-epoch linear warmup, decaying from 3×10^{-4} to 10^{-6} . Early stopping monitors validation loss with a patience of 100 epochs out of a 300-epoch budget.

The Lightweight version

Table 3.1 describes the differences between two proposed models. The lightweight version uses lighter backbones, smaller embedding dimensions and number of temporal

frames, leading to slightly lower accuracy at significantly lower parameter count.

Table 3.1: Major changes between Proposed Efficient and Proposed Lightweight models

Component	Proposed Efficient	Proposed Lightweight
Embedding dimension d	192	128
Temporal frames T	12	8
RGB backbone	EfficientNet-B2	MobileNetV3-Small
Depth backbone	EfficientNet-B0	MobileNetV3-Small
Skeleton backbone	3-layer GCN	2-layer GCN
Inertial backbone	$2 \times \text{Conv1D} + \text{BiLSTM}$	$1 \times \text{Conv1D} + \text{BiLSTM}$
Ambient backbone	$2 \times \text{Conv1D} + \text{BiLSTM}$	$1 \times \text{Conv1D} + \text{BiLSTM}$

Chapter 4

Results

This chapter presents the experimental results of the proposed multimodal activity recognition framework. Two model variants are evaluated: the Proposed Efficient Model (full architecture) and the Proposed Lightweight Model (an even further reduced-parameter variant designed for resource-constrained deployment). For each benchmark dataset, classification performance is reported using all available modalities jointly, followed by a modality ablation study in which each modality is used independently. All results are reported in terms of test accuracy (%), number of trainable parameters, and model size in megabytes (MB). The chapter concludes with an analysis of inference efficiency across all three datasets. The models have been implemented using PyTorch and trained on NVIDIA GeForce GTX 1080 Ti GPU with 11.7 GB of RAM.

4.1 Results on UTD-MHAD

This section reports the classification results obtained on the UTD Multimodal Human Action Dataset (UTD-MHAD). Table 4.1 presents the overall performance of both proposed models when all four modalities are fused, while Table 4.2 details the single-modality ablation results. First, the LOSO CV based will be shown and then, non-LOSO random-split results will be also provided.

Table 4.1: Combined multimodal classification results on UTD-MHAD (LOSO)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	93.8 $\pm$ 5.4 std	14,666,930	56.14
Proposed Lightweight Model	93.1 $\pm$ 3.6 std	3,677,648	14.18

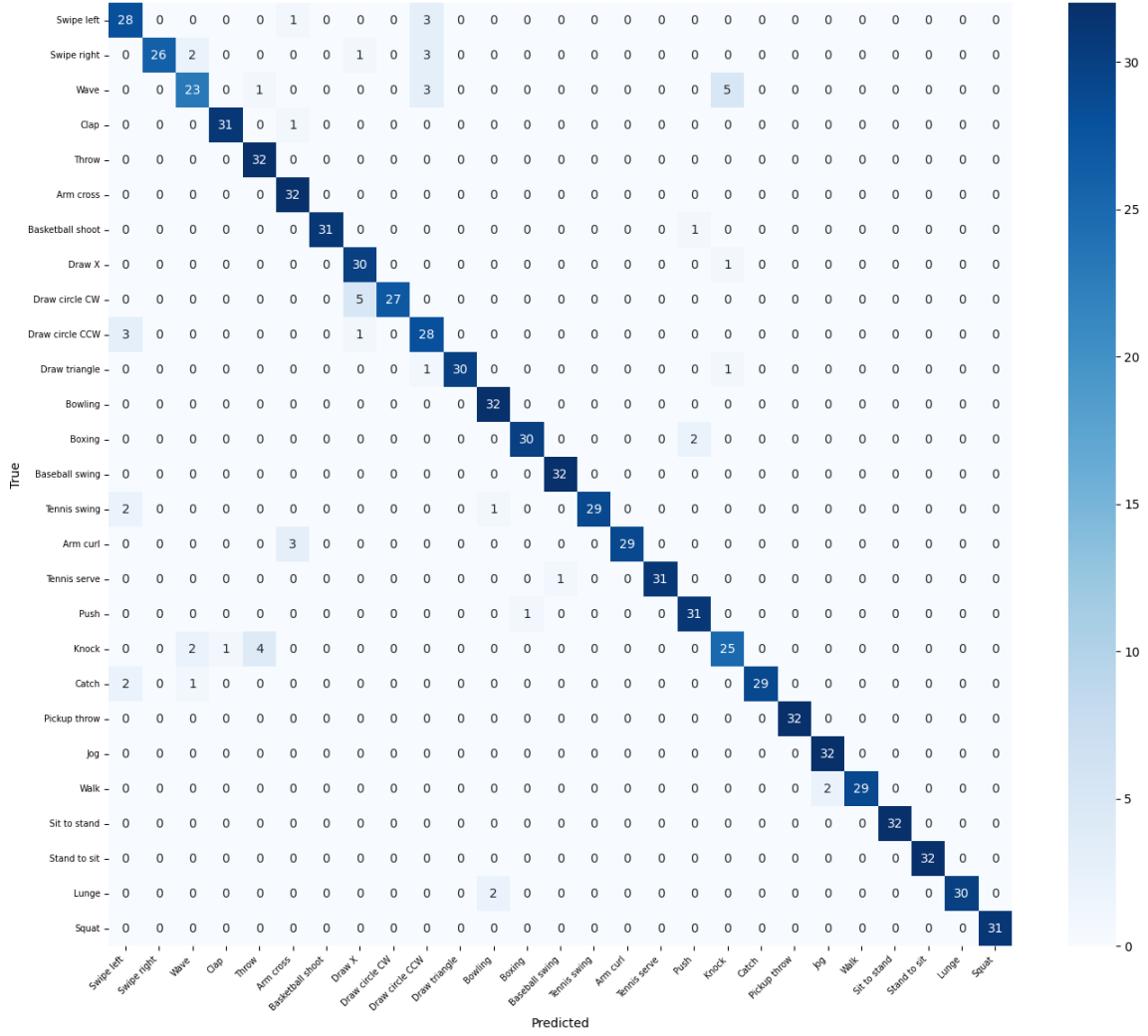


Figure 4-1: Proposed Enhanced performance on all combined modalities UTD-MHAD (LOSO)

When leveraging all available modalities (RGB Video + Depth + Skeleton + Inertial), the Proposed Efficient Model achieves a test accuracy of 93.8%, with 14.6M

trainable parameters and a model size of approximately 56 MB per fold. The Proposed Lightweight Model, designed for efficiency, attains a slightly lower 93.1% accuracy while reducing the parameter count by almost 4 times to 3.67M and the model size to only 14 MB. This demonstrates that a significant reduction in model complexity can be achieved with a controlled trade-off in classification performance.

Table 4.2: Single-modality results on UTD-MHAD (LOSO)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
RGB Video	80.7 ± 6.3	10,179 M	39.15	69.5 ± 10.4	2.527 M	9.74
Depth	92.2 ± 4.5	5,391 M	20.74	85.7 ± 4.4	1,514 M	5.83
Skeleton	92.3 ± 2.8	1,098 M	4.2	86.9 ± 4.2	513,213	1.96
Inertial	86.3 ± 7.6	1,132 M	4.33	78.7 ± 4.4	549,325	2.10

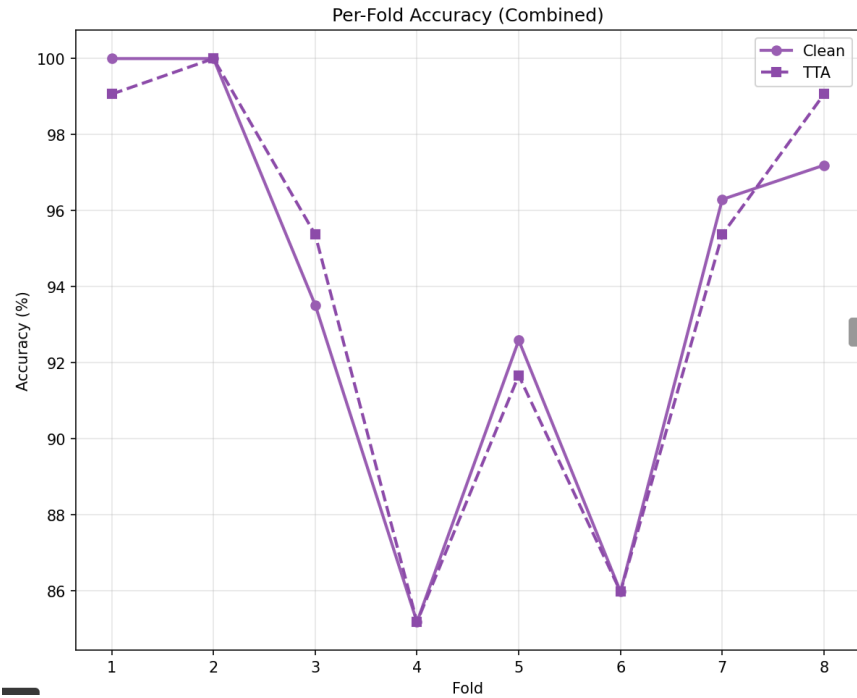


Figure 4-2: UTD-MHAD Proposed Efficient Model Per Fold Accuracy results

Table 4.3: Combined multimodal classification results on UTD-MHAD (Random-Split CV)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	98.8 $\pm$ 0.8 std	14,666,930	56.14
Proposed Lightweight Model	98.7 $\pm$ 1.5 std	3,677,648	14.18

Table 4.4: Single-modality results on UTD-MHAD (Random-Split CV)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
RGB Video	94.2 $\pm$ 2.1	10,179 M	39.15	88.3 $\pm$ 1.6	2.527 M	9.74
Depth	99.2 $\pm$ 1.1	5,391 M	20.74	97.1 $\pm$ 0.9	1,514 M	5.83
Skeleton	97.9 $\pm$ 1.0	1,098 M	4.2	92.8 $\pm$ 2.0	513,213	1.96
Inertial	97.6 $\pm$ 0.7	1,132 M	4.33	90.0 $\pm$ 1.6	549,325	2.10

Besides LOSO, another training method was used: 80-20% split on whole dataset into training and testing and training set is further split into 80-20% parts for training and validation. The cross-validation average across 5 folds was applied to obtain results. This was done to differentiate results and to compare results more consistently with prior work, because the majority of them present their non-LOSO results.

The single modality study reveals that individual backbones of each modality works correctly in activity recognition task. Among the single-modality experiments on UTD-MHAD, the Skeleton and Depth sensor modalities achieves the highest accuracy of 92.3% and 92.2% respectively with the Proposed Efficient Model (the same pattern happening on Lightweight version), suggesting that EfficientNet-B0 and MobileNetV3-Small backbones pretrained on ImageNet1K improves the sole depth accuracy and the inclusion of additional eigen joint-based features of skeleton modality provides the most discriminative features for the activity classes in this dataset and improves the skeleton-only results. Conversely, the RGB video and Inertial modalities yields lower single-modality accuracies of 69.5%-80.7% and 78.7%- 86.3% , indicating

limitation of EfficientNetb2 and MobileNetv3-small models in RGB-only recognition task, compare to more heavy ones: Resnet18 or Resnet50 based results, showed in Table 4-3. The consistent performance gap between the multimodal fusion results (Table 4.1) and the single-modality results confirms the complementary nature of the modalities and validates the fusion strategy employed by the proposed framework. The non-LOSO results on both of my proposed models proved to be more accurate than several of prior work, like Vision-based MMF [22], MAF-Net [30], XTinyHAR [37] and baseline DHCNN-FAM-Mamba model [3], only falling to UCFFormer results [31]: 98.8% and 98.7% on my efficient and lightweight models accordingly, versus 99.01% on their work. Single modality results on random-split CV on UTD-MHAD were consistently very accurate for any provided modality: having all of them accuracy larger than 88.3% and the majority (5/8) combinations having more than 94.2% in accuracy.

The table 4.6 shows the comparison between previous works on action recognition tasks on UTD-MHAD dataset. The combined accuracy is not the strongest in comparison with others, but there was not work previously focused on maintaining modality ablation lightweight model. The SOTA model is UCFFormer, but their results are not maintaining LOSO strategy, thus some data leakage could happen during their training. My testing strategy utilizing cross-validation subject-split LOSO cross validation across all subjects, making it it more strict in terms on final accuracy results. Per-Fold accuracy results are shown in Figure 4-2. These facts makes this proposed work standout from any of prior solutions. The single modality accuracy results are presented in Figure 4-3. Although my proposed approach does not beat UCFFormer and Vision-based Multimodal Framework (Vision-based MMF) combined accuracy results (99.04 % and 98.8 % vs 93.8 % at best), it could be stated that neither of the two approaches actually used all modalities in training and testing. In addition to that, my depth-only results on both configurations are better than Vision-based MMF (92.2% and 85.7% vs 71.91%) , and UCFFormer does not use it all. It should be also noticed from Table 4.6, that other two approaches are heavy ones, they do not include any lightweight-oriented strategies and weighs and number of parameters

more than my proposed solution. The single modality results per-fold of Efficient model and Lightweight model are presented on Figure 4-4, Figure 4-5 respectively.

In Table 4.5, you can see the per-modality feature retention after 50% RFE is applied on the whole combined feature set. Inertial and depth are selected more as individual modalities (71–75% and 56–67%, respectively), while around 20% to 40% of RGB and skeleton features are selected, confirming that the four streams carry complementary rather than redundant information and that the fused accuracy of 93.8% cannot be attributed to any single source.

Table 4.5: Per-modality feature retention after RFE feature selection (50% retention) across two model configurations.

Modality	Efficient (%)	Lightweight (%)
RGB	22.7	39.3
Depth	66.5	55.9
Skeleton	39.4	30.3
Inertial	71.4	74.6

Notably, the Proposed Lightweight Model maintains competitive accuracy in most single-modality settings (sometimes even improving on accuracy) while substantially reducing the parameter count and model size.

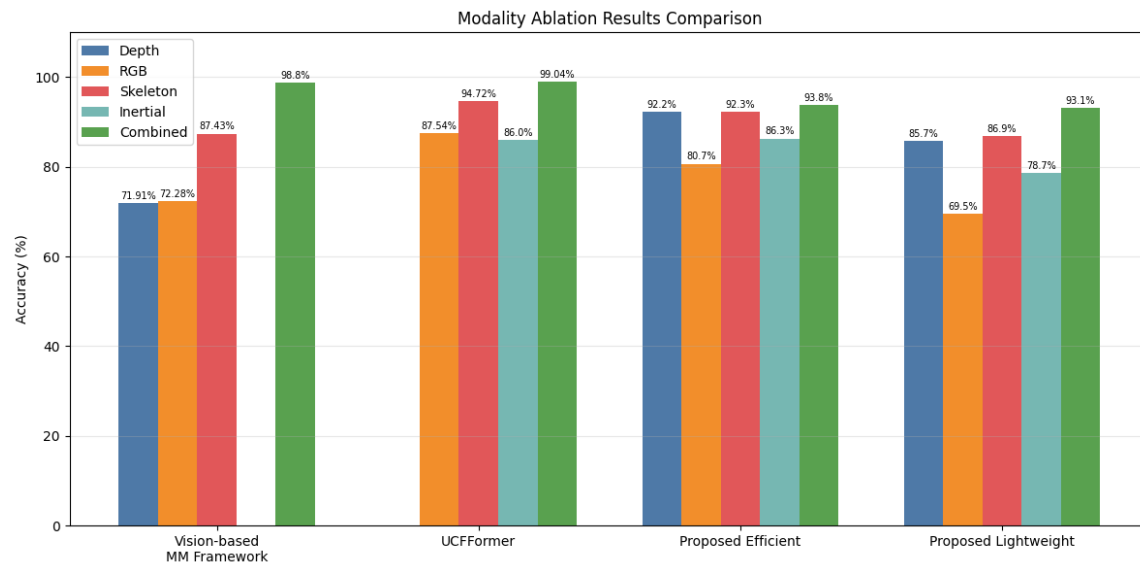


Figure 4-3: Modality Ablation Results on UTD-MHAD

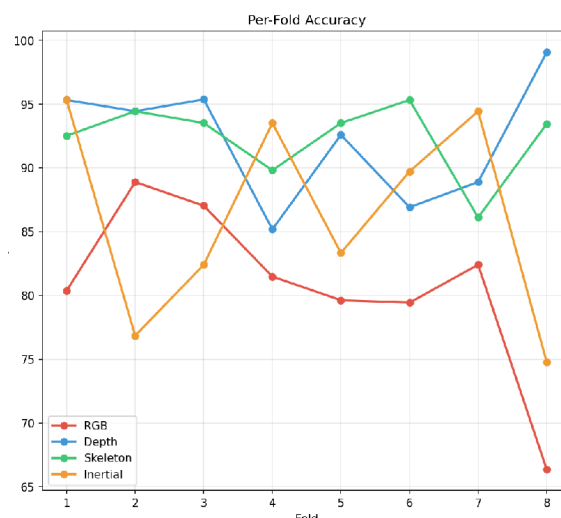


Figure 4-4: UTD-MHAD Single Modality results per fold of Efficient Model

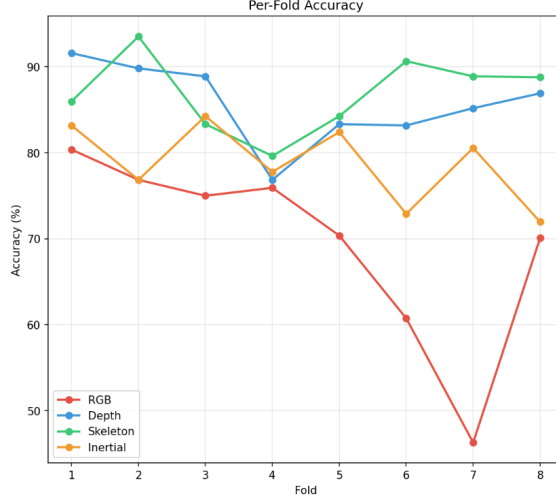


Figure 4-5: UTD-MHAD Single Modality results per fold of Lightweight Model

Table 4.6: Prior work results on UTD-MHAD comparasion

Study	Year	Lightweight	Modality Ablation Tests	Results	# of Params	Model size	Modalities used
Vision-based MM Framework [22]	2022	No	Yes	98.88	ResNet50-based (25.5M+)	–	RGB, Depth, Skeleton
XTinyHAR [14]	2025	Yes	No	98.71	0.62 M	2.45	Inertial, Skeleton
MAF-Net w(Bi-LSTM/ST-GCN)[30]	2025	No	No	91.5 / 92.0	29.4M / 31.1M	–	RGB, Skeleton
DHCNN-FAM-Mamba [3]	2025	Yes	No	96.04	8.08M	–	RGB, Depth, Skeleton
UCFFormer [31]	2025	No	Yes	99.01	33.72M	206.57	RGB, Skeleton, Inertial
Proposed Efficient LOSO CV	2026	Yes	Yes	93.8%	14.66M	56.14	All
Proposed Lightweight LOSO CV	2026	Yes	Yes	93.1%	3.6M	14.18	All
Proposed Efficient Random Split CV	2026	Yes	Yes	98.8%	14.66M	56.14	All
Proposed Lightweight Random Split CV	2026	Yes	Yes	98.7%	3.6M	14.18	All

Table 4.7: Prior works with modality ablation results on UTD-MHAD

Study	Year	Depth	RGB	Skeleton	Inertial	Combined
Vision-based MM Framework [22]	2022	71.91%	72.28%	87.43%	–	98.8%
UCFFormer [31]	2025	–	87.54%	94.72%	86.0%	99.04%
Proposed Efficient model LOSO CV	2026	92.2	80.7%	92.3%	86.3%	93.8%
Proposed Lightweight model LOSO CV	2026	85.7	69.5%	86.9%	78.7%	93.1%
Proposed Efficient model Random Split CV	2026	99.2	94.2%	97.9%	97.6%	98.8%
Proposed Lightweight model Random Split CV	2026	97.1	88.3%	92.8%	90.0%	98.7%

4.2 Results on HWU-USP

This section reports the classification results obtained on the HWU-USP Multimodal Activity Recognition Dataset. Table 4.8 presents the overall performance of both

proposed models when all four modalities are fused, while Table 4.9 details the single-modality ablation results.

Table 4.8: Overall multimodal classification results on HWU-USP (LOSO)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	91.7 $\pm$ 10	14,770,052	56.82
Proposed Lightweight Model	83.3 $\pm$ 10.8	3,772,014	14.53

Table 4.9: Single-modality ablation results on HWU-USP (LOSO)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
RGB	93.1 $\pm$ 4.6	10,177 M	39.14	82.6 $\pm$ 14.6	2,526 M	9.74
Depth	94.4 $\pm$ 4.8	5,389 M	20.73	92.4 $\pm$ 3.9	1,523 M	5.82
Inertial	61.8 $\pm$ 7.6	1,140 M	4.36	68.8 $\pm$ 7.3	571,483	2.18
Ambient	56.2 $\pm$ 9.8	1,157 M	4.43	56.9 $\pm$ 10.3	584,251	2.23

Table 4.10: Overall multimodal classification results on HWU-USP (Random-Split CV)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	92.3 $\pm$ 2.66	14,770,052	56.82
Proposed Lightweight Model	83.3 $\pm$ 2.62	3,772,014	14.53

Table 4.11: Single-modality ablation results on HWU-USP (Random-Split CV)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
RGB	86.13 $\pm$ 4.6	10,177 M	39.14	74.3 $\pm$ 5.7	2,526 M	9.74
Depth	93.05 $\pm$ 4.8	5,389 M	20.73	90.2 $\pm$ 3.5	1,523 M	5.82
Inertial	56.97 $\pm$ 7.6	1,140 M	4.36	53.4 $\pm$ 10.9	571,483	2.18
Ambient	55.52 $\pm$ 9.8	1,157 M	4.43	45.9 $\pm$ 4.8	584,251	2.23

Table 4.12: HWU-USP single modality testing results

Study	Year	Depth	RGB	Inertial	Ambient	Combined
HWU-USP Original Paper [20]	2021	36.36	86.72	65.56	51.39	98.61%
Proposed Efficient LOSO CV	2026	94.4	93.1	61.8	56.2	91.7%
Proposed Lightweight LOSO CV	2026	92.4	82.6	68.8	56.9	83.3%
Proposed Efficient Random Split CV	2026	93.05	86.13	56.97	55.92	92.3%
Proposed Lightweight Random Split CV	2026	90.2	74.3	53.4	45.9	83.3%

Table 4.13: HWU USP: Comparison with state-of-the-art.

Study	Year	Lightweight	Modality-Agnostic	Accuracy	# Params	Size (MB)	Modalities
HWU-USP Original Paper [20]	2021	No	Yes	98.61%	27M+	–	All
CNN-GRU Classifier [4]	2024	No	No	97.9%	–	–	RGB, Depth, Inertial
Biosignal-RGB Fusion [2]	2025	No	No	89.0%	–	–	RGB, Inertial
DHCNN-FAM-Mamba [3]	2025	Yes	No	97.12%	8.08M	–	RGB, Depth, Inertial
Proposed Efficient LOSO CV	2026	Yes	Yes	91.7%	14.7M	56.82	All
Proposed Lightweight LOSO CV	2026	Yes	Yes	83.3%	3.7M	14.53	All
Proposed Efficient Random Split CV	2026	Yes	Yes	92.3%	14.7M	56.82	All
Proposed Lightweight Random Split CV	2026	Yes	Yes	83.3%	3.7M	14.53	All

The poor Inertial results confirm the nature of the data and it’s just the wrong sensor for this specific task. HWU-USP activities are defined by what object you’re interacting with and what environment you’re in, not by how your body moves. For example, there would be almost no difference in inertial measurements when making a sandwich, a cup of tea, or a bowl of cereal. Therefore, relying solely on inertial data in this case is hard to predict activity, because the difference in intensity is minimal. This claim is further strengthened by the Table 4.12, which compares the single modality results of our proposed method to those of the original HWU-USP paper [20]. The random split CV-based results from the same table on single modality underperformed on HWU USP, maybe indicating some mistake on this dataset, but the overall combined accuracy did not fall and at least stayed the same for the lightweight model and slightly increased for the efficient one: 92.3% vs 91.7%.

4.3 Results on MSR Action 3D

This section reports the classification results obtained on the MSR-Action3D Dataset. Table 4.16 presents the overall performance of both proposed models when all modalities are fused, while Table 4.17 details the single-modality results.

Table 4.14: Overall multimodal classification results on MSR Action 3D (LOSO)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	81.9 $\pm$ 11.5	5,395,803	20.75
Proposed Lightweight Model	82.9 $\pm$ 11.3	2,170,743	8.34

Table 4.15: Single-modality ablation results on MSR Action 3D (LOSO)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
Depth	83.9 $\pm$ 12.5	5,342 M	20.55	82.6 $\pm$ 9.2	2,126 M	8.17
Skeleton	84.7 $\pm$ 8.9	1,057 M	4.05	81.8 $\pm$ 12.1	1,048 M	4.01

Table 4.16: Overall multimodal classification results on MSR Action 3D (Random Split CV)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	95.0 $\pm$ 2.11	5,395,803	20.75
Proposed Lightweight Model	94.77 $\pm$ 2.34	2,170,743	8.34

Table 4.17: Single-modality ablation results on MSR Action 3D (Random Split CV)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
Depth	94.09 $\pm$ 3.00	5,342 M	20.55	92.73 $\pm$ 2.34	2,126 M	8.17
Skeleton	95.91 $\pm$ 2.11	1,057 M	4.05	97.5 $\pm$ 1.11	1,048 M	4.01

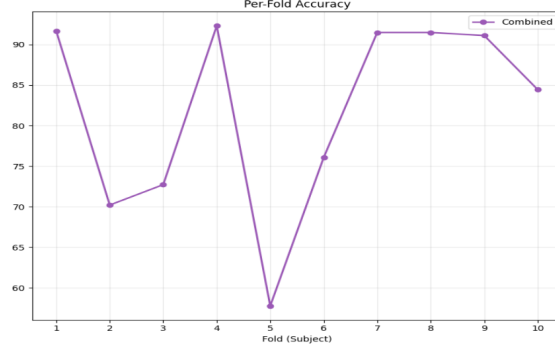


Figure 4-6: Per fold accuracy on MSR Action 3D dataset under LOSO CV

Table 4.18: MSR Action 3D: Comparison with prior work

Study	Year	Lightweight	Modality-Agnostic	Accuracy	# Params	Size (MB)	Modalities
MSR Action 3D Original Paper [16]	2010	Yes	No	74.7%	–	–	All
KNN-based model [29]	2017	Yes	No	92.3%	–	–	All
Proposed Efficient LOSO CV	2026	Yes	Yes	81.9%	5.4M	20.75	All
Proposed Lightweight LOSO CV	2026	Yes	Yes	82.9%	2.2M	8.34	All
Proposed Efficient Random Split CV	2026	Yes	Yes	95%	5.4M	20.75	All
Proposed Lightweight Random Split CV	2026	Yes	Yes	94.7%	2.2M	8.34	All

It can be stated that results on the MSR Action 3D dataset are the least accurate across all datasets used in this study, with only 81.9% and 82.9% on efficient and lightweight model respectively. It could be explained by the high variance of Fold 5 under the LOSO CV protocol, where test subject number 5 is the sole test set, pictured in Figure 4-6. While I have not found the explanation of this outlier, the alternative training approach proved the feasibility of the proposed model. Under Random-Split CV-based results, the model thrived and obtained 95% and 94.77% accuracy for efficient and lightweight model, respectively, as per table 4.16. The table 4.17 proves single modality results are also relatively consistent and achieve high recognition (all 92.73 % and larger), highlighting the skeleton data more, as expected, since MSR Action 3D dataset has 2D and 3D variations of this modality. If we compare my results with prior it could be seen that my LOSO results beat only the original dataset paper, however non-LOSO results are beating both of the comparable works, shown in table 4.18.

4.4 Results on MM-Fit

Lastly, in this section, we have the classification results obtained on MM-Fit Dataset. Table 4.21 shows the overall performance of both proposed models; Table 4.22 presents the single-modality results.

Table 4.19: Overall multimodal classification results on MM-Fit (LOSO)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	98.7 ± 1.9	1,187,687	4.54
Proposed Lightweight Model	98.1 ± 3.0	589,137	2.25

Table 4.20: Single-modality ablation results on MM-Fit (LOSO)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
Inertial	88.1 ± 11.7	1,134 M	4.34	88.0 ± 10.9	0.552 M	2.11
Skeleton	97.6 ± 3.0	1,096 M	4.19	98.6 ± 2.1	0.512 M	1.96

Table 4.21: Overall multimodal classification results on MM-Fit (Random Split CV)

Model	Accuracy (%)	Parameters	Size (MB)
Proposed Efficient Model	99.51 ± 0.40	1,187,687	4.54
Proposed Lightweight Model	99.35 ± 0.32	589,137	2.25

Table 4.22: Single-modality ablation results on MM-Fit (Random Split CV)

Modality	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
Inertial	92.37 ± 1.22	1,134 M	4.34	91.07 ± 2.45	0.552 M	2.11
Skeleton	98.7 ± 1.09	1,096 M	4.19	99.35 ± 0.6	0.512 M	1.96

Table 4.23: MSR Action 3D: Comparison with prior work

Study	Year	Lightweight	Modality-Agnostic	Accuracy	# Params	Size (MB)	Modalities
MM-Fit Original Paper Standard-CNN [27]	2021	No	No	96.37%	–	–	All
XTinyHAR [14]	2025	Yes	No	98.55%	–	–	Inertial
Proposed Efficient LOSO CV	2026	Yes	Yes	98.7%	1.2M	4.54	All
Proposed Lightweight LOSO CV	2026	Yes	Yes	98.1%	589K	2.25	All
Proposed Efficient Random Split CV	2026	Yes	Yes	99.51%	1.2M	4.54	All
Proposed Lightweight Random Split CV	2026	Yes	Yes	99.35%	589K	2.25	All

The MM-Fit results were the most accurate, and it could be explained by relative simplicity if the dataset and my proposed model overperformed the original dataset paper, as seen in table 4.23. However, the non-LOSO results outperformed both similar studies.

In general, the major improvement across all datasets in both combined and single modalities configurations has been recorded for the Random Split-based CV results compared to LOSO CV, as expected, since the data leakage of all subjects, which does not happen in LOSO, tends to improve results.

Table 4.24 consolidates the multimodal fusion results across all four benchmark datasets.

Table 4.24: Summary of multimodal combined results across all datasets.

Dataset	Proposed Efficient Model			Proposed Lightweight Model		
	Acc.(%)	Params	Size(MB)	Acc.(%)	Params	Size(MB)
UTD-MHAD	93.8	14.6M	56.43	93.1	3.6M	14.18
HWU-USP	91.7	14.7M	56.82	83.3	3.7M	14.53
MSR Action3D	81.9	5.4M	20.75	82.9	2.2M	8.34
MM-Fit	98.7	1.2M	4.54	98.1	0.6M	2.25

The Proposed Efficient Model consistently achieves the highest accuracy on each dataset, with an exception of MSR Action 3D, while the Proposed Lightweight Model offers a favourable trade-off between accuracy and computational cost. Across all benchmarks, the lightweight variant reduces the number of parameters by approximately 70% on average, with an average accuracy decrease of only 2.89 percentage

points.

4.5 Results without RFE

All previous experiments were conducted with RFE (50% feature retention) on the whole feature pool. For comparison purposes, in this section, we present the main results without RFE, on all features without selection: Table 4.25 summarizes them. As we can see, in 4 out of 8 situations, the introduction of RFE actually improved the classification accuracy, showing that there might have been some feature redundancy in the original concatenated vector. While it does not prove that the RFE is universal method, leading to overall improvement, it should be noted that introduction of RFE in the model was not the main goal, and Also MSR Action 3D results are an outlier in our case, with bad LOSO results due to extra low performance in only fold, so their non-improvement using RFE should be taken with this details in mind.

Table 4.25: Model performance comparison with and without RFE

Dataset	Efficient		Lightweight	
	RFE	No RFE	RFE	No RFE
UTD-MHAD	93.8	94.4	93.1	92.0
HWU-USP	91.7	87.5	83.3	88.2
MSR Action 3D	81.9	87.5	82.9	83.3
MM-Fit	98.7	98.1	98.1	97.6

4.6 Inference Efficiency Analysis

To evaluate the practical deployability of the proposed models, this section reports inference timing results measured on NVIDIA GeForce GTX 1080 Ti GPU with 11.7 GB of RAM and edge device: Raspberry Pi 4 Model B with 8 GB RAM. For each computer, a subset of 10 test samples is processed over the same number of folds as in training, and the average per-sample inference time is computed. The total inference pipeline is decomposed into two stages:

1. **Feature preprocessing and extraction:** all modality-specific transformations and feature extraction prior to classification.
2. **Classification:** the forward pass through the fusion and classification layers.

Tables 4.26- 4.29 summarise the inference timing results for all four datasets. It could be seen that lightweight model is always delivers faster results with GPU results faster than Raspberry Pi 4 ones in 23.64 times on average on Efficient and 17.61 times on Lightweight model. The difference between inference times, while being the same model lies in the dataset utilization: the amount of modalities in UTD-MHAD and HWU-USP (4 in total) involves more backbones to be working, while the 2 modality-datasets: MSR Action 3D and MM-fit utilizes only 2 backbones, which makes their computation faster. MSR Action 3D is little bit slower due to the fact that depth backbone is being used there, which involves computation of depth motion maps on many depth images, compared to MM-fit deals with only skeleton and inertial modalities: easiest ones in terms of data preprocessing computational cost. They demonstrate that the Proposed Lightweight Model achieves a notable reduction in total per-sample inference time compared to the full Proposed Model.

Table 4.26: Average per-sample inference time (in ms) on UTD-MHAD dataset

Model	Computer	Preproc. (ms)	Classif. (ms)	Total (ms)
Efficient	GPU	24.58	95.6	120.8
	Raspberry Pi 4B	562.36	57.0	619.37
Lightweight	GPU	15.45	69.41	84.86
	Raspberry Pi 4B	123.5	41.4	164.9

Table 4.27: Average per-sample inference time (in ms) on HWU-USP dataset

Model	Computer	Preproc. (ms)	Classif. (ms)	Total (ms)
Efficient	GPU	102.63	1.151	103.7
	Raspberry Pi 4B	1071.99	4.071	1076.07
Lightweight	GPU	48.40	1.03	49.43
	Raspberry Pi 4B	980.44	3.161	983.6

Table 4.28: Average per-sample inference time (in ms) on MM-Fit dataset

Model	Computer	Preproc. (ms)	Classif. (ms)	Total (ms)
Efficient	GPU	5.78	0.06	5.84
	Raspberry Pi 4B	29.129	136.114	165.243
Lightweight	GPU	4.21	0.06	4.27
	Raspberry Pi 4B	25.886	87.321	113.207

Table 4.29: Average per-sample inference time (in ms) on MSR Action 3D dataset

Model	Computer	Preproc. (ms)	Classif. (ms)	Total (ms)
Efficient	GPU	14.828	10.754	25.582
	Raspberry Pi 4B	1147.427	151.5	1298.987
Lightweight	GPU	9.374	8.868	18.241
	Raspberry Pi 4B	255.042	147.847	402.889

4.7 Summary

4.8 Statistical significance test

The null hypothesis states that there is no statistically significant difference in performance between the Efficient and Lightweight models. To prove that, paired t-tests

and Wilcoxon tests were performed on each dataset’s Efficient vs Lightweight per-fold results. If the value of $p < 0.05$, it would make the difference significant. Both a parametric (paired t-test) and non-parametric (Wilcoxon) test were used to ensure robustness, as the Wilcoxon test makes no assumption about the normality of the fold-level score distributions.

Table 4.30: Statistical significance test

	Paired t-test	Wilcoxon test
UTD-MHAD	0.7416	0.6406
HWU-USP	0.0559	0.1250
MSR Action 3D	0.7794	0.9102
MM-Fit	0.4650	0.6250

According to results from table 4.30 all four datasets p-value is less than 0.05, which proves that the difference between Efficient and Lightweight is statistically not significant. These results support the use of the Lightweight model as a viable alternative, achieving comparable performance with reduced computational cost.

4.9 Summary

This chapter presented a comprehensive evaluation of the proposed multimodal activity recognition framework across four benchmark datasets: UTD-MHAD, HWU-USP, MSR Action 3D, and MM-Fit. The key findings are as follows:

1. Single Modality testing results are good (9/12 modalities across 4 datasets on Efficient model achieves $> 83.9\%$ accuracy) and other low performance is the result of HWU-USP nature of activities.
2. Multimodal fusion consistently outperforms any single modality in isolation, confirming the value of integrating complementary data sources.
3. The Proposed Efficient Model achieves the highest accuracy on 3/4 datasets, demonstrating the effectiveness of the designed architecture.

4. The Proposed Lightweight Model provides a strong accuracy–efficiency trade-off, reducing model size and parameter count in 70% while maintaining competitive performance with only 2.025% accuracy decrease on average.
5. The inference analysis reveals that both proposed variants are suitable for near-real-time deployment on edge devices.

These results collectively support the viability of the proposed approach, the robustness to missing modalities with competitive results on single modalities testing.

This work presents several contributions to the field of multimodal human activity recognition. Inspired by the DHCNN-FAM-Mamba architecture of Avinash et al. [3], modality-agnostic fusion framework was proposed that extends the original design through per-modality temporal modeling prior to any cross-modal interaction. Unlike their work, who concatenate all modality features into a unified representation before applying the core processing blocks, our architecture assigns each modality a dedicated specialized backbone — EfficientNet-B2 for RGB appearance, MobileNetV3-Small for optical flow, a three-layer Graph Convolutional Network for skeleton, and a DHCNN-augmented Bidirectional LSTM for inertial and ambient signals — allowing each modality to develop its own temporal representation independently before fusion. The simpler backbone choices and embedding size are used in the lightweight version, which has been statistically proven to be insignificant and therefore, feasible. This design enables true modality agnosticism, whereby the model handles any given modality or combination of modalities without architectural modification or retraining. Furthermore, two architectural additions absent from prior work were introduced. First, a RFE stage using GradientBoosting selects the most discriminative 50% of projected features per modality at training time, reducing noise in the feature space before fusion without adding parameters to the deployed model. Second, a dedicated Cross-Modal Mamba Fusion block enables structured information exchange between modalities through a principled cross-input construction, in contrast to the simple concatenation strategy in [3]. RGB processing was enhanced by incorporating person-aware preprocessing through YOLOv8n bounding box detection and dense

Farneback optical flow computation, providing the visual backbone with spatially focused appearance features and explicit motion representations that jointly encode both what the subject looks like and how they are moving. Finally, all experiments are conducted under the Leave-One-Subject-Out cross-validation protocol across four benchmark datasets, evaluating generalization to entirely unseen subjects. This evaluation strategy is stricter than the random split protocols used by Yang et al. [31] and Avinash et al. [3], making the obtained reported results more conservative but more reflective of real-world deployment conditions.

Chapter 5

Conclusion

The results demonstrate that the modality-robust model for multimodal HAR can exist and be deployable on the edge device. Even if the results are less accurate than SOTA papers, it could be justified by the more harsh testing strategy: LOSO with k-fold cross-validation and low memory size of the model, low inference time and smaller number of parameters used. The results of 93.8% and 93.1 % on UTD-MHAD on Efficient and Lightweight version of proposed model demonstrate the effectiveness of the approach. The single-modality results, on average, achieve approximately 85% and 82% per modality on UTD-MHAD, proving that the model is modality-agnostic and robust to missing modalities due to its backbone-based architecture. The total per-sample inference times on Raspberry Pi4 show that the time lies in near-real-time response and that it is suitable for edge deployment, making it a lightweight-friendly solution. The results on HWU-USP datasets are beating the baseline (HWU-USP original paper) in terms of single modality results (ambient, inertial, depth) and model size and number of parameters (14.7 M vs 27 M+). The 3D results of MSR Action (81.9% in efficient versions and 82.9% in lightweight) are consistent; the performance of a single modality is also consistently good (each modality achieving more than 81% in isolation), especially considering the large number of subjects (20). Performance on MM-Fit is also remarkably good (98.7% and 98.1% on Efficient and Lightweight) on both combined and single modality experiments. If we take a look at non-LOSO (random-split based CV results), the recognition there is also highly accurate in both

efficient and lightweight models: for UTD-MHAD

We have tested the proposed framework on 4 different datasets with differing activity natures: activities of daily living on UTD-MHAD and MSR Action3D, kitchen activities on HWU-USP, and physical exercises on MM-Fit. This also proves the generalizable and adaptable nature of our proposal to different HAR domains.

Future work may involve evaluating the model’s feasibility on larger datasets, exploring alternative feature selection methods beyond RFE, and experimenting with different fusion strategies. One limitation of the proposed model is its inability to correctly assign a bounding box to a person performing an action when two individuals are present in the camera frame. A potential solution would be to select the largest bounding box as the primary source of action recognition, which remains an area to be addressed in future iterations of the model.

Bibliography

- [1] Yawar Abbas and Ahmad Jalal. Drone-based human action recognition for surveillance: a multi-feature approach. In *2024 International Conference on Engineering & Computing Technologies (ICECT)*, pages 1–6. IEEE, 2024.
- [2] Iqra Aijaz Abro, Haifa F Alhasson, Shuaa S Alharbi, Mohammed Alatiyyah, Dina Abdulaziz Alhammadi, Ahmad Jalal, and Hui Liu. Advanced biosignal-rgb fusion with adaptive neurofuzzy classification for high-precision action recognition. *Ieee Access*, 2025.
- [3] Baruri Sai Avinash, Sakshi Thakur, and Binod Kumar. Multimodalmamba: An efficient dhcn-fam-mamba architecture for human activity recognition. *IEEE Sensors Journal*, 2026.
- [4] Mouazma Batool, Moneerah Alotaibi, Sultan Refa Alotaibi, Dina Abdulaziz Alhammadi, Muhammad Asif Jamal, Ahmad Jalal, and Bumshik Lee. Multimodal human action recognition framework using an improved cnngnu classifier. *IEEE Access*, 12:158388–158406, 2024.
- [5] Sizhen Bian, Mengxi Liu, Vitor Fortes Rey, Daniel Geissler, and Paul Lukowicz. Tinierhar: Towards ultra-lightweight deep learning models for efficient human activity recognition on edge devices. In *Proceedings of the 2025 ACM International Symposium on Wearable Computers*, pages 163–169, 2025.
- [6] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015.
- [7] Xue-wen Chen and Jong Cheol Jeong. Enhanced recursive feature elimination. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, pages 429–435. IEEE, 2007.
- [8] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.
- [9] Qisheng He, Nicholas Summerfield, Ming Dong, and Carri Glide-Hurst. Modality-agnostic learning for medical image segmentation using multi-modality

- self-distillation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.
- [10] Md Amran Hossen and Pg Emeroylariffion Abas. Machine learning for human activity recognition: State-of-the-art techniques and emerging trends. *Journal of Imaging*, 11(3):91, 2025.
 - [11] Md Mofijul Islam and Tariq Iqbal. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10285–10292. IEEE, 2020.
 - [12] Md Mofijul Islam and Tariq Iqbal. Mumu: Cooperative multitask learning-based guided multimodal fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:1043–1051, 06 2022.
 - [13] Samee Ullah Khan, Maryam Sultana, Sufyan Danish, Deepak Gupta, Norah Saleh Alghamdi, Suchang Woo, Dong-Gyu Lee, and Sangtae Ahn. Multi-modal feature fusion for human activity recognition using human centric temporal transformer. *Engineering Applications of Artificial Intelligence*, 160:111844, 2025.
 - [14] Ismail Lamaakal, Chaymae Yahyati, Yassine Maleh, Khalid El Makkaoui, Ibrahim Ouahbi, Ahmed A Abd El-Latif, Mariam Zomorodi, and Basma Abd El-Rahiem. A tiny inertial transformer for human activity recognition via multimodal knowledge distillation and explainable ai. *Scientific Reports*, 15(1):42335, 2025.
 - [15] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey: S. li, h. tang. *International Journal of Computer Vision*, 134(3):103, 2026.
 - [16] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010.
 - [17] Rex Liu and Xin Liu. Mu-mae: Multimodal masked autoencoders-based one-shot learning. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 253–259. IEEE, 2024.
 - [18] Shenghuan Miao, Ling Chen, Rong Hu, and Yingsong Luo. Towards a dynamic inter-sensor correlations learning framework for multi-sensor-based wearable human activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), September 2022.
 - [19] Takuya Nakabayashi and Hideo Saito. Multimodal human activity recognition on edge devices. *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 136–140, 2024.

- [20] Caetano M Ranieri, Scott MacLeod, Mauro Dragone, Patricia A Vargas, and Roseli AF Romero. Human activities with videos, inertial units and ambient sensors. *Dryad Digital Repository*, 2021.
- [21] Dustakar Surendra Rao, L Koteswara Rao, Vipparthi Bhagyaraju, and Goh Kam Meng. Enhanced depth motion maps for improved human action recognition from depth action sequences. *Traitement du Signal*, 41(3), 2024.
- [22] Beddiar Djamila Romaisa, Oussalah Mourad, and Nini Brahim. Vision-based multi-modal framework for action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5859–5866. IEEE, 2021.
- [23] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. In *German conference on pattern recognition*, pages 281–297. Springer, 2018.
- [24] Fatemeh Shafizadegan, Ahmad R Naghsh-Nilchi, and Elham Shabaninia. Multimodal vision-based human action recognition using deep learning: a review. *Artificial Intelligence Review*, 57(7):178, 2024.
- [25] Muhammad Bilal Shaikh, Douglas Chai, Syed Muhammad Shamsul Islam, and Naveed Akhtar. From cnns to transformers in multimodal human action recognition: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(8):1–24, 2024.
- [26] Jungpil Shin, Najmul Hassan, Abu Saleh Musa Miah, and Satoshi Nishimura. A comprehensive methodological survey of human activity recognition across diverse data modalities. *Sensors*, 25(13):4028, 2025.
- [27] David Strömbäck, Sangxia Huang, and Valentin Radu. Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–22, 2020.
- [28] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.
- [29] Qingqiang Wu, Guanghua Xu, Longting Chen, Ailing Luo, and Sicong Zhang. Human action recognition based on kinematic similarity in real time. *PloS one*, 12(10):e0185719, 2017.
- [30] Dongwei Xie, Xiaodan Zhang, Xiang Gao, Hu Zhao, and Dongyang Du. Maf-net: A multimodal data fusion approach for human action recognition. *PloS one*, 20(4):e0319656, 2025.

- [31] Kyoung Ok Yang, Junho Koh, and Jun Won Choi. Ucfformer: Recognizing human actions from multimodal sensors using unified contrastive fusion transformer. *Neurocomputing*, page 131374, 2025.
- [32] Tianjin Yang, Zhenjie Hou, Jiuzhen Liang, Yuwan Gu, and Xin Chao. Depth sequential information entropy maps and multi-label subspace learning for human action recognition. *IEEE Access*, 8:135118–135130, 2020.
- [33] Xiaodong Yang and YingLi Tian. Effective 3d action recognition using eigen-joints. *Journal of Visual Communication and Image Representation*, 25(1):2–11, 2014.
- [34] Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3063–3071, 2022.
- [35] Muhammad Yaseen. What is yolov8: an in-depth exploration of the internal features of the next-generation object detector (2024). *Accessed: Sep, 10, 2025*.
- [36] Xianfa Zhou, Jiabin Yuan, Lili Fan, Xuewei Niu, Keke Zha, and Xu Liu. Msmft: Multi-stream multimodal factorized transformer for human activity recognition. *IEEE Sensors Journal*, 25(6):10402–10416, 2025.
- [37] Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. Tinyhar: A lightweight deep learning model designed for human activity recognition. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, pages 89–93, 2022.