This article has been accepted for publication in IEEE Embedded Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LES.2023.3298738



# DynaFuse: Dynamic Fusion for Resource Efficient Multi-modal Machine Learning Inference

Hamidreza Alikhani[1], Anil Kanduri[2], Pasi Liljeberg[2], Amir M. Rahmani[1], Nikil Dutt[1] [1] *Dept. of CS, University of California, Irvine, USA,* [2] *Dept. of Computing, University of Turku, Finland*
hamidra@uci.edu, spakan@utu.fi, pakrili@utu.fi, a.rahmani@uci.edu, dutt@uci.edu

***Abstract*—Multi-modal machine learning (MMML) applications combine results from different modalities in the inference phase to improve prediction accuracy. Existing MMML fusion strategies use static modality weight assignment, based on the intrinsic value of sensor modalities determined during the training phase. However, input data perturbations in practical scenarios affect the intrinsic value of modalities in the inference phase, lowering prediction accuracy, and draining computational and energy resources. In this work, we present DynaFuse, a framework for dynamic and adaptive fusion of MMML inference to set modality weights, considering run-time parameters of input data quality and sensor energy budgets. We determine the insightfulness of modalities by combining design-time intrinsic value with the run-time extrinsic value of different modalities to assign updated modality weights, catering to both accuracy requirements and energy conservation demands. The DynaFuse approach achieves up to 22% gain in prediction accuracy and an average energy savings of 34% on exemplary MMML applications of human activity recognition and stress monitoring in comparison with state-of-the-art static fusion approaches.**

***Index Terms*—Multi-modal machine learning, energy efficiency, run-time systems**

## I. Introduction

Smart eHealth applications such as human activity recognition [1], pain and stress monitoring [2] use multi-modal machine learning (MMML) algorithms for combining supplementary and complementary information across heterogeneous sensor modalities [3]. MMML fusion methods such as early (e.g., feature aggregation) and late fusion (e.g., priority voting) consider the relative *intrinsic value* i.e., the significance of a modality in contributing towards prediction accuracy, to prioritize among different modalities [4]. On the other hand, multi-modal sensing in practical settings is often prone to input data perturbations with different noise components, motion artifacts, and missing data due to battery drain and physical failure of sensory devices. Input data perturbations affects prediction accuracy of MMML models, and drain computational and energy resources by processing data in a garbage-in garbage-out fashion [5].

Existing MMML methods determine modality weights based on their intrinsic value in the training phase, and use these fixed values for fusion in the inference phase [6], [7]. Figure 1 shows the workflow of training and inference using static and dynamic modality weight assignment for fusion. As shown in Figure 1 (a), MMML models are trained over pre-processed quality data under ideal scenarios. However, inference in real-time faces external challenges such as perturbed input data and battery drain of sensory modalities

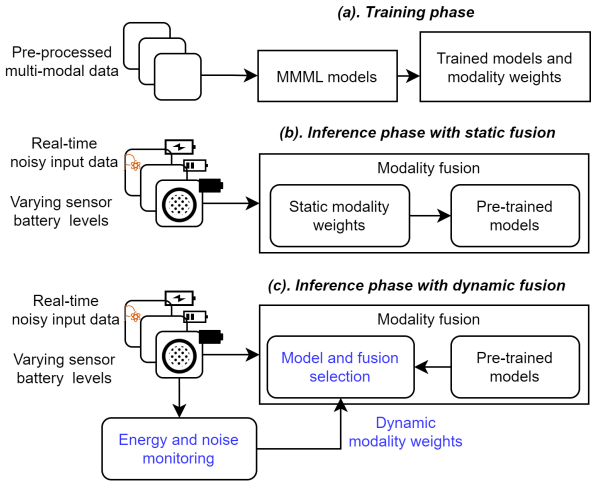


Fig. 1. Static versus dynamic fusion in handling run-time system parameters

(as shown in Figure 1 (b)), affecting the intrinsic value of modalities. Yet, state-of-the-art MMML inference strategies fuse modalities using fixed modality weights that are set at design time without considering external factors such as input data perturbations in practical scenarios [6]. This approach fails to prioritize modalities appropriately based on run-time variable dynamics, potentially degrading prediction accuracy and resource utilization. In practical scenarios, the intrinsic value of a modality is compounded at run-time with the *extrinsic value*, which includes factors such as input data perturbations and available battery levels. As illustrated in Figure 1 (c), considering input data quality and available energy budgets of sensor modalities enables re-prioritizing among modalities for fusion to improve prediction accuracy and minimize energy consumption.

In this work, we present DynaFuse, a framework for dynamic fusion of MMML inference by considering run-time parameters of input data quality and sensor energy budgets. We determine the insightfulness of modalities by combining the design-time intrinsic value with run-time extrinsic value of different modalities to assign updated modality weights. DynaFuse adapts the choices of modality weight assignment and model selection to cater to both accuracy requirements and energy conservation demands. We select appropriate models from a pre-trained model pool and adaptively fuse the results from different modalities using the updated priorities for resource-efficient inference, achieving an acceptable prediction

This article has been accepted for publication in IEEE Embedded Systems Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LES.2023.3298738



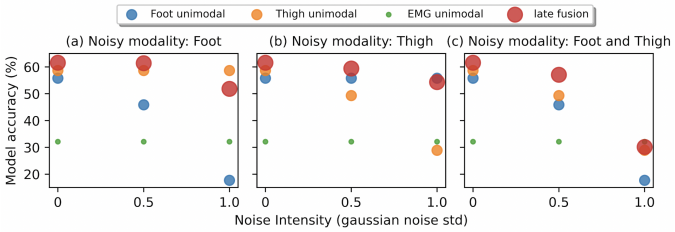


Fig. 2. Noise effect on uni-modal and multi-modal models' accuracy in Stress level detection and HAR applications.

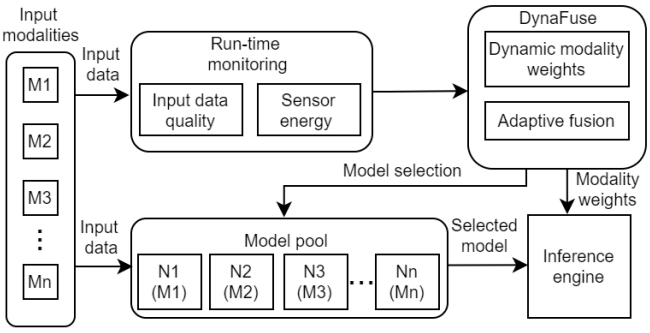


Fig. 3. Overview of the DynaFuse framework

accuracy within the lowest energy consumption. We present the relevant background and motivation for dynamic modality weight assignment in Section 2, the proposed framework in Section 3, an evaluation of our proposed approach in Section 4, followed by conclusions in Section 5.

## II. Motivation

We demonstrate the disparity in prediction accuracy between ideal and practical scenarios for an exemplary MMML application of Human Activity Recognition (HAR). The HAR application uses data from accelerometer/gyroscope (both foot and thigh) and Electromyogram (EMG) modalities to detect activity of the subject. Figure 2 illustrates the overall prediction accuracy when results from different modalities are combined under varying noise levels. We create different scenarios where input data from one or more modalities is noisy. In each scenario, we injected Gaussian noise with varied standard deviations (0 - no noise, 0.5- noise level-1, 1 - noise level-2) to demonstrate the effect of different noise levels on prediction accuracy. Figure 2 also shows the relative total sensing and compute energy consumption of each model, represented by the circle marker size. Figure 2 (a) shows the prediction accuracy of uni-modal models foot and thigh (around 55%) and EMG (33%), and the late fusion model (around 60%) with no noise. The accuracy of the foot uni-modal model drops to 45% at noise level-1 and to 20% at noise level-2. However, existing MMML fusion strategies do not consider such discrepancies and continue to use the static modality weights set under ideal conditions. This affects the prediction accuracy of the late fusion model, which drops to 50% at noise level-2, since the late fusion model still uses modality weights from ideal conditions. In Figure 2 (c), where both foot and thigh modalities are noisy, the EMG uni-modal becomes the most accurate model at noise level 2, completely contrasting the modality weight that would have been assigned under the no-noise scenario. Also, EMG consumes significantly lower energy than the other models, reinforcing the assignment of a higher priority. On the other hand, the foot and thigh uni-modal models prompt the need for lowering their corresponding modality weights, which were assigned under no-noise conditions.

Existing MMML fusion strategies do not consider run-time variable system parameters and rely on static modality weight assignment, degrading the prediction accuracy [7]. Other approaches that consider input data quality use a binary weight assignment to drop the entire noisy modality(s) [2], limiting the utility of multi-modal inputs. Further, none of the existing MMML strategies consider energy conservation while ensuring a certain level of prediction accuracy under noisy input data scenarios. We address the aforementioned limitations through dynamic and adaptive fusion using qualitative run-time re-prioritization of insightful modalities for resource-efficient inference.

## III. Dynamic Adaptive Fusion

Our proposed dynamic fusion framework (DynaFuse) targets resource efficient MMML inference, combining design time intrinsic and run-time extrinsic values of different modalities, and varying accuracy and energy demands. Figure 3 shows the end-to-end overview of DynaFuse with input data originating from different modalities, run-time monitoring, dynamic fusion module, model pool and inference engine. The framework is built on top of the *model pool* – which consists of pre-trained uni-modal models that are used to execute inference tasks. The *inference engine* executes inference tasks on input data from different modalities ($M1 - M_n$), fusing results from relevant uni-modal models from the model pool. The core DynaFuse components include *run-time monitoring* and *dynamic fusion modules* to monitor run-time parameters and make dynamic fusion decisions. The *run-time monitoring* monitors (i) input data quality to identify the noise level across modalities in terms of Signal to Noise Ratio (SNR) and motion artifacts (based on the methodology proposed in [2]), and (ii) energy consumption of different modalities. The *run-time monitoring* enables estimation of extrinsic value of different modalities in practical scenarios, which will be used to update the modality weights. The *dynamic fusion* module sets modality weights and selects appropriate models based on run-time monitored parameters by combining the intrinsic and extrinsic value of different modalities. The *inference engine* executes the selected model (uni-modal/multi-modal) with modality weights assigned by the dynamic fusion module.

**Dynamic fusion** We implement dynamic modality weight assignment in two phases viz., *resilience estimation* – to evaluate the affect of different noise levels on accuracy of uni-modal models, and *energy demand estimation* – to understand the energy demands for exploring accuracy-energy trade-offs. *Resilience estimation* We perform a profiling to assess the impact of different noise intensities on the accuracy of uni-modal models. During this phase, we collect noise-accuracy pairs

across different level and types of noise. The results obtained from the profiling are utilized to estimate the degradation in prediction accuracy of the models on real-time noisy input data streams. This estimation phase employs a linear function that incorporates the noise-accuracy pairs collected during the profiling phase. We assign initial normalized insightfulness weights ($w_i^{init}$) to the uni-modal models based on the estimated accuracies, as shown in Equation 1.

*Energy demand estimation* We update the normalized insightfulness weights (assigned in the resilience estimation phase) considering the energy demands. We define three modes of operation viz., *high accuracy*, *nominal*, and *energy-saving*. The *high accuracy* mode demands achieving highest possible prediction accuracy, the *nominal* mode requires achieving highest possible prediction accuracy within lowest possible energy consumption, while the *energy-saving* mode demands lowest possible energy consumption with an acceptable prediction accuracy gained by dynamic modality weight assignment. The corresponding changes in modality weights are implemented within the boundaries set by these modes. We define a configurable minimum accuracy threshold as $0.8\times$ of maximum achievable prediction accuracy under ideal scenarios. In high-accuracy mode, the model (either uni-modal or late fusion multi-modal with dynamic weights) with the highest estimated accuracy is chosen for inference. In the nominal modes, uni-modal model that suffices the accuracy threshold requirement is selected, while in the energy-saving mode, uni-modal model with the lowest energy consumption that meets the minimum accuracy threshold is selected. Selecting uni-modal models (in nominal and energy-saving modes) aligns with the modality weights assignment, where the weight for the non-selected models is set to zero. In certain scenarios where no uni-modal model satisfies the minimum accuracy requirement, an additional phase is implemented to enhance control over energy consumption. The *run-time monitoring* module dynamically updates inherent sensing energy consumption of different modalities based on their sampling rate. In nominal and energy-saving modes, the *top-k* uni-modal models (sorted based on energy consumption) are selected for the inference phase, while the weights of the remaining uni-modal models are set to zero. The value of *k* is set to 1 in energy-saving mode, $0.5\times$ *number of modalities* in nominal mode, and *number of modalities* in high accuracy mode. Selecting or ignoring uni-modal models for inference is done by masking $w_i^{init}$ with a binary value $b_i$ for each modality to get the final weight ($w_i^{final}$), as shown in Equation 1. Overall, we enhance the dynamic modality weight assignment by prioritizing energy efficiency while simultaneously meeting accuracy requirements.

$$w_i^{init} = \frac{acc_i^{est}}{\sum_{j=1}^{N} acc_j^{est}}, \quad w_i^{final} = w_i \cdot b_i, (1 \leq i \leq N) \quad (1)$$

## IV. Evaluation

**Workloads** We evaluate the efficacy of our proposed framework in terms of prediction accuracy and energy conservation, in comparison with the baseline strategy of static weight assignment and model selection [6]. For evaluation, we use two exemplary MMML applications viz., human activity recognition (HAR) and stress monitoring. For the HAR application, we use HuGaDB dataset [8] to train uni-modal models based on ideal input data scenarios. We use input data from accelerometer and gyroscope sensors attached to foot and thigh, and EMG sensors attached to front thighs. Neural models are used for training uni-modal models for each modality (two convolutional layers, followed by two dense layers). Input samples are processed with window size of four seconds and overlap of two seconds to extract temporal properties of the HAR application. For the stress monitoring application, we trained uni-modal models for PPG, EDA, and ECG modalities. WESAD [9] dataset was used for training the mentioned models. The features for PPG, EDA, and ECG modalities were extracted using tools provided in [10]. We extracted the *sensing energy consumption* of different modalities through empirical models (presented in [2]) based on the sampling rate and number of input channels of the modalities, and *compute energy consumption* based on the DL model size, and finally normalized the total energy consumption of modalities to enable re-prioritization of weights.

**Experimental Scenarios** For evaluating efficiency of DynaFuse, we define scenarios with varying noise levels amongst different modalities, and system-level demands (high-accuracy, nominal, and energy-saving modes). We inject different levels of noise in the HAR application by varying the standard deviation of Gaussian noise randomly ranging between 0-2. For the stress monitoring application, we used four noise levels of Motion Artifacts (no noise, 10%, 30%, and 50%). These setups simulate realistic noise intensities varying uni-modal models' accuracy from ideal to totally random classification during inference. Previous works [11], [2] have used these types of noise to demonstrate the effect of input data perturbation on ML models' accuracy. We experimented with 40 batches of input data, each comprising random levels of input noise, and system demands. Each batch consists of 300 samples in the HAR application and 100 samples in the stress monitoring application, where the batch size is empirically determined to emulate realistic noisy input behavior. Across the 40 batches, we randomly generated objectives such that 50% of the batches demand high accuracy, 30% demand nominal performance, and 20% demand energy-saving.

**Results** We evaluate our proposed DynaFuse approach against static modality weight assignment strategy [7] [2] for HAR and stress monitoring applications under different noisy conditions and accuracy and energy demands. Figure 4 shows modality weights that are dynamically assigned by the proposed DynaFuse strategy under varying run-time parameters across 40 batches of inputs, for HAR and stress monitoring applications. Also, we show the static weights that are assigned for each of the modalities at the design time. Our proposed DynaFuse approach updates modality weights by adapting to varying input data quality and accuracy and energy demands across different input batches. In contrast, the static modality weight assignment strategies continue to use the fixed modality weights, irrespective of input data quality and energy demands across different batches. It should be noted that the disparity between static and dynamic weights in stress monitoring application

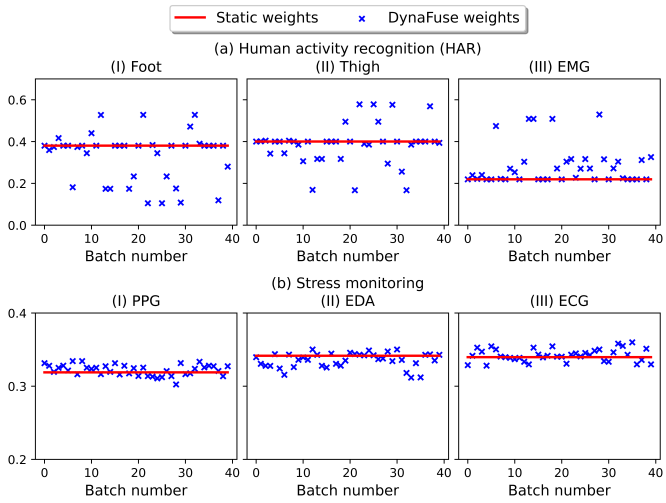


Fig. 4. Modality weights for HAR and Stress monitoring.

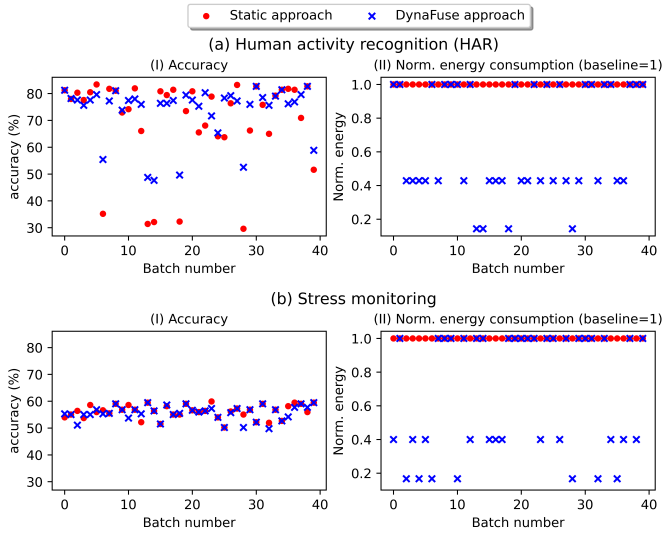


Fig. 5. Prediction accuracy and energy consumption for HAR and Stress monitoring.

is relatively lower in comparison with the HAR application. This can be attributed to the deviation among different input modalities in terms of intrinsic value. For instance, prediction accuracies of the uni-modal models (PPG, EDA, ECG) in the stress monitoring are relatively similar and inherently exhibit high degree of resilience to noise. Thus, the extent of accuracy gains using dynamic fusion depends on correlation between different input modalities of an application, although the energy savings depend on sensor properties.

The adaptivity of DynaFuse approach in updating modality weights is reflected in maximizing the prediction accuracy while minimizing the energy consumption. Figure 5 shows per-batch average prediction accuracy across 40 batches of inputs and corresponding energy consumption per-batch. In certain input batches, DynaFuse has either higher or similar prediction accuracy as the static modality weight assignment strategies with the HAR appliation. However, prediction accuracy of the static approach is better than DynaFuse in specific cases. This is due to the fact that DynaFuse approach deliberately sets modality weights and selects a fusion model that consumes low energy but with relatively lower prediction accuracy to meet the accuracy and energy demands. DynaFuse approach combines run-time input data quality with demands on accuracy and energy to select the most suitable fusion model, exploiting accuracy-energy trade-offs to ensure resource efficient MMML inference. For the HAR application, DynaFuse achieves upto 22% gain in prediction accuracy and average energy savings of 34% in comparison with the baseline. For the stress monitoring application, DynaFuse achieves upto 2.6% gain in prediction accuracy and an average energy savings of 33% in comparison with the baseline. DynaFuse approach has a significantly lower energy consumption in specific cases with demands on energy saving. The energy consumption of DynaFuse is on par with the static assignment in cases where the objective is to maximize the accuracy.

## V. Conclusions

We presented DynaFuse framework for resource efficient MMML inference through adaptive dynamic fusion of modalities, considering run-time input data quality and energy saving demands. We have evaluated our approach on human activity recognition and stress monitoring applicatios over different scenarios of variable input data quality, and accuracy and energy demands. The DynaFuse approach has upto 22% (HAR) gain in prediction accuracy and average energy savings of 34% (HAR) in comparison with state-of-the-art static modality fusion approaches. An intelligent reinforcement learning agent to contextualize system dynamics for model selection and modality weight assignment is planned for the future work.

## VI. Acknowledgements

This work was partially supported by NSF Smart and Connected Communities (S&CC) grant CNS-1831918, Nokia Foundation, and Kaute Saatio.

## References

[1] Zeeshan A and Naimul M K. Human action recognition using deep multilevel multimodal (m2) fusion of depth and inertial sensors. *IEEE Sensors Journal*, 20(1):1445–1455, jan 2020.

[2] E. K Naeini et al. Amser: Adaptive multimodal sensing for energy efficient and resilient ehealth systems. In *DATE*, pages 1455–1460, 2022.

[3] T Baltrusaitis, C Ahuja, and L-P Morency. Multimodal machine learning: A survey and taxonomy. 41(2):423–443, feb 2019.

[4] Chiori Hori et al. Attention-based multimodal fusion for video description. In *2017 IEEE ICCV*, pages 4203–4212, 2017.

[5] Mengmeng Ma et al. Smil: Multimodal learning with severely missing modality. In *AAAI Conference on Artificial Intelligence*, 2021.

[6] H Yu, T Vaessen, I Myin-Germeys, and A Sano. Modality fusion network and personalized attention in momentary stress detection in the wild. In *ACII*, pages 1–8. IEEE, 2021.

[7] H Yang et al. More to less (m2l): Enhanced health recognition in the wild with reduced modality of wearable sensors. In *Ann. Int. Conf. of IEEE EMBC*, pages 3253–3256, 2022.

[8] Roman C and Attila K-F. Hugadb: Human gait database for activity recognition from wearable inertial sensor networks. In *Int. Joint Conf. on the Analysis of Images, Social Networks and Texts*, 2017.

[9] Pothers Schmidt. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proc. of ACM Int. Conf. on Multimodal Interaction*, page 400–408, 2018.

[10] P van Gent, Haneen F, N Nes, and B van Arem. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. 2018.

[11] Rohan Banerjee et al. Noise cleaning and gaussian modeling of smart phone photoplethysmogram to improve blood pressure estimation. In *2015 IEEE ICASSP*, pages 967–971. IEEE, 2015.