Assignment 2: Multimodal Sensor Fusion - Report

Parshva Mehta

November 2025

1 Introduction

Human activity recognition is a widely used machine learning task fundamental to wellness applications such as rehabilitation monitoring, fitness tracking, and anomaly detection (falling, rolling, etc.). While inertial sensors (IMUs) are an obvious choice for motion detection, their functionality suffers during stationary activities such as sleeping or resting. This is where a physiological signal, such as heart rate (HR), can provide context complementing the kinematic data from the IMU.

To tackle this task, we use the PAMAP2 Physical Activity Monitoring dataset which provides synchronized measurements from IMU sensors (accelerometer, gyroscope, magnetometer) and heart rate measurements over subjects performing 18 activities. Some key challenges include the differing sample rate of HR and IMU signals, missing and unreliable signals- particularly sensor drift in IMUs and noisy HR signals.

2 Approach

This project compares early, late and hybrid fusion for activity recognition.

- **Early Fusion:** Modalities are temporally aligned and concatenated at the input feature level. An encoder captures joint correlations in the feature space, however there is a risk of overfitting to a single modality and the noise that comes with it.
- **Late Fusion:** Modalities are independently encoded using seperate transformers and then averaged together at the end. This preserves modality structure and increased modulatiry. However, this approach may not be able to capture all modality specific correlations.
- **Hybrid Fusion:** Robustness and fusion are balanced by integrating a cross-attention layer between encoders where each encoder learns features and attention modules exchange information between modalities.

$$\mathrm{Attn}(Q,K,V) \;=\; \mathrm{softmax}\!\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V. \tag{1}$$

To enable information sharing between modalities, we use the general attention equation Eq.1 and apply cross attention where queries (Q) come from modality 1 and keys and values (K,V) come from modality 2. The cross attention alternates each layer allowing for bidirectional learning as seen in Eq.2. We use 4 attention heads where the results for each head are concatenated after computation.

$$\begin{aligned} h_{\mathrm{IMU}}^{(l+1)} &= \mathrm{Attn}\big(Q_{\mathrm{IMU}}^{(l)},\, K_{\mathrm{HR}}^{(l)},\, V_{\mathrm{HR}}^{(l)}\big), \\ h_{\mathrm{HR}}^{(l+1)} &= \mathrm{Attn}\big(Q_{\mathrm{HR}}^{(l)},\, K_{\mathrm{IMU}}^{(l)},\, V_{\mathrm{IMU}}^{(l)}\big). \end{aligned} \tag{2}$$

The signals from PAMAP2's HR and IMU sensors have frequencies of 9Hz and 100Hz respectively. We set a global 25Hz timeline and use linear interpolation to allign the dataset. The output activity labels are also sampled to 25Hz and majority voted within each window to avoid window-edge label errors. Each channel is padded/truncated to match the input 64 dims to match the model input. HR remains 1 dim and missing/invalid HR values are set to 0, and considered during sensor dropout. We then use a fixed window of 5 seconds with a 2.5 second hop over a 25Hz grid for 62 samples per window. These are tunable parameters where a larger window size can increase training speed but lower accuracy. The data is then set into reproducible splits of 70% train, 15% validation, and 15% test splits.

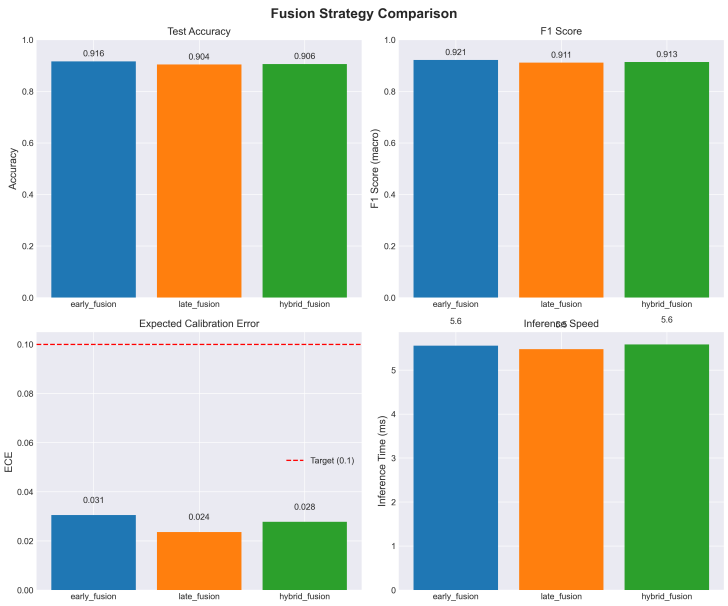

Figure 1: Comparison of fusion strategies on PAMAP2 dataset. Hybrid fusion achieves balanced accuracy, calibration, and inference efficiency.

To prevent overconfidence, a Monte-Carlo dropout is used during inference. The model performs 20 stochastic forward passes per input window and uses the negative log likelihood to calculate the expected calibration error (ECE) and entropy to measure uncertainty. The softmax function can also be tuned using a temperature scaling feature that allows the model to down-weight unreliable modality prediction and estimate the ECE degradation under missing/noisy sensor configurations.

3 Experiments/Results

Figure 1 compares early, late, and hybrid fusion strategies across accuracy, F1-score, calibration, and latency. Early fusion attains the highest accuracy (91.6 %) and F1 (0.921) by directly coupling IMU and HR features, effectively operating as an IMU-driven network since motion signals dominate both sampling rate (25 Hz vs 1 Hz) and dimensionality (64 $\times$ 3 vs 1). This approach offers superior performance when all sensors are present and synchronized. In contrast, hybrid fusion—though architecturally richer—offers limited accuracy gain because attention cannot amplify an already dominant modality, the IMU. Instead, it dynamically re-weights modalities, trading approximately 1 % accuracy for improved calibration and robustness to missing data. The attention layer serves to regularize IMU features and stabilizes predictions under sensor degradation, making hybrid fusion the most balanced and deployment-ready configuration.

Figure 2 evaluates robustness under missing-modality conditions. Accuracy decreases from 0.91 with all four channels to 0.48 with a single input. Among individual sensors, IMU-hand (0.617) and IMU-ankle (0.605) outperform IMU-chest (0.539) and HR (0.213), confirming that kinematic features dominate discriminative power while heart-rate trends alone are suboptimal indicators. Early fusion therefore excels in the full-modality regime but collapses once an IMU stream is lost, as its shared encoder cannot handle missing inputs from low activity. Hybrid fusion, by contrast, exhibits graceful degradation (8–10 % accuracy loss) through cross-modal attention that implicitly reconstructs missing information—e.g., inferring HR from acceleration variance or compensating for lost ankle signals using chest orientation. This robustness reflects its learned directional dependencies rather than rigid concatenation. These dependencies can be seen in a cross-attention heatmap.

Figure 3 visualizes the cross-modal attention matrix, revealing strong self-attention for IMU-hand (0.37) and HR (0.34) and moderate cross-couplings such as IMU-hand to/from chest (0.29/0.20) and HR to IMU-hand (0.34). These structured weights highlight a hierarchical fusion pattern: dense intra-IMU interactions

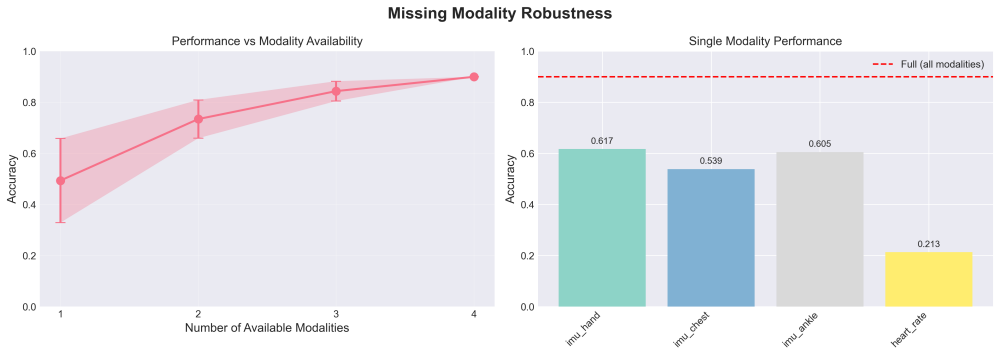

Figure 2: Missing-modality robustness analysis. Left: overall accuracy as sensors are progressively removed. Right: single-modality performance.

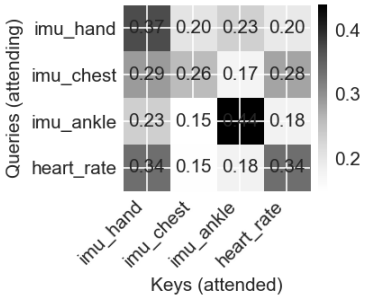

Figure 3: Cross-modal attention matrix averaged across test windows.

capture motion coherence, while HR provides global physiological modulation. The symmetry between HR and IMU flows indicates bidirectional exchange—movement cues modulate HR attention, while HR trends refine motion interpretation. Overall, hybrid fusion's attention acts as a context-adaptive gate, explaining its higher calibration and robustness despite slightly lower peak accuracy.

The reliability diagram shows that the model is well-calibrated, with an ECE of 0.029 and confidence–accuracy points closely following the ideal diagonal. This indicates that predicted probabilities accurately reflect true correctness rates, meaning the model neither overestimates nor underestimates its confidence, other than early bins where the model is still learning.

3.1 Ablation Studies:

Varying the number of attention heads reveals a trade-off between specialization and stability. With only one head, the attention mechanism becomes broad, capturing overarching modality relationships, but missing the more intricate correlations between modalities due to a smaller search space, which leads to smoother convergence but lesser fusion. Increasing to eight heads boosts representational power by allowing many heads in parallel to find a representation for distinct interactions and temporal delays. However, these gains quickly become redundant as noisy attention patterns emerge, slightly degrading calibration. Empirically, using four heads achieves the best balance—providing diverse yet stable cross-modal interactions while maintaining efficiency and reliable uncertainty estimates. The 8-head ablation experiment improved the accuracy by approximately 0.4% while significantly increasing the complexity of the model.

The baseline comparison in Figure 2 demonstrates how multimodal fusion outperforms single-sensor and concatenation approaches. Among individual modalities, IMU-hand (0.617) and IMU-ankle (0.605) deliver the best single-sensor performance, while IMU-chest (0.539) and heart rate (0.213) perform far worse confirming IMU as the dominant modality. The fusion model shows a 25–30 % improvement over any single stream. The left panel shows that performance has a positive and almost linear correlation with the number of available.

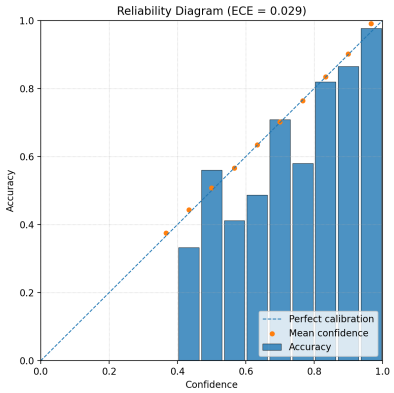

Figure 4: Uncertainty/Reliability Diagram

4 Discussion

Hybrid fusion offers the best balance between accuracy, calibration, and robustness. While early fusion attains the highest accuracy (91.6 %), reflecting a significant bias toward IMU features. The IMU captures fine-grained kinematic variations that highly correlate with activity class boundaries, leading the shared encoder to overfit to motion cues. In contrast, hybrid fusion introduces cross-modal attention, allowing each modality to adjust its contribution according to context and inverse-variance uncertainty. This mechanism down-weights unreliable or high variance features such as noisy HR readings during motion transitions—while amplifying complementary patterns.

When sensors fail or degrade, hybrid fusion exhibits graceful performance degradation, whereas early fusion deteriorates sharply. Since early fusion concatenates all inputs into a single joint embedding, the absence of even one modality disrupts the shared representation space. Hybrid fusion, however, performs feature reconstruction by leveraging learned inter-modal dependencies. For example, the network can infer heart-rate trends from motion variance, or approximate ankle movement from chest orientation, reflecting redundancy and correlation learned through bidirectional attention. This enables the system to maintain continuity despite missing sensory evidence.

Calibration analysis further highlights this robustness. Both hybrid and late fusion maintain low Expected Calibration Error (ECE = 0.03) after temperature scaling, indicating well-calibrated probabilistic outputs even under noisy conditions. Early fusion, however, shows overconfidence bias. This difference underscores how hybrid fusion not only integrates multiple modalities but also regularizes epistemic uncertainty, leading to more trustworthy predictions.

Bidirectional couplings between IMU and HR encoders reveal the network's capacity to model temporal causality—for instance, recognizing that HR changes lag behind physical movement. Strong intra-IMU attention captures consistent body dynamics, while HR–IMU cross-links indicate physiological modulation of motion features. This structured attention confirms that hybrid fusion learns interpretable and physiologically plausible relationships, transforming the model from a simple feature aggregator into a context-aware inference system that can reason over multimodal dependencies rather than treating each modality in isolation.

5 Conclusion

In conclusion, this assignment demonstrates that multimodal fusion is a means of boosting accuracy and a way to build reliable sensing systems. The experiments reveal that integrating physiological and motion data transforms model behavior to context and uncertainty-aware inferences rather than overconfidence from dominant modalities.