Assignment 2 – Multimodal Sensor Fusion with Attention

Ritwika Das (rd935)

November 10, 2025

1 Introduction

Multimodal sensor fusion is important in human activity recognition, where multiple wearable sensors (accelerometers, gyroscopes, heart rate) are combined to capture complex motion patterns. A single sensor can be noisy or incomplete, whereas complementary signals improve accuracy and robustness, which is useful for health and fitness applications where sensors may drop out.

We use the **PAMAP2** dataset, which records 18 activities with three IMUs (hand, chest, ankle) and a heart rate sensor. PAMAP2 is realistic but challenging because sensors can be missing, sampling rates differ, and deep models can become overconfident. This motivates comparing **Early**, **Late**, and **Hybrid (attention-based)** fusion, and adding **uncertainty** measurements to make sure predictions are calibrated.

2 Approach

Architecture Overview

We implement three pipelines: Early Fusion (concatenate encoder outputs → MLP), Late Fusion (per-modality classifier → learned weighting), and Hybrid Fusion (project to common space → cross-modal attention → adaptive fusion → classifier).

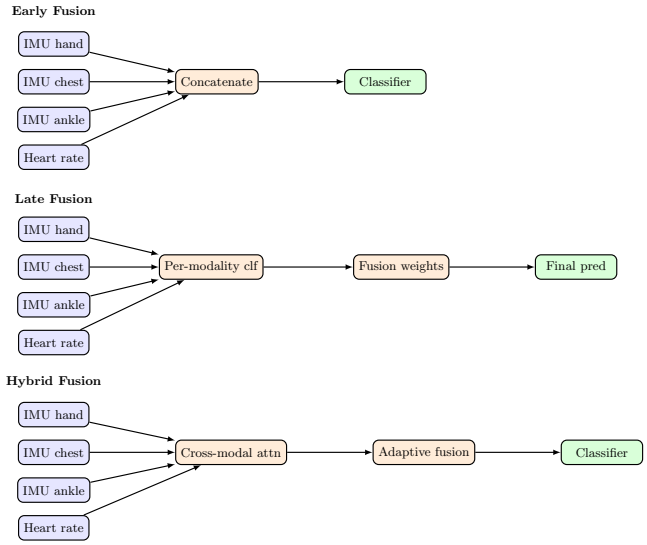

Figure 1: Overview of the three fusion pipelines: Early (feature-level), Late (decision-level), and Hybrid (attention-based).

Attention Mechanism

For query modality A attending to modality B:

$$\text{Attn}(A \to B) = \text{softmax}\left(\frac{Q_A K_B^\top}{\sqrt{d_k}}\right) V_B, \tag{1}$$

with $Q_A = W_Q q_A$, $K_B = W_K k_B$, $V_B = W_V v_B$. We apply this pairwise and fuse with learned weights conditioned on the modality mask so missing sensors get weight 0.

Temporal Alignment

Because PAMAP2 sensors have different sampling rates, we resample/ interpolate to a fixed window for each modality (IMUs padded/truncated, heart rate upsampled). This keeps the model simple and makes fusion possible.

Uncertainty Quantification

We report Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \big|\text{acc}(S_b) - \text{conf}(S_b)\big|. \tag{2}$$

This tells us how close confidence is to true accuracy.

3 Experiments & Results

Fusion Strategy Comparison

The fusion plot (`fusion_comparison.png`) shows all three strategies are very close (within 1%): Late = 0.932, Early = 0.931, Hybrid = 0.926. This is likely because preprocessing (resampling + normalization) made the modal features highly correlated, so attention had little extra structure to learn.

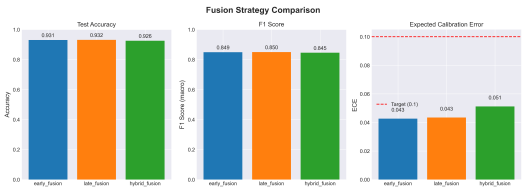

Figure 2: Accuracy/F1/ECE for Early, Late, Hybrid fusion.

Missing-Modality Robustness

The missing-modality plot (`missing_modality.png`) shows graceful degradation: accuracy rises from $\sim$0.28 (1 sensor) to $\sim$0.92 (4 sensors). The chest IMU alone reaches $\sim$0.89, so it is the most informative. Similar preprocessing across IMUs made them partially redundant, so losing one did not break the model.

Figure 3: Performance vs. number of available sensors.

Attention Visualization

The attention heatmap (`attention_viz.png`) shows IMUs attending to each other (esp. chest $\leftrightarrow$ ankle), and very little attention to heart rate. So the Hybrid model learned sensible motion-based dependencies, even if the gain over Early/Late was small.

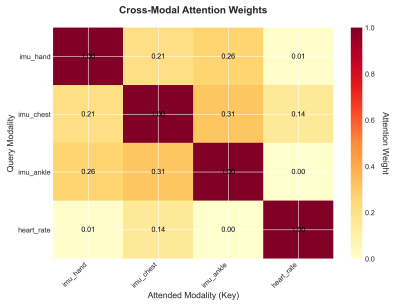

Figure 4: Cross-modal attention learned by Hybrid fusion.

Calibration

The reliability diagram (`calibration.png`) shows ECE ≈ 0.05 and curves close to the diagonal. All three fusion types are similarly calibrated, which again suggests the shared preprocessing and training setup stabilized them.

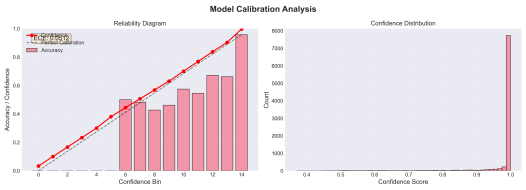

Figure 5: Calibration curve (confidence vs. accuracy).

Ablation

Attention heads. The plot (`attn_heads_accuracy.png`) shows 1 head is best; more heads slightly hurt accuracy. This means PAMAP2's cross-modal structure is simple.

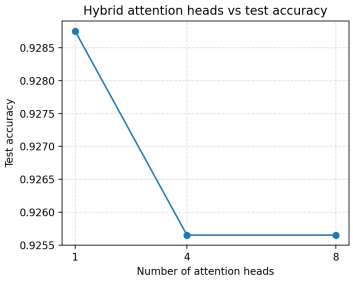

Figure 6: Accuracy for Hybrid with different attention heads.

Attention vs. Concatenation Early (0.931) > Hybrid (0.926), so attention added interpretability but not performance, likely because preprocessing homogenized the features.

Baseline Comparison

Model	Acc.	F1	ECE
Single IMU-hand	0.928	0.838	0.057
Early (concat)	0.931	0.849	0.043
Hybrid attention	0.926	0.845	0.051

Table 1: Baseline vs. multimodal models.

Even though differences are small, Early Fusion improves over the single-modality baseline and achieves better calibration, confirming that combining modalities is useful. Hybrid stays competitive but does not surpass Early/Late, likely because preprocessing produced highly redundant modality embeddings.

4 Analysis & Discussion

Late Fusion was the strongest overall, probably because it combines per-modality decisions and avoids overfitting to shared features. Hybrid did not improve much because preprocessing made the modalities too similar, so attention had little extra signal. Missing-modality tests showed graceful degradation and confirmed the chest IMU dominates. All models were well-calibrated (ECE $\approx$ 0.04–0.05). Attention maps were sensible (IMU-to-IMU, HR mostly ignored), so attention is still useful for interpretability. For deployment, Late Fusion is the safest choice under this preprocessing.

5 Conclusion

This work compared Early, Late, and Hybrid fusion on PAMAP2. All achieved high accuracy and low ECE, with Late Fusion slightly best. Early is fine when all sensors are present; Late is better for real-world settings with possible dropouts; Hybrid is useful when modalities are more heterogeneous or attention interpretability is needed. Future work: relax preprocessing to preserve modality differences, add temporal cross-attention, and explore uncertainty-aware fusion.