Assignment A2 Report

1. Introduction

This assignment requires implementation of encoders, cross-attentions, fusion strategies, and uncertainty metrics. Uncertainty aware fusion is also implemented as an extra credit.

1.1 Dataset

PAMAP2 is chosen for fast training (light weight sensor data).

It's processed with data_process.py with the following settings:

```
SPLITS = {"train": 0.7, "val": 0.15, "test": 0.15}
SEQ_LEN = 100      # sequence length (in samples)
STRIDE = 50        # sliding window stride (in samples)
```

We take sequence as input since the encoder is LSTM.

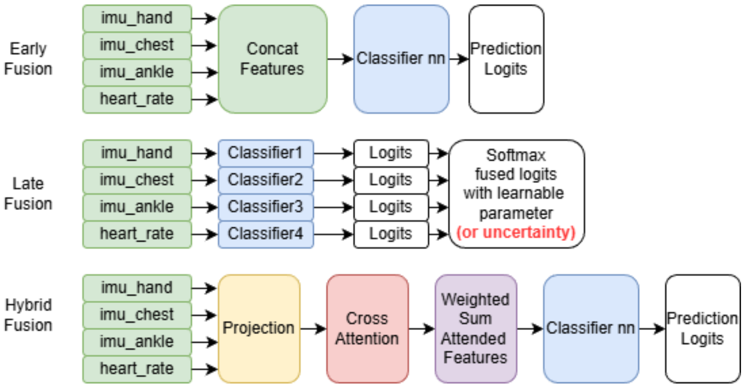
1.2 Key challenges

This task takes data from multiple sensors as input (i.e., multiple modalities), and the key challenge lies in effectively fusing them. In addition, input samples may contain NaNs due to missing modalities or unaligned sampling rates, which lead to naive models ineffectiveness.

2. Approach

2.1 Architecture diagram

We implemented three strategies: early, late and hybrid fusion. Note that uncertainty aware fusion is considered as a variant of late fusion approaches.



2.2 Cross Attention Module: Key Steps

step1 Give query features $Q \in \mathbb{R}^{B \times d_q}$, key features $K \in \mathbb{R}^{B \times d_k}$ and value features $V \in \mathbb{R}^{B \times d_k}$, we first project them into a shared hidden space:

$$Q' = QW_Q, \quad K' = KW_K, \quad V' = VW_V$$

step2 The hidden space is divided into multiple heads: $d_{hid} = h \times d_{head}$

For each head, compute attention score which measures the similarity between modality A query and keys from modality B:

$$S = \frac{Q'K'^{\top}}{\sqrt{d_{head}}}$$

step3 Apply mask if provided: $S_{masked} = S \odot M$

step4 Apply softmax to obtain attention weights: $A = \text{softmax}(S)$

step5 Weighted sum over values: $Z_h = AV'$

step6 Concatenate and then output Z:

$Z = \text{Concat}(Z_1, Z_2, \ldots, Z_h)W_O$

2.3 Temporal alignment

The dataset PAMAP2 has plenty of NANs due to this sampling rate issue. We tested two approaches:

Down sampling: simply delete all samples with any NANs, which makes the amount of data becomes too small.

Adaptive Modality Input (selected): Each encoder independently handles input sequences containing NaNs. For example, if the IMU ankle sensor has missing values for certain samples, columns 38–54 in the input sequence will be NaN. The encoder replaces these NaNs with zeros, producing feature representations with zero values for the missing data. The feature inputs passed to the fusion module are thus already processed by the encoders, with zeros indicating missing modalities.

That is, mask in fusion modules was implemented but not used. We simply handle temporal alignment or missing modality in encoder stage. The fusion modules will receive zero values when sensor is down.

2.4 Uncertainty quantification

We used temperature scaling class that learns a single scalar parameter T. The calibrate method applies the calibration by returning scaled logits, then apply softmax to obtain calibrated probabilities.

3. Experiment & Results

3.1 Fusion Comparison Table (with full modality)

We run this experiment when all modalities are available (though in samples some modality is None because of sampling rate difference) to show the performance when all sensors are working.

Fusion Strategy	Accuracy	F1_macro	Expected Calibration Error
Early	0.9550523161888123	0.9502402394254089	0.013009258545935154
Hybrid	**0.9612283706665039**	**0.9586350083845425**	0.017492420971393585
Late	0.953336775302887	0.9502622720343585	**0.012846475467085838**

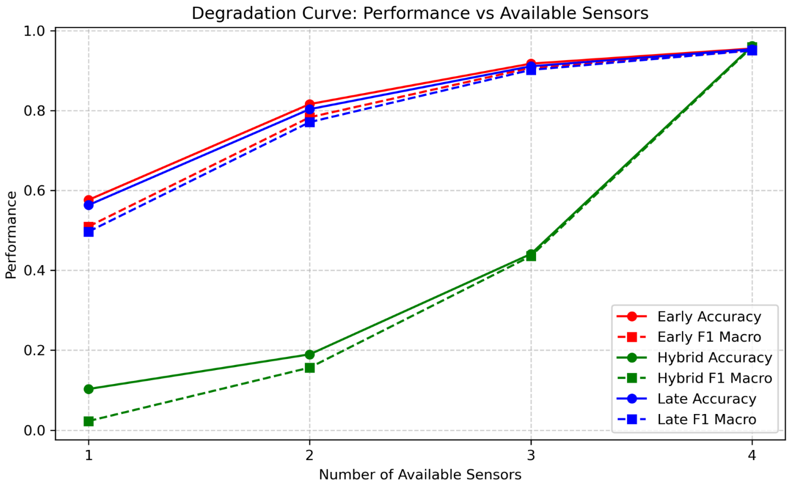
This result shows that Hybrid fusion strategy has the best accuracy and F1 score, though it has a higher ECE, implying it's not calibrated.

3.2 Missing Modality Plot

3.2.1 Degradation Curve

In order to examine the robustness of our model, we test the model on test set with all combinations of available sensors to mock real deployment where sensor could fail. We calculate the average across the combinations with the same number of available sensors. Our result shows that hybrid fusion strategy is less robust to sensor failure. This is likely due to the cross-attention mechanism used in hybrid fusion: when one modality fails, the effective information from cross-attention decreases quadratically. However, the reason could also be some bug in my hybrid fusion implementation.

For 1 sensor fail, we achieved 3.7% and 4.3% accuracy drop in average, when using early and late fusion strategy. For 3 sensor fail, the average accuracy drop is 37.9% and 39.0% when using early and late fusion strategy. This accuracy could also be interpreted as the accuracy when using one modality.



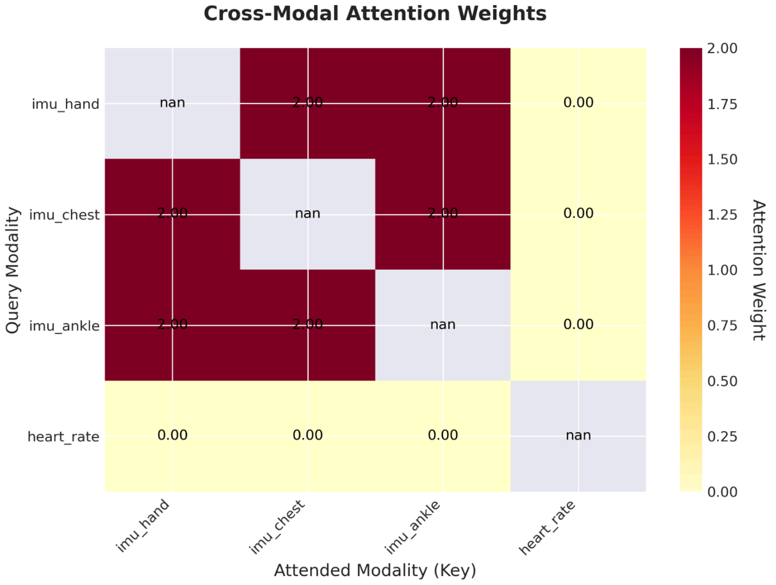
3.2.2 Modality Importance

	imu_hand	imu_chest	imu_ankle	heart_rate
Early Fusion	0.25309911220353265	0.3550396621992872	0.2510661224310637	-0.1407951031661165
Late Fusion	0.2415349277312563	0.34711272964225326	0.27462644786972445	-0.13672589475676597
Hybrid Fusion	0.26564155719111043	0.3651644422422404	0.3330887602843422	-0.03610524028230713

Our results shows for all fusion strategies, the chest sensor is the most important, and the heart rate is less important. It's notable that the variance of importance is smaller in hybrid fusion, with higher importance score in heart_rate and imu_ankle. This is likely due to the cross-attention mechanism, which takes multiple modalities as its input.

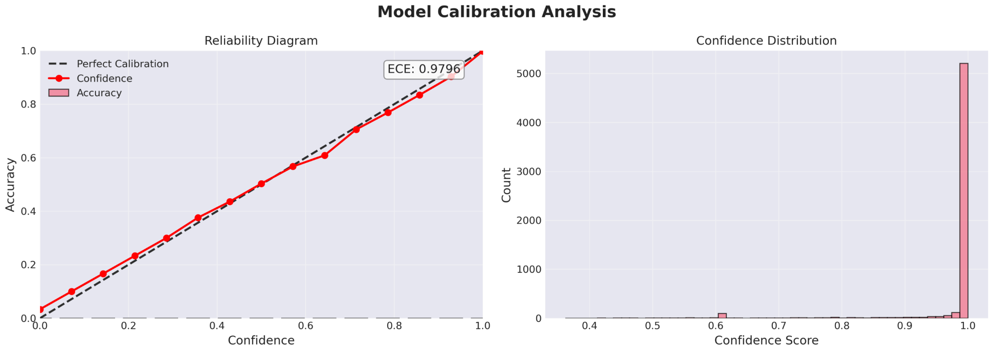
3.3 Attention Visualization

The following attention visualization figure shows the relationship between modalities. From this figure, we could know that heart rate is not an important modality in our cross-attention, which is consistent with our previous conclusion from importance score.



3.4 Calibration Diagram

Although the model’s confidence curve almost perfectly aligns with the ideal calibration line (Perfect Calibration), indicating an overall well-calibrated trend, the ECE value remains relatively high. This discrepancy mainly arises from the highly imbalanced confidence distribution—most samples have prediction confidences concentrated near 1.0, as shown on the right. Since ECE is computed as a weighted average over all bins, even small deviations in the dominant high-confidence region can substantially inflate the overall value. Therefore, while the model appears visually well-calibrated, the numerical metric is elevated due to the uneven confidence distribution.



4. Analysis & Discussion

Which fusion strategy wins?

The early fusion strategy wins. It has slightly lower accuracy vs hybrid with all modalities, but it maintain the accuracy when the sensor fails. This is likely because the pamap2 is a relatively simple and small dataset. The difference in modality is not as significant as other multimodal datasets with image, video, text and audio. The dataset itself only contains numbers, which implies already encoded. My intuition is that a simple MLP with proper hyper-parameter tuning strategy and large enough number of parameters could easily handle this task.

How does performance degrade?

For early and late fusion, it's graceful. For hybrid fusion, it's catastrophic. The reason is similar to the reason in 4.1: hybrid fusion is not a must in such task, and won't help increasing the robustness of the system.

Is the model well-calibrated?

Yes. We have a ECE from 0.012 to 0.017, implying it's well-calibrated.

What do attention patterns reveal?

Heart rate modality is not important. This is also revealed by the importance score.

Limitations

The hybrid fusion didn't work. However, our early and late fusion strategy shows good enough performance for this task, which makes it not meaningful to further develop the hybrid fusion model. Our system is already capable of predicting the motion status, even challenged with sensor failures in real environment.

5. Conclusion

In this assignment, we implemented early, hybrid and late fusion model, and the encoders/attention needed for these fusion models. Our result shows early and late fusion is good enough to handle the pamap2 task. To the best of our knowledge, for this pamap2 task, using early or late fusion strategy is already good enough, while using hybrid fusion only increase the complexity of system without improving the real performance.

We believe further works should focus on scaling laws or data quality for the motion prediction task. The scale of model and data is way more important than the complex model design.