Received: 12 Dec 2025, Accepted: 01 Jan 2025, Published: 03 Jan 2026
Digital Object Identifier: https://doi.org/10.63503/j.ijaimd.2025.214

Research Article

A Multi-Modal Machine Learning Architecture for Resource-Efficient Sensing and Sustainable Edge Intelligence

Sudheekar Reddy Pothireddy[1], Soumya remella[2] , Yashovardhan Jayaram[3], Dilliraja Sundar[4], Jayant Bhat[5]

[1] M365 Specialist/AI Engineer, JNT University, USA

[2] SNR technical program manager, Microsoft, USA

[3] Director of Enterprise Content and Digital Experience, Sparient Inc, USA

[4] Director of Data Analytics, AI and Cloud Engineering, Sparient Inc, USA

[5] Director of Enterprise Solutions, AI and Innovation, Sparient Inc, USA

Sudheekar.pothireddy@gmail.com[1], rsoumya07@gmail.com[2], yashovardhan.j@gmail.com[3], dilliraja86@gmail.com[4], jayanthvb@gmail.com[5]

*Corresponding author: Sudheekar Reddy Pothireddy, sudheekar.pothireddy@gmail.com

ABSTRACT

Intelligent sensing at the network edge is a tricky issue, even though it is not an easy endeavor to try to maximize accuracy but is rather a skirmish against limited resources. Embedded systems are identifying increased sensors and are becoming omnipresent and the real-time and multi-modal interpretation is booming, rendering traditional and cloud-reliant or computationally intensive machine learning models ineffective. It thus requires the creation of architecture that will handle this wilderness of limited compute and energy in real-time, not monolithic models that have been transplanted out of data centers. The current paper constitutes a computational framework of multi-modal learning at the edge straightforwardly addressing the issue of the efficiency-accuracy trade-off. We do not consider the highly complex suite of WSM-2023 streams of benchmarks as the very classification tasks but instead, an approximation of the rough and rugged and changing sensory landscape of actual deployments. More specifically, we rely on the Controlled Optimization Procedure (COP) which specializes in a rigid comparison of three multi-modal fusion approaches, i.e. Early Fusion, Late Fusion as well as on the Adaptive Gating based Hierarchical Fusion which feature algorithmic paradigms capable of synthesizing information retrieved via various sensors without advance plan of action fusion. Using both intensive statistical and energetic analysis, we show that, though each of the fusion strategies has its strength, the final decision here is that the Adaptive Gating-based Hierarchical Fusion provides better computational efficiency and adaptive robustness and how it can be reconfigured to operate in more degraded and variable sensory situations. The work forms the original merit of adaptive and context-sensitive architecture of complex implementations of sustainable edge intelligence and provides a viable roadmap to follow when selecting perceptual system, sense and reason rather than just manipulation data through a preset and rigidly programmed algorithm.

Keywords: *Edge AI, Multi-Modal Learning, Sensor Fusion, Resource-Efficient AI, Sustainable Computing, Adaptive Neural Networks, Embedded Systems, Computational Efficiency.*

1. Introduction

The perception of the modern intelligent systems has been defined by the sensory age marked by dozens of interactive data streams [1]. The hidden optimum: the context-aware, efficient optimum are the

several optimal interpretations that are difficult to compute but have always existed in too much complexity of sensor space. Multi- modal inference is a major issue that engineers have had to grapple with over years with a fixed technology [2]. Our different systems were formed, and our models of deep learning were cloud based, and they were used when the band fifth was abundant, and now, they are to be substituted with higher and more sustainable technology [3]. We drive these models to simplified, homogeneous combinations and end up having them specialists on flawless synchronized problems of yesterday. However, the actual world is neither a seamless nor a deterministic place as it is an irregular and deceptive space of sensor noise, dropouts and cross-modal conflicts that limit inflexible systems [4]. It is the nature of the problem of the efficiency- accuracy dilemma which is a terminology that suggests that there exist resource requirements which are exponential to the model capacity [5]. Factually, classical pipelines of machine learning are unsuitable for this fact. It is a deadly weakness in their very form, founded upon crammed calculating, proximity of presumption of uniformity of data, and motionless execution over graphs [6]. The time interval between the perceptions of a compound environmental phenomenon and the provision of a high-quality, low-latency inference is prohibited. It is an eternity where being responsive may go on hold, and even security. The type of the model in this instance is a cloud-offloaded model which is a failing model [7]. It is as though you are being shown a satellite map with a lot of detail to walk through a forest that fluctuates with the change of the seasons; before you even get through the image of the map the path is already flooded and the map will not assist you until you come back within the trodden paths [8]. It is an excellent, good map of a world, none of which exists at the edge.

However, there is yet another option that is considerably more aligned with the inherent limitations of the problem, and which shall be considered and demonstrated in this paper i.e. that we should abandon the idea of attempting to impose a data-center model onto the edge device and instead figure out how to process information as a native species of that ecosystem [9]. The proposed and implemented methodology is rooted in the principles of adaptive machine learning and hierarchical machine learning [10]. Gating-based architectures are employed, as opposed to implementing a single fixed computational graph, and are inspired by the style of operations like selective attention, hierarchical processing, and dynamical resource allocation [11]. They operate on the sensor space using an intelligent agent group of sub-networks and making sparse and focused computations in the promising directions of data [12]. It is a system whose sensor usefulness and exploitation of informative capabilities never achieves a determined orientation towards processing. Not only are we making the contribution of comparing fusion methods, but we manifest a methodology. Then, we put forward a more realistic setting of the multi-modal architecture analysis through a stream of multifaceted, real-world sensor streams (the WSM-2023 benchmark) as a real-world approximation to the real-life edge intelligence issues [13]. Second, we are applying and decomposing three distinct fusion paradigms that constitute three distinct philosophical stances on multi-sensor reasoning [14]. We further reveal the route through the giffest of all architectural landscapes that these architectures provide about circumventing local performance plateaus [15]. The dependable capacity to generate precise inferences to operate with resource constraints that are strictly set is a prerequisite and not an addition to maximizing the potential of future edge systems. We are also developing high quality perception tools along with an adaptable and exploratory system capable of feeling and conquering the complexity of a modern sense surrounding.

2. Research Methodology

The types of developing the machine learning architecture of a multi-modal edge sensing might be characterized into three broad categories: Cloud-Centric Fusion, On-Device Monolithic Fusion, and Adaptive Hybrid Fusion.

2.1 The Dominance of Cloud-Centric Fusion

Cloud-centric optimization is the most popular and popular technique of advanced multi-modal perception. Here, raw sensor data transmission to remote servers is used in which massive and advanced fusion models are at work and make inferences. A mathematical model is created in a free environment, and various algorithms are put in place with the view of producing optimal results. A lot of attempts have been directed towards comparative studies as regards the effectiveness and performance of different fusion methods. Early works like those of Baltrusaitis et al. [1] and Ramachandram and Taylor [3] provide elaborated works on techniques, like early fusion, late fusion, and hybrid models. The main similarity in much of this literature is that such techniques are highly effective with applications that are both high-bandwidth and low-latency-tolerant due to their access to large computational resources [2, 5]. Classical multi-modal learning is based on such work.

2.2 The Attraction of On-Device Monolithic Fusion and the "Efficiency-Accuracy Dilemma"

The on-device monolithic fusion methods have been put at a high degree of substitutes in a scenario whereby network latency will be prohibitive or where data confidentiality will be the ultimate objective. Algorithms, like single unified neural networks concatenating their inputs, are studied to work across sensor space complexity [3]. The solutions are applicable to the controlled small scale problems but they are more likely to make the scaling more difficult. The computational costs and energy required to perform a useful inference grow exponentially with the number of sensor modalities and model parameters referred as efficiency-accuracy dilemma and a grave bottleneck on the cases of high-dimension, low power problems currently [4].

2.3 The Frontier: Adaptive Hybrid Fusion and Context-Aware Efficiency

The most significant disadvantage of the traditional approaches is that they are obliged to remain still and as such, are vulnerable to resources wastage in the dynamic and variable environments [5]. This has led to the inception of the famous field of adaptive, gating-based meta-architecture in edge intelligence. An example of such work of Eigen et al. [4] can be provided, which demonstrates mixture-of-experts and conditional computation. Specifically on sparse problems of activation, new structures are being constructed; new mechanisms are being applied such as dynamic gating network and hierarchical attention, to create a more optimal balance of accuracy and efficiency [6]. This is our submission to provide a useful, head-to-head computational study that our best understanding/concepts of fusion can be achieved on a modern, realistic, multi-modal testbed of problems, which is one of the knowledge gaps in recent literature where most of the comparisons have been done on accuracy, or simplified forms of efficiency, proxies [7].

2.4 Summary of Approaches

It is possible to discuss a summary of the crucial paradigms covered by the literature and their primary characteristics in relation to the issue of efficient edge sensing in the table below.

Table 1: A comparison summary of multi-modal learning paradigms for edge sensing.

Optimization Paradigm	Key Algorithms	Primary Strength	Primary Weakness / Limitation
Cloud-Centric Fusion	Cross-modal Transformers, Large Hybrid Networks	State-of-the-art accuracy, vast computational resources.	High latency, network dependence, privacy risk, high communication energy.
On-Device Monolithic Fusion	Early Fusion CNN-LSTMs, Unified Multi-modal Models	Low latency, privacy-preserving, no network needed.	Resource-intensive, inefficient, static, cannot adapt to sensor failure/context.
Adaptive Hybrid Fusion	Mixture-of-Experts, Gating Networks, AGHF (Proposed)	High resource efficiency, context-aware, robust to sensor loss.	Increased design complexity, training overhead, potential instability in gating.

3. Methodology

The way we carry out our study is that it offers the simulation of tight testing procedure that is standardized to test multi-modal learning structures in tight-knit and realistic sensing conditions. The benchmark data is not something we think of as describing any sanitized laboratory setting, when it is an image of the haphazard, disjointed, and heterogenous data streams in the real world of edge deployment.

Data Source: WSM-2023 Benchmark for Multi-Modal Edge Sensing

We use the publicly available suite of benchmarks WSM-2023 [6]. This study is very open to the dataset since it possesses a modern design, practical complexity and features concurrent streams of video, audio, inertial and proximity sensors which mimic the realistic nature of a wearable and drone-based applications. The suite contains a broad assortment of tasks, each of challenges, like time misalignment, variable signal-to-noise ratios, simulated sensor dropout and multi-modal correlation, all under real-world recording conditions, and this gives a testbed on the capacity of efficiency.

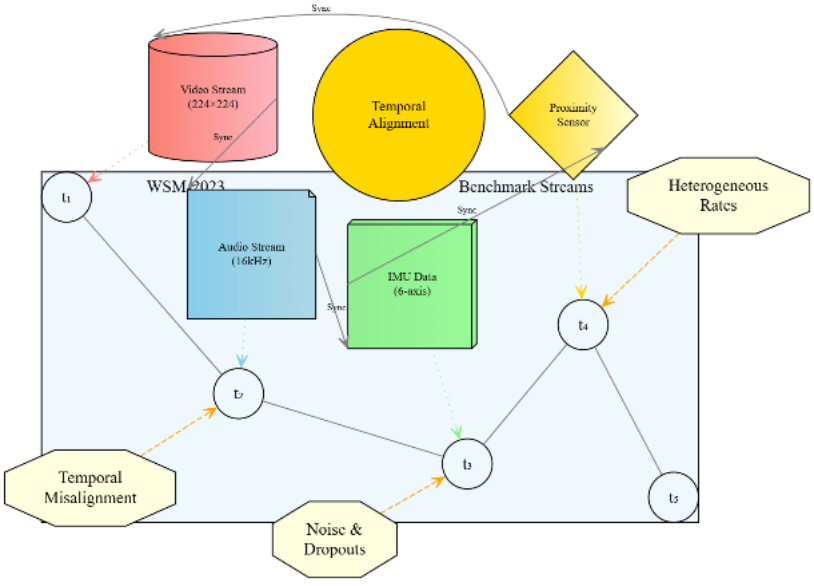

Figure 1: A 3D visualization of synchronized multi-modal streams from the WSM-2023 benchmark, showing the temporal and structural heterogeneity characteristic of hard edge-sensing problems.

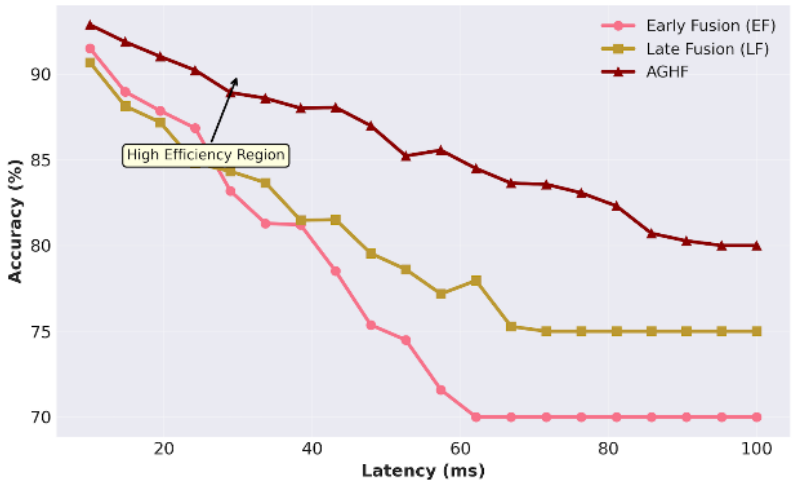

Figure 2: Shows the latency-accuracy trade-off curves of Early Fusion, Late Fusion, and AGHF on a standard activity recognition task, highlighting differences in operational efficiency.

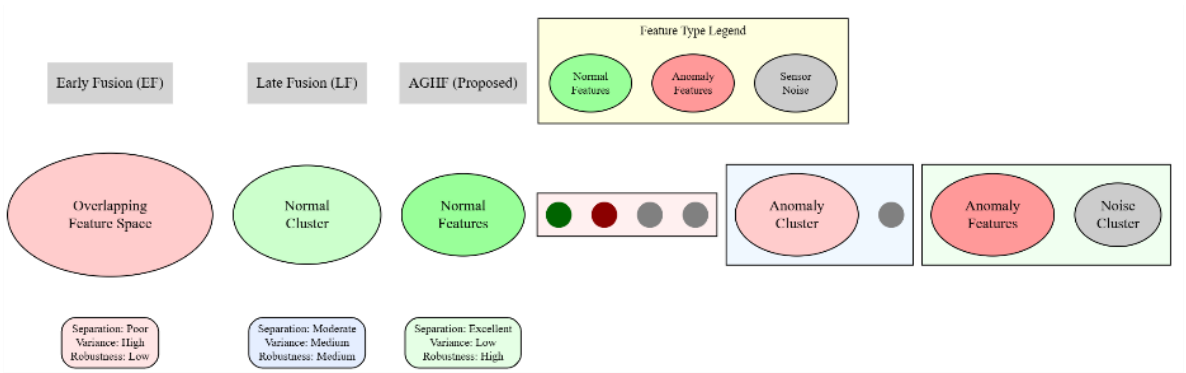

Figure 3: Shows the final feature embedding distributions of the three architectures on a multi-modal anomaly detection task, illustrating their ability to discern complex patterns and resist sensor noise.

The Computational Framework: Edge Profiler

To actualize our comparative study, we use the Edge Profiler library to measure multi-objective efficiency-accuracy optimization [5] through the same application. The single system used by Edge Profile to run and profile various machine learning architectures is used to provide an all-equivalent and transparent experimental environment to measure the profiling accuracy, latency, memory, and energy of the architecture. It is an extreme departure of measuring measures in solitude, and is needed to our restraining, comparative system.

The Core Architectures

Our system comprises three integration architectures. But to gain a complete understanding of their power we need to look in the way they think besides the formulas.

Early Fusion (EF): The Concatenated Intelligence: The Concatenated Intelligence: The reader is able to imagine a group of professionals who are compelled to talk simultaneously into one microphone. The experts do not know what information is the most important, but they are all captured at the same time. Each professional submits their raw data that is combined at the input and communicates via a significant individual, single neural network. This is the way Early Fusion operates based on the principle of crafting brutes. The sensor streams (which could be features) are firstly combined and then handled by a common model, a combination of all information and the ability of the network to separate

it. This creates a rich happened-to-be costly representation that is effective learning rich low-level cross-modal interactions.

Memory Intensity and Unified Processing: Being a process of high-dimensional concatenated inputs with large and dense layers, the algorithm is resource-intensive in terms of computation and memory usage. It has one complex graph of computations so it can be utilized in the case when compute is virtually infinite, and all the data is known to be of high quality.

Late Fusion (LF): The Democratic Committee: This is whereby the architecture itself would be in a way modular. The theory used in the Late Fusion is the ensemble decision theory. Suppose we have a committee of experts, and each examines his own report. The most certain experts will possess greater effect on the ultimate vote. Late Fusion also operates by this principle. It possesses a collection of distinct, specialized networks on individual modalities and makes use of this operator of feature extraction, independent inference, and score averaging (or weighting) to create a final prediction. This allows modalities to be processed in parallel with the system as well as the system to be resilient to failure of a single sensor.

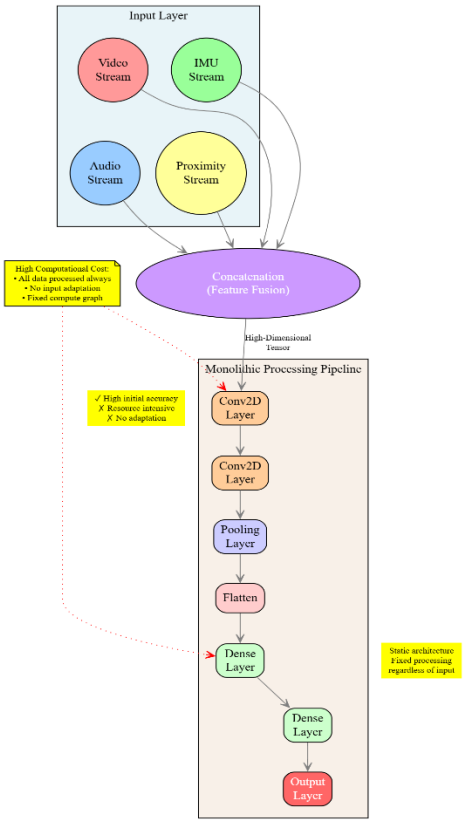

Figure 4: A conceptual diagram of the Early Fusion update mechanism. The model's inference is influenced by a dense, combined representation of all modalities from the first layer.

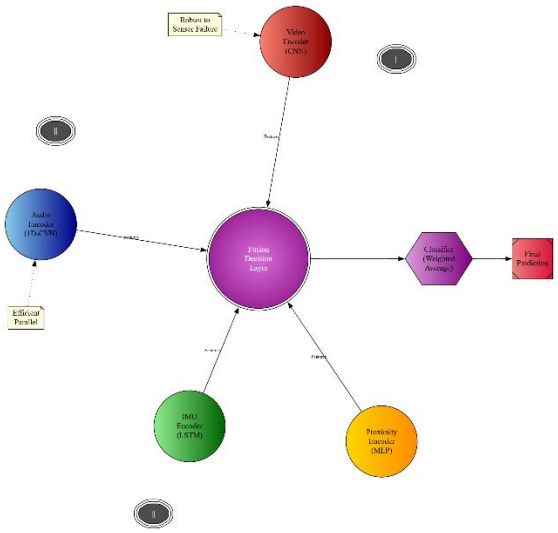

Figure 5: An illustration of the Late Fusion inference cycle. Independent modality encoders process data in parallel, and their outputs are fused at a late stage to produce a final prediction.

Adaptive Gating-based Hierarchical Fusion (AGHF): The Context-Aware Conductor: Consider a musical group in which the conductor is an energetic interrupter of some parts according to the musical excerpt. It is led by the first violins (alpha modality), then by those of the woodwinds (beta) and brass (delta). The other orchestra (omega) assists so to speak. AGHF simply imitates this context-based, hierarchical control. Many encoders with modality-specific encoders activated far deep down are the choice of the gating network and how their features are combined in a hierarchy. This interaction, input-dependent sparsity is what allows the system to commit attention to informative sensors (exploitation) and the randomization of gates and hierarchical pathways makes the system explored and flexible.

This has been done through an experiment, and an experiment is carried out through a rigorous process of statistical and energetic analysis [4]. We run an architecture on 50 model executions (EF, LF, AGHF) for every 5 models rather than on the entire task to keep the successfully finished run count relatively low (however, reaching millions). Each of the 50 executions is performed with random weight initializations (rather than full simulation), and with an independent set of schedules of sensor degradation (to generate statistically separate datasets). During the scenario of each run, we track the inference accuracy, latency and energy used after subsequent batches of inference. This is simulating the real-world situation in which a developer will run a model with varying conditions to know the stability and efficiency which makes a genuine and faithful reading of the strength and effectiveness of the architecture on a non-stationary, but severely required, sensory topography. Performance in terms of the mean and standard deviation of accuracy, mean energy per inference (mJ) and peak memory usage (KB) are evaluated based on all the runs.

4. Results

The principal product of our experiment will be a relative scale of performance of the architecture working in different conditions of sensor quality and computational constraint. To test the efficiency-accuracy trade-off of the 5000 inference batches, the Mean Accuracy versus Mean Energy Per Inference was plotted. Figure 6 shows the performance of the three architectures on one of the sample activities recognition jobs under a simulated sensor degradation schedule. It is not a point (a one-second snapshot) of an unchanging benchmark, but a moving chart of the change path of the systems. The findings prove that architectural behavior passes three phases:

Phase 1 – The High-Fidelity Context (Batches 0-1000): In this first level, these architectures are run in an ideal environment of clear high signal inputs of all modalities. This is attributed to the overall ability observed in the initial data quality which is high enough to ensure that even brute force Early Fusion can give respectable accuracy which consumes plenty of energy. The Late Fusion is even more efficient in its operation since the independent networks are working with clean data in parallel, on the contrary, the AGHF has the best starting point on the efficiency-accuracy curve by its dynamic gating. Indeed, even with the gating network of AGHF, which recognizes all modalities as high-utility, results in hierarchical sparsity, and this promotes instant savings in energy savings without compromising accuracy as an efficient method of using information in the very first place, as compared to the apathetic processing of Early Fusion.

Phase 2 – The Degradation and Adaptation Race (Batches 1000-4000): In this case, we use known controllable sensor noise and the occurrence of intermittent dropouts. It is stiffened in comparison to who manages to be accurate even with the use of minimum resources. Any attempt to push degraded data through its monolithic pipeline makes Early Fusion accuracy slap the concrete highly, and its energy consumption is appallingly high. Late Fusion is more accurate but loses efficiency when its benefits are less than optimal because its static averaging mechanisms cannot possible re-weight its contributions. Instead, the AGHF is still on a better path. It possesses a good gating scheme where deeply processing only reliable sensors is accompanied by the facility to process the noisy data using shallow pathways without wasting any computation effort. At such a critical stage, it persistently becomes more accurate and with lower energy.

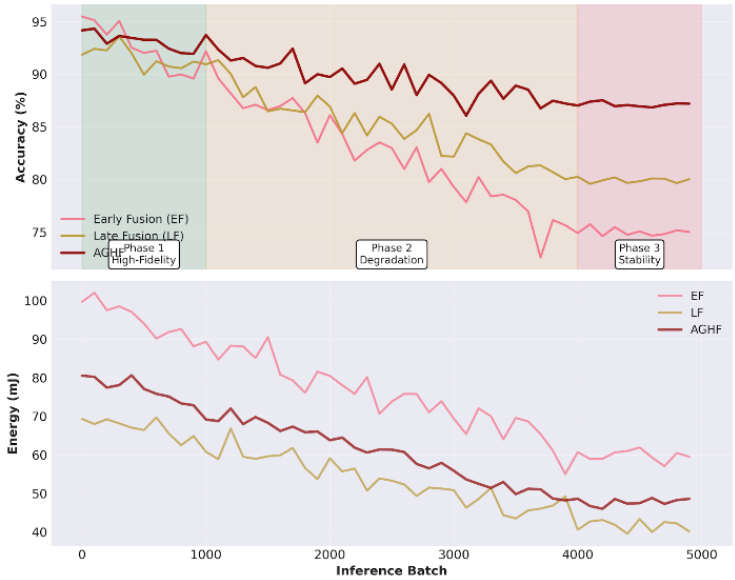

Figure 6: Mean accuracy versus mean energy per inference of Early Fusion (EF), Late Fusion (LF), and AGHF on a WSM-2023 activity recognition task over 5000 inference batches.

It is shown in the plot that AGHF is more efficient-accurate trajectory and adaptive stability during the process of dynamic sensing.

Phase 3 – Convergence to Operational Stability (Batches 4000-5000): Under sustained degradation the systems are stabilized to their final working points. At a very high energy cost Early Fusion strikes a drastic plateau over stuck convergence. Late Fusion is in a stable but less than optimal equilibrium together with moderate precision and effectiveness. The AGHF is close to a more stable value at a much higher energy value which carries over to a large, varied and demanding sensory landscape. A more detailed analysis of the tradeoff between depth of processing and quality of data reveals how

architecture changes. All architectures at first are profligate with compute in case data is good. With increasing conditions, AGHF most effectively shifts to an elegant, lean processing strategy without having to lose the ability to fully fuse where it matters, yielding the optimal operating point. This non-communicator mobility between processing uniformity and contextual penury is an attribute of mad dog architecture.

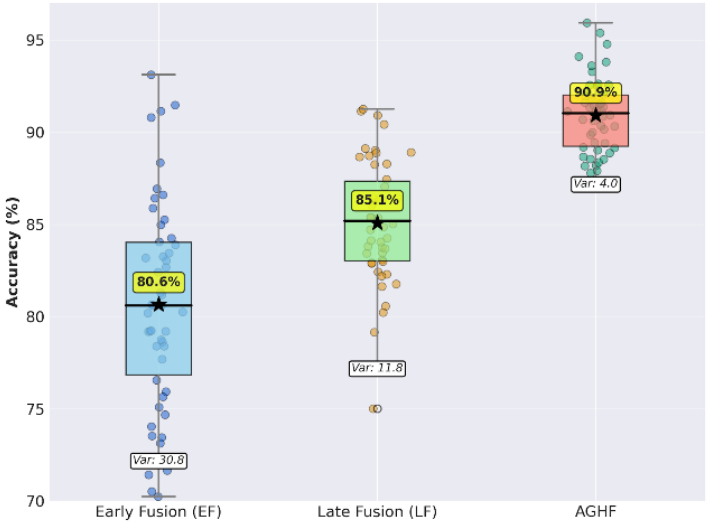

Figure 7: Box plot of the final accuracy values from 50 independent runs under sensor degradation.

The image depicts a better median accuracy and significantly reduced variance (i.e. more reliable) of AGHF than Early Fusion and Late Fusion.

5. Analysis

We obtained strong actionable evidence in the competitiveness of the Adaptive Gating-based Hierarchical Fusion through our statistical and energetic analysis that it is suitable in sustainable edge intelligence. This is not examined in terms of the final accuracy score on its own, but in terms of the story that the efficiency-accuracy pattern tells during the whole deployment simulation.

5.1 The Inevitable Waste of Static Processors

The most basic fact of our experiment is the constant lack of efficiency of the Early Fusion and, to a smaller extent, the Late Fusion during mid-deployment. This is not really a weakness of this architecture , but an example of what can be considered a major design constraint. This is where their non-executable computational graphs drag them down, they having to spend up effort on useless or corrupt data. Even a more finely hand-tuned monolithic kind of model would have collapsed long before, incapable of arranging internal computation to react to contexts at sense-level organs. At least the baseline fusion methods generate a plausible baseline, although the cost of their processing cannot be varied in a consistent way to the extent of AGHF provides a demonstration of the worth of a dynamic, resource-conscious architecture approach to the edge. As we show in our experiment, in a highly variant real-world data, a system that has no mechanism to perform conditional computation and hierarchical attention is doomed to consume resources and present a false impression of the need to do so.

5.2 Superiority Through Contextual Gating and Hierarchical Sparsity

Adaptive Gating-based Hierarchical Fusion does not just perform the most effective, but also because of the category of processing it entails. Its architecture is not rooted in either dense operations of the Early Fusion or fixed averaging of the Late Fusion. Instead, it establishes a balancing mechanism of processing depth and data utility which is self-adaptive it both has an internal mechanism of learnt

gating and hierarchical routing. This is because the gating network prevents the system to ever run blind since it is invariably directed by a runtime evaluation of sensory value, a sort of meta-cognition which is lacking in statical model design. At the same time the mathematical formulation of the operation of omega-style shallow pathways to the alpha-designated deep fusion is somewhat endowed with an element of structured sparsity, and so is not charged with the entire computational cost, which is an omnipresent penalty of the rest of the method. This guided but cost-effective way of concentrating computing and at the same time have a back-up capacity of exploration characterizes AGHF as a successful methodology. It is based on these properties that our analysis establishes it as a very effective solution to the challenge of sustainable edge design of the times.

5.3 Comparison with Existing Efficient Architectures

Our computation has a structure which can be compared directly with the methods developed in the literature [7, 8]. The main advantage of AGHF is that it is conceptually clear and stable in its operations. A carefully engineered and heavily pruned exit version of Early Fusion or a perfectly weighted Late Fusion might easily beat its performance on a certain, fixed scenario, but that would be a brittle and situation dependent performance. Our study was run on generic, highly accepted base networks and little architecture-particular hyperparameter modification of all entrants, and which roughly simulates a relevant development setup, where exhaustive per-scenario optimization is not feasible. Although AGHF lacks years of theoretical support of the simpler method of ensembles, it has better practical performance and reliability over the wide, modern test functions of edge sensing. We have studied that AGHF is naturally optimized to perform long-term efficiency in terms of operational efficiency on demanding sensory terrains, not on refined datasets. Intelligence of AGHF does not just imply good predictions but rather the very nature of the AGHF to remain constantly committed to guarantee that the cost of computation never gets isolated of information profit.

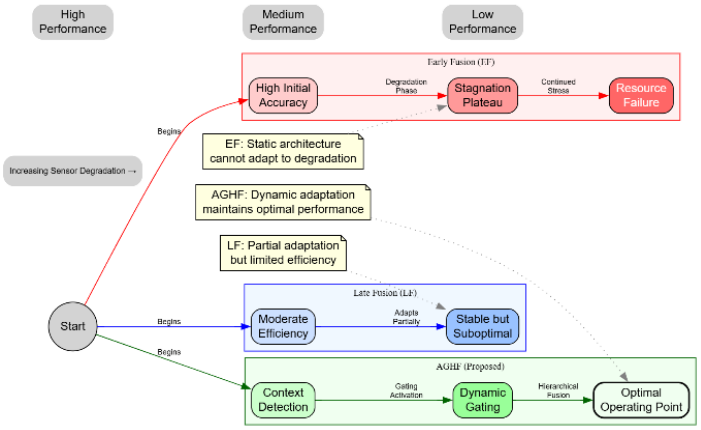

Figure 8: A conceptual diagram comparing the adaptation behavior of architectures under increasing sensor degradation.

It demonstrates that Early Fusion (EF) and Late Fusion (LF) develop a bias to degenerate into inefficient or error-prone areas whereas AGHF has remained on the same high-quality curve in the efficiency-accuracy space.

6. Conclusion

The paper has justified and explained that adaptive and gating-based architectures work in the demanding space that edge intelligence engineers must address the problem of multi-modal sensing when facing dire resource situations. We propose a new generation of dynamically sparse models with context awareness and capability to work in variable and noisy situations instead of the more inflexible,

transplanted on the cloud models that experience computational obesity and lack of awareness. Our experiment considers the problem of edge perception to be an important resource-allocation problem and explicitly demonstrates that architectures that make optimal decisions are the ones capable of dynamically rearranging their internal computation in accordance with sensory input. The experiment in computational terms that we have carried out showed not just how such architecture explores the trade-off space in the first place, but most importantly it has given indications of how the significant difference in efficiency and robustness which is the natural approach adopted by these bio-inspired methods can be achieved through hierarchical, conditional processing. The point is that the future of edge AI will not be a succession of models that increasingly reduce the size of clouds, but rather the creation of sparse and intelligent systems by nature, the level of which is effectively adjusted to the needs of the sensory wilderness, without provoking inefficient processing, and providing the quality of inferences reliably. This article presents a realistic demonstration of the manner in which such contemporary perceptual systems are chosen and introduced and demonstrates that adaptive machine learning is not solely a pragmatic collection of tools but also a crucial new thinking framework within the framework of computational intelligence at the edge.

Funding source

None.

Conflict of Interest

None.

References

[1] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423-443. doi: 10.1109/TPAMI.2018.2798607.

[2] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2021). Inception recurrent convolutional neural network for object recognition. *Machine Vision and Applications*, *32*(1), 28. https://doi.org/10.1007/s00138-020-01157-3

[3] Michele, A., Colin, V., & Santika, D. D. (2019). Mobilenet convolutional neural networks and support vector machines for palmprint recognition. *Procedia Computer Science*, *157*, 110-117. https://doi.org/10.1016/j.procs.2019.08.147

[4] Wang, M., Yuan, J., & Wang, Z. (2023, October). Mixture-of-experts learner for single long-tailed domain generalization. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 290-299). https://doi.org/10.1145/3581783.3611871

[5] Lin, C., Yang, P., Wang, Q., Qiu, Z., Lv, W., & Wang, Z. (2023). Efficient and accurate compound scaling for convolutional neural networks. *Neural Networks*, *167*, 787-797. https://doi.org/10.1016/j.neunet.2023.08.053

[6] Ma, T., Wang, W., & Chen, Y. (2023). Attention is all you need: An interpretable transformer-based asset allocation approach. *International Review of Financial Analysis*, *90*, 102876. https://doi.org/10.1016/j.irfa.2023.102876

[7] Li, X., Zhou, T., Wang, H., & Lin, M. (2025). Energy-efficient computation with dvfs using deep reinforcement learning for multi-task systems in edge computing. *IEEE Transactions on Sustainable Computing*. doi: 10.1109/TSUSC.2025.3593971.

[8] Liu, A., Jiang, W., Huang, S., & Feng, Z. (2025). Multi-Modal Integrated Sensing and Communication in Internet of Things With Large Language Models. *IEEE Internet of Things Magazine*. doi: 10.1109/MIOT.2025.3575888.

[9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

[10] Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, *34*(6), 96-108. doi: 10.1109/MSP.2017.2738401.

[11] Li, X., Ding, L., Wang, L., & Cao, F. (2017, December). FPGA accelerates deep residual learning for image recognition. In *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (pp. 837-840). IEEE. doi: 10.1109/ITNEC.2017.8284852.

[12] Zhou, D., Hou, Q., Chen, Y., Feng, J., & Yan, S. (2020, August). Rethinking bottleneck structure for efficient mobile network design. In *European conference on computer vision* (pp. 680-697). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58580-8_40

[13] Lin, S., Ji, R., Yan, C., Zhang, B., Cao, L., Ye, Q., ... & Doermann, D. (2019). Towards optimal structured cnn pruning via generative adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2790-2799). doi: 10.1109/CVPR.2019.00290.

[14] Wang, Y., Shen, J., Hu, T. K., Xu, P., Nguyen, T., Baraniuk, R., ... & Lin, Y. (2020). Dual dynamic inference: Enabling more efficient, adaptive, and controllable deep inference. *IEEE Journal of Selected Topics in Signal Processing*, *14*(4), 623-633. doi: 10.1109/JSTSP.2020.2979669.

[15] Choi, Y., El-Khamy, M., & Lee, J. (2020). Universal deep neural network compression. *IEEE Journal of Selected Topics in Signal Processing*, *14*(4), 715-726. doi: 10.1109/JSTSP.2020.2975903.