ISCA: Intelligent Sense-Compute Adaptive Co-optimization of Multimodal Machine Learning Kernels for Resilient mHealth Services on Wearables

Hamidreza Alikhani[1], Anil Kanduri[2], Emad Kasaeyan Naeini[1], Sina Shahhosseini[1], Pasi Liljeberg[2], Amir M. Rahmani[1], and 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, ekasaeya@uci.edu, sshahhos@uci.edu, pasi.liljeberg@utu.fi, a.rahmani@uci.edu, dutt@uci.edu

***Abstract*— mHealth services use multi-modal machine learning (MMML) models to process physiological and contextual data for automated decision making. Run-time input data perturbations degrade the prediction accuracy of MMML models, while continuous sensing, transmission, and processing of such noisy data drains the energy resources of wearable devices. Identifying qualitative input data and dropping non-insightful modalities can improve prediction accuracy and energy efficiency simultaneously. We propose a ISCA: a sense-compute adaptive co-optimization framework that employs reinforcement learning to jointly determine sensing and compute configuration settings which minimizes energy consumption while providing accuracy guarantees. Our approach considers run-time noise levels to selectively sense specific modalities, followed by selecting MMML models that are suitable for the chosen modality combination. We demonstrate the effectiveness of our solution using an exemplar mHealth application of pain assessment over various noise levels. Our solution achieves up to 23% improvement in prediction accuracy compared to Noise-agnostic method, and 42% energy savings in comparison with state-of-the-art selective sensing frameworks.**

Keywords: **Multi-modal machine learning, Efficient inference, Wearable computing, Internet of Things, Energy efficiency**

I. Introduction

mHealth applications such as sleep monitoring, pain and emotion recognition, activity tracking etc., require inputs from multiple sensor modalities for holistic data-driven decision making. Smart mHealth applications use Multi-modal Machine Learning (MMML) algorithms to fuse supplementary and complementary information across different sensor modalities for accurate predictive results [1]–[3]. There is an increasing demand for on-device and edge-only inference to guarantee low-latency resilient mHealth services and ensure privacy of users' sensitive health data. However, wearable devices have stringent compute capabilities and energy budgets to run data and compute intensive MMML algorithms.

On the other hand, continuous data acquired by wearable sensors in *everyday settings* is prone to perturbations such as unreliable signal quality, noisy components, and motion artifacts. Processing such perturbed input data drains compute and energy resources of wearable devices on un-insightful computations. Further, running inference on garbage data affects the confidence and prediction accuracy of MMML models. Existing mHealth strategies ignore/replace perturbed input samples through data filtering and pre-processing, outlier detection, and signal quality assessment [4]. These techniques focus on data quality rather than sensing efficiency, resulting in significant energy drain by continuously sensing data and eventually trashing the perturbed data. Advanced mHealth applications have adopted selective sensing i.e., sensing a subset of input samples based on run-time context-awareness [5], and relevance of a given sensor modality to achieve overall prediction accuracy [2]. While these approaches reduce the penalty of sensing and processing garbage data, they are agnostic to subsequent impact on prediction accuracy of MMML models. Effectively, input data perturbations from the sensing phase influences the efficiency of the compute phase, in terms of energy consumption and prediction accuracy. However, existing mHealth applications optimize multi-modal sensing and computation disjointly [6], affecting resilience of mHealth services and energy efficiency of wearables.

Developing mHealth services that are resilient to input data perturbations while minimizing energy consumption necessitates end-to-end sense-compute co-optimization. This approach (i) selectively senses data from insightful modalities and (ii) intelligently selects appropriate MMML models suitable for given input data and feature sets. Selective sensing minimizes sensing energy consumption and improves input data quality, and intelligent MMML model selection maximizes prediction accuracy while handling noisy input data. Further, sense-compute co-optimization attenuates total input data volume at the source, reducing also the communication energy penalty.

Our *ISCA intelligent sense-compute adaptive co-optimization* approach addresses these challenges through *continuous monitoring* of sensor modalities to detect input data perturbations, *selective feature aggregation* to isolate reliable inputs, and *model selection* to choose suitable MMML models for given input modalities and feature vectors. Selecting the optimal sense-compute configuration settings while considering run-time stochastic variance of multi-modal input data perturbations and diversity of MMML models is an NP-hard problem [7]. For such multi-constraint problems with conflicting objectives, Reinforcement Learning (RL) has been used as an effective approach for optimal decision making [8]. In the context of mHealth services, RL can be employed for sense-compute

TABLE I: Summary of MMML-based solutions for mHealth services.

	Related Works				
	[1]	[9], [10]	[5]	[11]	**ISCA**
Selective Sensing	✗	✓	✗	✓	✓
Noise-Awareness	✗	✗	✓	✓	✓✓
Network Awareness	✗	✗	✗	✗	✓
App Flexibility	✗	✗	✗	✓	✓
Platform Agnostic	✗	✗	✗	✗	✓

co-optimization decisions to maximize prediction accuracy while minimizing energy consumption [7]. In this work, we propose an RL guided sense-compute co-optimization framework for MMML based mHealth services. We design an RL agent to jointly configure *(i) sensing parameters* i.e., selecting/prioritizing among different sensor modalities and setting their sampling rates, and *(ii) model selection* i.e., to select a MMML model that is appropriate for the updated sensing configuration. The RL agent makes optimal sense-compute decisions by exploring accuracy-energy trade-off space, while considering run-time input data perturbations among different sensor modalities and prediction accuracy of different MMML models. Table I summarizes the key contributions of our ISCA approach in comparison with state-of-the-art mHealth orchestration strategies.

ISCA's novel contributions are as follows:

- A scalable sensor-edge sense-compute co-optimization framework for delivering resilient MMML-based smart mHealth services, capable of understanding input data discrepancies and optimizing end-to-end energy consumption
- Design and implementation of an RL agent for online decision making on setting sensing configuration, qualitative feature selection, weighted prioritization of input modalities, and input-driven MMML model selection
- Design of an adaptive rule-based controller to enforce the RL agent's decisions on feature, modality, and model selection, and sensing configuration for improving energy efficiency and prediction accuracy
- Evaluation of the proposed framework's efficiency on an exemplar pain assessment mHealth case study.

II. Motivation and Significance

We present the significance of sense-compute co-optimization approach for multi-modal mHealth services through an exemplar case study of pain monitoring application. The pain monitoring application acquires physiological data from different modalities viz., Photoplethysmography (PPG), Electrodermal Activity (EDA), and Electrocardiography (ECG) sensors to capture the autonomic nervous system activity against pain.

Sense-compute pipeline in mHealth services Figure 1 shows an end-to-end pipeline of multi-modal data acquisition, feature extraction, selective feature aggregation, and MMML model selection. In this example, PPG, EDA, and ECG sensors are sampled at 500Hz (2 channels), 4Hz, and 64 Hz, respectively. Relevant features from each input modality are extracted and fused at an early stage into a single feature vector. A relevant MMML model then predicts the

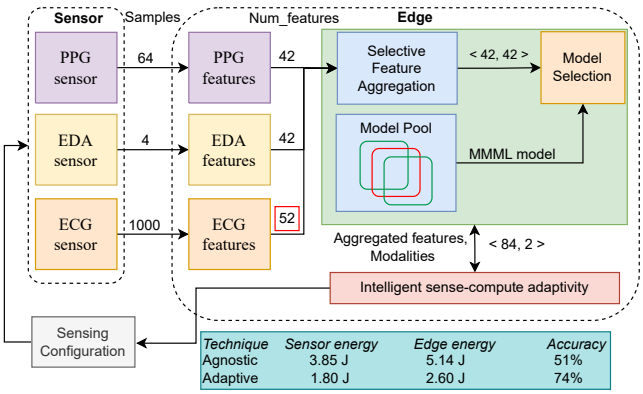

Fig. 1: Processing with modality selection (selective feature aggregation and model selection) and additional intelligence for adaptive sense-compute

pain levels using the aggregated feature vector. In practical scenarios, one or more modalities can be distorted due to noises caused by motion artifacts, physical damages, battery shutdown. In this example, we consider the ECG modality to be noisy due to motion artifacts. A naive baseline approach that is agnostic to input data perturbations processes noisy ECG data, resulting in a 51% prediction accuracy (shown at the bottom of Figure 1). On the other hand, an adaptive sense-compute optimization approach (i) reduces the feature volume of noisy ECG modality by lowering the sampling window, followed by (ii) selecting a MMML model that is suitable for the updated feature vector. This approach improves the prediction accuracy to 74%, which can be attributed to deliberate dropping of features from the noisy ECG modality. Further, lowered sampling of the noisy ECG modality also reduces the sensing energy consumption to 1.8J (from 3.85J with the baseline) and communication energy consumption at the edge to 2.6J (from 5.1J with the baseline). This demonstrates both accuracy and energy gains with the sense-compute co-optimization approach.

Dynamic accuracy-energy trade-off space Figure 2 shows the accuracy-energy trade-offs for the pain monitoring application under different levels of input data perturbations in the form of motion artifacts. In Figure 2, combination of different modalities is represented as a bit sequence, where 1= modality considered, 0= modality dropped. For example, the sequence `110` represents PPG and EDA modalities being considered in the MMML model and ECG modality being dropped. As shown in Figure 2, MMML models with different modality combinations provide different levels of accuracy at various levels of noise. For instance, the modality combination `011` (EDA+ECG) provides a higher accuracy (92%) with no noise, while the accuracy of this combination drops to 85% with 10% noise. Each modality combination consumes a different amount of sensing energy, represented by the size of the circle marker. Although the combination of `111` offers better accuracy, the energy consumption by using input data from all three modalities is higher. Also, certain uni-modal combinations (for example, `100`) provide relatively higher accuracy with significantly lower energy consumption. Both the modality combinations `100` (PPG) and `101` (PPG+ECG) provide higher accuracy even with

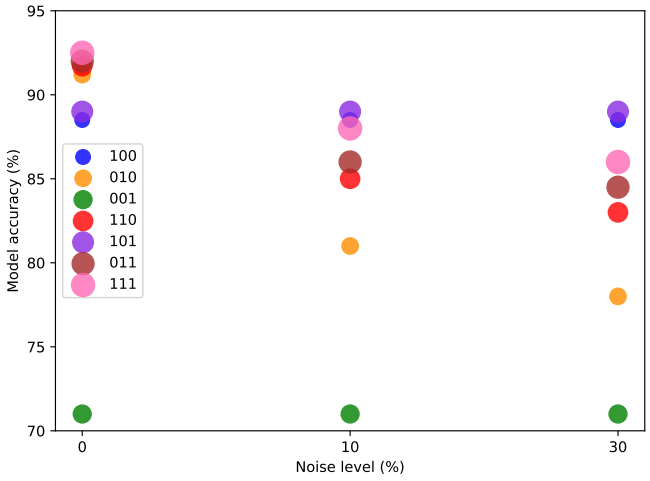

Fig. 2: Accuracy and energy trade-offs of different modalities with varying noise levels

increasing noise levels, since the noisy component in this example originates from the EDA modality. Furthermore, the modality combination `111` yields a relatively lower accuracy at the same noise level despite using data from all modalities, being affected by the noisy EDA modality, which explicitly shows the problem of garbage-in garbage-out. In real-world settings, sensor modalities can be prone to input data perturbations in a stochastic fashion, making the trade-off space dynamic and unpredictable. Selecting a suitable sense-compute configuration at run-time under such uncertainty becomes a complex challenge. For instance, the most insightful sensor modality becoming unreliable at run-time changes the weighted priority of other modalities, necessitating changes to MMML model selection and consequently affecting prediction accuracy and energy consumption. In our ISCA approach, we addresses these non-intuitive choices on sense-compute co-optimization under varying system dynamics through an intelligent reinforcement learning agent.

III. Sense-Compute Co-optimization Framework

In the following, we present our proposed intelligent sense-compute adaptive co-optimization framework (ISCA) and design of reinforcement learning agent that guides the sense-compute configuration settings.

A. System Architecture

Figure 3 shows an overview of the ISCA framework, with a pipeline of sensor devices at the *sensor layer* and computing resources at the *edge layer*. The sensor layer provides multi-modal sensing capabilities needed for target mHealth applications. The edge layer comprises of data handling, orchestration modules for intelligent sense-compute adaptivity, and a model pool with MMML models for inference.

Data Gateway captures raw inputs from different sensor modalities. Physiological signals from these modalities can contain different noise intensities, affecting specific segments of the input signal window, from zero (no noise) to 100% (highly noisy). *Data Processing* carries out pre-processing and feature extraction of the input signals collected from the sensors. This process involves synchronizing signals from various modalities, filtering and cleaning the input signals, and incorporating additional components essential for subsequent affective signal processing (ASP) pipeline, such as peak detection and data normalization. The ASP pipeline encompasses necessary physiological signals (e.g., ECG, EMG, and EDA), which are crucial for feature extraction in typical mHealth applications (e.g., stress and pain monitoring). Subsequently, informative features are extracted from previously pre-processed data. We extract handcrafted features in both time and frequency domains, along with additional automatic features extracted for all modalities. This comprehensive process not only accelerates performance of machine learning models for target mHealth application, but also facilitates the signal quality assessment process. *Quality Assessment* module monitors system parameters, disruptive sensory events, and data quality to analyze contextual information. Specifically, it evaluates the quality of signals and their extracted features by tracking key parameters from the sensing phase, and identifies events and triggers for jointly optimizing both sensing and sense-making.

Intelligent Sense-Compute Adaptivity module analyzes inputs from the Quality Assessment for configuring sensing and selecting relevant features, modalities, and MMML models. Our framework includes a *Model Pool*, which consists of a set of pre-trained models for various combinations of modalities and aggregated features. The compute configuration decisions are enforced by selecting the appropriate MMML model with the targeted modality and feature combinations from the *Model Pool*. The *Inference Engine* executes an instance of the selected model to run the inference for predictive results. Sense-compute configuration settings are determined by the RL agent, as described in the following.

B. Reinforcement Learning Agent Design

Reinforcement learning (RL) is widely used to automate intelligent decision making based on experience. Information collected over time is processed to formulate a policy which is based on a set of rules. Each rule consists of three major components viz., (a) state, (b) action, and (c) reward. Among the various RL algorithms [8], Q-learning has low execution overhead, which makes it a perfect candidate for runtime invocation. Specifically, model-free RL techniques operate with no assumptions about the system's dynamic or consequences of actions required to learn a policy. Model-free RL builds the policy model based on data collected through trial-and-error learning over epochs [8]. In this work, we design model-free RL agent to enable *Platform agnostic* and *Application flexible* sense-compute optimization decisions, where the agent finds optimal configuration through trial-and-errors during training phase with no assumption on *Platform* and *Application*. Figure 3 depicts the high-level block diagram for our RL agent. The RL agent is invoked at runtime for intelligent orchestration decisions. Our agent is composed as follows:

This article has been accepted for publication in IEEE Design & Test. 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/MDAT.2024.3469828

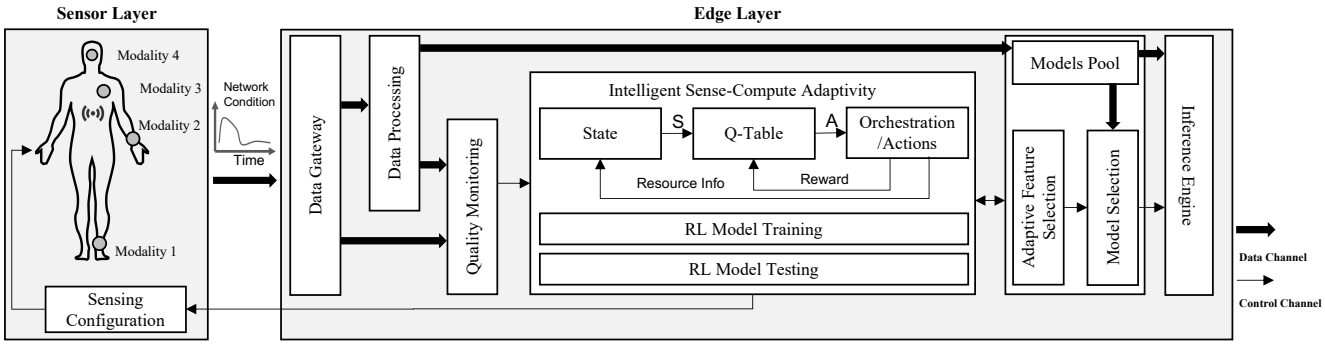

Fig. 3: ISCA System Architecture Overview.

State Space: Our state vector is composed of modality availability, feature volume per each modality, and noise level. Table II shows the discrete values for each component of the state. Modality availability (represented as MA) is a binary value that states our framework either collects sensory data for that modality or is idle. Feature volume (FM) states what percentage of the feature volume for each modality is incorporated for the further analysis after data processing. FM is considered as discrete value between 0 to 100%. Network condition (represented as NC) shows how the sensors are connected to the edge device. We consider three different network connections which demonstrates different data transfer speed and power consumption. Noise level (represented as NL) states what percentage of each modality is noisy. The state vector at time step τ is defined as follows:

$$S_\tau = \{FM_1, FM_2, FM_3, MA_1, MA_2, MA_3, NL, NC\} \tag{1}$$

Action Space: The action vector consists of increase/decrease feature volume for each modality, and configuring the sampling rate as: 1) 1x(default), 2) 0.8x, and 3) 0.6x of the default sampling rate, respectively. Consequently, we have 9 different possible settings for action at each time step, which includes selecting the feature volume for each of the three modalities and setting the sampling rate for that specific modality. The action vector at time step τ is defined as follows:

$$A_\tau = \{A_1, A_2, A_3, ..., A_9\} \tag{2}$$

Reward Function: The reward function is defined as the negative total energy consumption including edge and sensor devices. In our case, the agent seeks to minimize the energy consumption. To ensure the agent minimizes the energy consumption while satisfying the accuracy constraint, the reward R is calculated as follows:

$$\begin{aligned} &\text{if } \overline{\text{Accuracy}} > \text{constraint:} \\ &\qquad R_\tau \leftarrow -\text{Energy} \\ &\text{else:} \\ &\qquad R_\tau \leftarrow -\text{Max Energy} \end{aligned} \tag{3}$$

To apply the accuracy constraint, the minimum possible reward is assigned when the accuracy threshold is violated. On the other hand, when the selected action satisfies the accuracy constraint, the reward is negative energy consumption

TABLE II: State Discrete Values

State	Discrete Values	Description
FM_i	0,35%,70%,100%	Modality Feature Volume
MA_i	0,1	Modality Availability
NL	0, 20%, 50%	Modality Noise Percentage
NC	Regular, Moderate, Weak	Network Condition

Algorithm 1 Q-Learning Algorithm

1: **while** system is on **do**
2: **From *Resource Monitoring*:**
$S_\tau \leftarrow$ State at step τ
3: **if** $RAND < \epsilon$ **then**
4: Choose random action A_τ
5: **else**
6: Choose action A_τ with largest $Q(S_\tau, A_\tau)$
7: **end if**
8: Monitor total energy consumption
9: Calculate reward R_τ
10: **From *Resource Monitoring*:**
$S_{\tau+1} \leftarrow$ State at step $\tau + 1$
11: Choose action $A_{\tau+1}$ with the largest $Q(S_{\tau+1}, A_{\tau+1})$
12: **To *Updating Qtable*:**
$Q(S_\tau, A_\tau) \leftarrow$
$Q(S_\tau, A_\tau) + \alpha[R_\tau + \gamma Q(S_{\tau+1}, A_{\tau+1}) - Q(S_\tau, A_\tau)]$
13: $S_\tau \leftarrow S_{\tau+1}$
14: **end while**

in that time step. Algorithm 1 defines our agent's logic with the epsilon-greedy Q-Learning:

Line Description

3: The agent determines the current system state from the resource monitors.
4-8: Either the state-action pair with the highest Q-value is identified to choose the next action to take, or a random action is selected with probability ϵ.
9-10: The selected action is applied and normal execution resumes. The reward R_τ for the execution period is calculated based on measured consumed energy.
11-12: Based on the resource monitors, the new state $A_{\tau+1}$ is identified, along with the state-action pair with highest Q-value.
13: The Q-value of the previous state-action pair is updated.
14: The current state is updated, and the loop continues.

Hyper-parameter Tuning An RL agent has a number of hyper-parameters that impact its effectiveness (e.g., learning rate, epsilon, discount factor, and decay rate). The ideal

values of parameters depend on the problem complexity, which in our case scales with the number of modalities and noise level. In order to determine the learning rate and discount factor, we evaluated values between 0 and 1 for each hyper-parameters. We observed that a higher learning rate converges faster to the optimal, meaning the more the reward is reflected to the Q-values, better the agent works. We also observed that a higher discount factor is better. This means that the consecutive actions have a strong relationship, so that giving more weight to the rewards in the near future improves the convergence time. The selected configuration for hyper-parameters is as follows: $\alpha = 0.99, \gamma = 0.7$, decay rate $= 0.1, \epsilon_{initial} = 0.1$

IV. Evaluation

In this Section, we present our evaluation of the proposed ISCA adaptive sense-compute co-optimization approach over a case study of pain assessment application under scenarios with diverse noise levels and different networks.

A. Experimental Setup

Pain Assessment Application The pain assessment application requires continuous monitoring of multiple sensor modalities including ECG, EMG, and EDA signals. We collected data from face-, chest-, and wrist-worn devices via an eight-channel biopotential acquisition system for EMG and ECG recordings and Empatica E4 for EDA recordings. We used 500 Hz ECG (2 channels), 500 Hz EMG (6 channels), and 4 Hz EDA sensor modalities during our experiments. We pre-processed each modality and segmented them into 60 second windows. We use different uni-modal and multi-modal pre-trained models for the pain assessment application. These models are trained using iHurt Pain DB, a multi-modal dataset from postoperative patients in hospital from 20 patients [12]. After pre-processing, we extract a set of unique features from each modality viz., 52 features of ECG, 52 features of EMG, and 42 features of EDA.

Platform Our platform for evaluation comprises a sensory node - to collect physiological input modalities of ECG, EMG, and EDA from the subject, and an edge node - to control sensing and computation and execute the inference. We deployed the ISCA framework on ODROID-XU3 with an octa-core Exynos processor as the edge device.

Evaluation Scenarios We evaluate the ISCA approach under different levels of input data noise (0-50%), across three different networks (Wifi, 4G, 3G), exposing diverse scenarios of energy and accuracy exploration targets. We augmented the raw input training data with real-life noises such as BW and MA at various portions like 20% and 50% of data, to handle potential noisy components in physiological signals for the prediction model. We evaluate the accuracy and energy efficiency of the pain monitoring application under each of these scenarios. We trained the RL agent of the ISCA framework with accuracy-energy data extracted by profiling the pain assessment application under varying noise levels and network types. Within the current ISCA framework, target of the RL agent is to minimize energy consumption while satisfying the accuracy constraints.

B. Experimental Results

We present energy consumption and prediction accuracy of the pain assessment application under different noise levels over different network types. We compare the evaluation metrics our proposed ISCA approach against (i) Noise-agnostic baseline (agnostic to input data perturbations) and (ii) framework presented in [11] named AMSER. Figure 4 shows the total energy consumption of the pain assessment application with 0%, 20%, and 50% noisy data over different network types viz., (a) WiFi, (b) 4G, and (c) 3G networks. We present the total energy consumption as a summation of energy consumed in sensing, uploading the data from sensor to edge, downloading data at the edge, and computation at the edge. In Figure 4, each of these individual energy splits are shown as a stacked bar (with different shades) for Noise-agnostic (in orange), AMSER (in blue), and ISCA (in green). In Figure 4 (a-c), we show the modality combination used by the Noise-agnostic, AMSER, and ISCA methods on top of the energy bars. The notation used for modality combination is a 3 digit number, with each digit corresponding to ECG, EMG, and EDA respectively. Each digit represents the selected feature volume for a modality such that `0`= 0% features (dropping the modality), `1`=35%, `2`=70%, and `3`=100% of the features considered. For example, the modality combination `300` indicates utilizing all the features from ECG sensor and dropping the whole EMG and EDA modalities, while the modality combination `203` indicates using 70% of ECG features, dropping the EMG, and utilizing 100% of the feature volume from EDA. In each of the scenarios, we set a minimum accuracy constraint of 70%, 65%, and 60% for no noise, 20% noise, and 50% noise scenarios, respectively. In each case, we determined the minimum accuracy constraint as the median of accuracy levels achieved across all possible sense-compute decisions with different modalities and feature volume combinations. For this purpose, we used the offline profiling data extracted while training the RL agent. As shown in Figure 4, our proposed ISCA approach has significant energy savings with average saving of 72% and 42% in comparison with Noise-agnostic and AMSER methods (across different noise levels and network types), respectively. This is attributed to the intelligent sense-compute configuration, where the ISCA approach configures the sampling rate of each input modality by considering varying noise levels at run-time. This enables selecting specific feature volumes and modality combinations that guarantee target accuracy level while minimizing energy consumption at sensing, communication, and edge layers. For instance, in Figure 4(a), the ISCA approach selects modality combinations of `333`, `203`, and `201` for 0, 20% , and 50% noise levels respectively. In each of these cases, specific modalities are dropped/re-configured, leading to energy savings within accuracy constraints. Whereas, the Noise-agnostic approach continues to sense in full throttle across all modalities `333` irrespective of the run-time noise, resulting higher energy consumption, while also degrading prediction accuracy. It is noteworthy to mention that although AMSER framework is capable of selective sensing, since it is platform

This article has been accepted for publication in IEEE Design & Test. 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/MDAT.2024.3469828

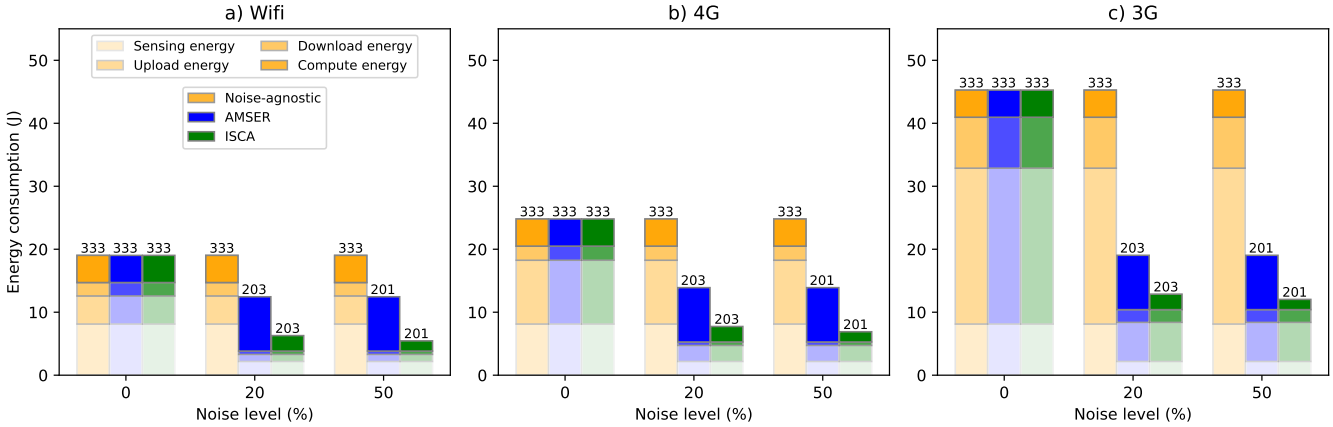

Fig. 4: Total energy consumption (sensor + edge) of ISCA (in green) in comparison with Noise-agnostic method (in orange) and AMSER [11] (in blue) in different noise levels for (a) WiFi (b) 4G and (c) 3G networks

and network agnostic, it needs to re-evaluate various models' performance with different modality combinations in each scenario so it can pick the best one in terms of accuracy and energy consumption, which results in higher compute energy consumption at edge (compute energy). On the other hand, ISCA picks the best performing modality combination for inference from the trained Q-table at run-time without compute energy overhead.

Table III shows the prediction accuracy of the proposed ISCA approach against Noise-agnostic and AMSER for different noise levels. With 20% noisy data, the baseline accuracy drops from 81% to 46.5%, while the ISCA approach provides an accuracy of 70.15% through intelligent modality selection. Similarly, at 50% noise level, the baseline slides to an accuracy of 41.6%, whereas ISCA delivers 60.2% accuracy, demonstrating the resilience of our proposed sense-compute co-optimization approach. In each of these cases, the proposed ISCA approach meets the accuracy constraints with a significantly lower energy consumption (4), while the baseline fails to meet the accuracy constraints and also drain energy across the system. Our evaluations on other network types (4G and 3G) follows a similar trend of gains with ISCA compared to Noise-agnostic and AMSER framework, while meeting the minimum accuracy requirements. It should be noted that prediction accuracy of a given modality combination remains the same across different networks. This can be related to limited dimension of state and action space of the RL-agent, which will be explored in other multi-modal applications as a future work of this paper.

TABLE III: Accuracy gain with ISCA(%)

Approach	No noise	20 % noise	50 % noise
Noise-agnostic	81.1	46.54	41.6
ISCA	81.1	70.15	60.29

Overhead Analysis We evaluate the time required by the proposed agent for the training phase to identify an optimal policy. Figure 5 shows the training phase under different accuracy constraints using Q-Learning algorithm. This shows that when training a model from scratch, the reward converges after about about 170 inference runs on average.

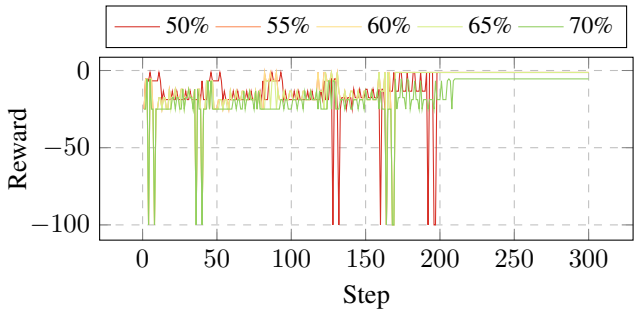

Fig. 5: Training overhead for Q-Learning algorithm under different accuracy constraints

However, increasing the accuracy constraint leads to a more complex problem and therefore increase in convergence time.
Runtime Overhead: To demonstrate the viability of mobile inference deployment, we evaluate the ISCA runtime overhead. The performance overhead of RL algorithm in ISCA is, on average, 20 μs for training, excluding the time for inference execution. It corresponds to 1.2% of the lowest inference latency. In addition, when using the trained Q-table, the overhead can be reduced to 7.3 μs with only 0.3% overhead. This result means it takes 18.1 μs to measure the inference results, calculate the reward, and update the Q-table. The energy overhead is only 1% and 0.2% of the total system energy consumption, when training the Q-table and exploiting the trained Q-table, respectively.

V. Conclusion

We proposed ISCA, a novel RL-based intelligent and adaptive multi-modal sense-compute co-optimization framework for energy efficient and resilient mHealth applications. Our approach monitors input signal and feature quality to configure sense-compute settings that reduce non-informative data, improve energy efficiency, while meeting accuracy constraints. We demonstrated the effectiveness of the ISCA framework with an exemplar mHealth application of pain assessment, achieving up to 42% energy savings in comparison with state-of-the-art selective sensing methods, and up to 23% accuracy gains in comparison with the baseline method.

VI. Acknowledgements

This work was partially supported by Nokia Foundation and Kaute Saatio, Finland.

References

[1] M. Kächele *et al.*, "Multimodal data fusion for person-independent, continuous estimation of pain intensity," in *International Conference on Engineering Applications of Neural Networks*. Springer, 2015, pp. 275–285.

[2] H. Alikhani, A. Kanduri, P. Liljeberg, A. M. Rahmani, and N. Dutt, "Dynafuse: Dynamic fusion for resource efficient multi-modal machine learning inference," *IEEE Embedded Systems Letters*, 2023.

[3] A. Kanduri, S. Shahhosseini, E. K. Naeini, H. Alikhani, P. Liljeberg, N. Dutt, and A. M. Rahmani, "Edge-centric optimization of multi-modal ml-driven ehealth applications," in *Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Use Cases and Emerging Challenges*. Springer, 2023, pp. 95–125.

[4] E. K. Naeini, F. Sarhaddi, I. Azimi, P. Liljeberg, N. Dutt, and A. M. Rahmani, "A deep learning–based ppg quality assessment approach for heart rate and heart rate variability," *ACM Trans. Comput. Healthcare*, vol. 4, no. 4, nov 2023.

[5] D. Amiri *et al.*, "Context-aware sensing via dynamic programming for edge-assisted wearable systems," *ACM HEALTH*, 2020.

[6] S. Liu *et al.*, "On-demand deep model compression for mobile devices: A usage-driven model selection framework," in *MobiSys*, 2018.

[7] S. Shahhosseini, D. Seo, A. Kanduri, T. Hu, S.-S. Lim, B. Donyanavard, A. M. Rahmani, and N. Dutt, "Online learning for orchestration of inference in multi-user end-edge-cloud networks," *ACM Transactions on Embedded Computing Systems (TECS)*, 2022.

[8] R. Sutton *et al.*, *Reinforcement learning: An introduction*. MIT press, 2018.

[9] B. Pourghebleh *et al.*, "Data aggregation mechanisms in the internet of things: A systematic review of the literature and recommendations for future research," *Journal of Network and Computer Applications*, 2017.

[10] C. Perera *et al.*, "Context aware computing for the internet of things: A survey," *IEEE communications surveys & tutorials*, 2013.

[11] E. K. Naeini *et al.*, "Amser: Adaptive multimodal sensing for energy efficient and resilient ehealth systems," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2022, pp. 1455–1460.

[12] Naeini *et al.*, "Prospective study evaluating a pain assessment tool in a postoperative environment: Protocol for algorithm testing and enhancement," *JRP*, 2020.

Hamidreza Alikhani is a PhD student at the University of California, Irvine. His research interests include Efficient inference specifically for Large Foundation Models, and context aware and resilient multi-modal machine learning.

Anil Kanduri is a University Lecturer and Senior Researcher at Department of Computing, University of Turku, Finland. He received PhD in Computer Systems from University of Turku in 2018. His research interests are broadly in resource efficient edge AI.

Emad Kasaeyan Naeini earned his PhD degree from University of California, Irvine. His research interests include Internet of Things (IoT), data analytic, and health monitoring.

Sina Shahhosseini Sina Shahhosseini earned his PhD in Computer Engineering from the University of California, Irvine, in 2023. His research focuses on Efficient AI, specializing in optimizing deep learning models.

Pasi Liljeberg is a full Professor and Head of the Department, at Dept. of Computing, University of Turku, Finland. He received PhD in ICT from University of Turku in 2006. His research interests include AI for healthcare and medical IoT.

Amir M. Rahmani is a Professor of Nursing and Computer Science at the University of California, Irvine. His research interests include mHealth, wearable and mobile computing, machine learning, and affective computing

Nikil Dutt is a Chancellor's Professor of Computer Science at University of California, Irvine. His research interests includes computational self-awareness principles for adaptive and resilient system design, and smart IoT-enabled healthcare technology.