Inferência Adaptativa Multimodelo para Dispositivos de Borda com Recursos Limitados

João Pedro B. Lima[1,2], **Atslands R. Rocha**[1]

[1]Programa de Pós-Graduação em Engenharia de Teleinformática (PPGETI), Universidade Federal do Ceará (UFC)

[2]Instituto Atlântico

Abstract. *The application of machine learning models on edge devices requires strategies that balance accuracy and power consumption. In this work, we propose and evaluate a multi-model adaptive inference approach composed of convolutional neural networks to classify three classes in the CIFAR-10 dataset. The results showed that the adaptive strategy achieved higher accuracy than individual models when implemented on an ESP32 microcontroller. We observed that energy consumption is more closely related to the time the device remains active than to the complexity of the models. These results indicate that the approach can improve accuracy, but its energy savings depend on deployment conditions.*

Resumo. *A aplicação de modelos de aprendizado de máquina em dispositivos de borda exige estratégias que conciliem acurácia e consumo energético. Neste trabalho é proposta e avaliada uma abordagem de inferência adaptativa multimodelos composta por redes neurais convolucionais para classificar três classes do dataset CIFAR-10. Os resultados mostraram que a estratégia adaptativa alcançou maior acurácia que os modelos individuais quando implementada em um microcontrolador ESP32. Observou-se ainda que o consumo energético tem uma maior relação com o tempo em que o dispositivo permanece ativo do que com a complexidade dos modelos. Esses resultados indicam que a abordagem pode melhorar a acurácia, mas seus ganhos energéticos dependem das condições de implantação.*

1. Introdução

No cenário de *Edge Artificial Intelligence (Edge AI)*, modelos de aprendizado de máquina são executados diretamente em dispositivos de borda, reduzindo a dependência de processamento em nuvem e favorecendo aplicações com requisitos de baixa latência, menor consumo de banda e maior privacidade. Esse paradigma tem ganhado relevância em sistemas de Internet das Coisas (*Internet of Things*, IoT), onde dispositivos embarcados precisam tomar decisões localmente. Entretanto, nesses cenários, restrições de memória, de capacidade de processamento e de autonomia energética, naturais de dispositivos IoT, como a ESP32, impõem limites severos ao uso de modelos complexos [Lonkar and Malarkodi 2025], exigindo estratégias que permitam adaptar o custo da inferência às condições reais de execução. Abordagens que exploram múltiplos modelos com diferentes níveis de complexidade surgem como uma alternativa promissora para equilibrar desempenho e eficiência computacional, adicionando modelos mais robustos apenas quando necessário [Zawish et al. 2022] [Wang et al. 2023].

Uma das estratégias mais relevantes nessa direção é a inferência adaptativa baseada em confiança [Wang et al. 2023]. Nessa abordagem, a saída produzida por um modelo mais simples é avaliada por meio de uma métrica de confiança, como a margem entre as classes mais prováveis. Quando essa margem é suficientemente alta, a inferência é encerrada. Caso contrário, um modelo mais complexo é acionado para refinar a decisão. Trabalhos como o de [Wang et al. 2023] propõem esse tipo de mecanismo com foco na redução do consumo energético em dispositivos embarcados, indicando que a seleção progressiva de modelos pode representar uma alternativa eficiente à execução fixa de um único modelo de alta complexidade.

Entretanto, a adoção dessa estratégia em microcontroladores reais envolve desafios adicionais que nem sempre são capturados em formulações conceituais ou em ambientes experimentais mais controlados. Em plataformas com *clock* fixo, como a ESP32, o consumo energético pode estar mais diretamente relacionado ao tempo total em que o dispositivo permanece ativo fora do estado de *deep sleep* do que propriamente à complexidade teórica do modelo executado. Além disso, fatores práticos como limitações de memória, processo de quantização, custo de alternância entre modelos e características da biblioteca utilizada na implantação podem alterar significativamente o comportamento esperado da abordagem adaptativa.

Este trabalho propõe uma estratégia de inferência adaptativa multimodelo baseada na margem de confiança da predição. Os modelos são organizados em uma cadeia de complexidade crescente, e a cada etapa calcula-se a diferença entre as duas maiores probabilidades produzidas pelo modelo corrente. Quando essa margem supera um limiar previamente definido, a predição é aceita, caso contrário, a inferência avança para o próximo modelo da cadeia. Dessa forma, busca-se reduzir ativações desnecessárias de modelos mais custosos, preservando a acurácia quando a decisão de um modelo mais simples não é suficientemente confiável.

Os resultados obtidos mostram que a estratégia adaptativa foi capaz de alcançar acurácia superior à dos modelos individuais avaliados, evidenciando que a combinação entre modelos com diferentes capacidades pode produzir efeitos complementares na classificação. Por outro lado, observou-se que esse ganho não se traduziu, necessariamente, em melhor desempenho energético quando comparado à execução isolada do modelo mais complexo, o que revela limitações práticas importantes da abordagem no *setup* experimental adotado. Esses achados indicam que, em sistemas embarcados restritos, o projeto de mecanismos adaptativos deve considerar não apenas a qualidade preditiva dos modelos, mas também os custos adicionais introduzidos pela própria dinâmica de seleção e execução multimodelo.

Dessa forma, as principais contribuições deste trabalho são:

- a proposta de uma estratégia de inferência adaptativa multimodelo para execução em dispositivos embarcados, implementada em uma ESP32 com modelos de redes neurais convolucionais de diferentes complexidades;
- a avaliação experimental dos efeitos do tempo de inferência, do modo de baixo consumo *deep sleep* e da implantação dos modelos sobre o consumo energético do sistema;
- a investigação do compromisso entre acurácia e eficiência energética na abordagem proposta, destacando seus benefícios e limitações no cenário experimental

considerado.

2. Trabalhos Relacionados

A revisão sistemática realizada por [Gill et al. 2024] discute avanços recentes em *Edge AI*, com ênfase em estratégias de aprendizagem adaptativa e de eficiência energética. O artigo destaca tendências como compressão de modelos, aprendizado incremental e execução seletiva de modelos, todas voltadas à redução do consumo sem degradação significativa de desempenho. Essa revisão oferece um panorama teórico que justifica a abordagem adotada neste estudo, particularmente a estratégia de execução seletiva condicionada ao nível de confiança do modelo.

As propostas mais próximas à esse estudo são aquelas baseadas em inferência adaptativa, *early-exit* e seleção dinâmica de modelos. [Wang et al. 2023] propõem o Adaptive Resolution Inference (ARI), no qual uma inferência de menor custo é executada inicialmente, e a margem em relação ao limiar de decisão determina se uma versão mais completa do modelo deve ser acionada. Embora este trabalho adote inspiração semelhante no uso das margens de confiança, a abordagem aqui proposta difere por utilizar múltiplos modelos CNN independentes, com arquiteturas e custos distintos, ao invés de alternar apenas entre diferentes resoluções ou precisões de uma mesma arquitetura.

Trabalhos como [Korol and Beck 2025] exploram o conceito de *early-exit* associado à divisão IoT-Edge e a CNNs podadas, permitindo que determinadas entradas sejam classificadas em saídas intermediárias da rede. Essa classe de abordagem atua dentro da própria arquitetura neural, adicionando pontos de saída antecipada ou combinando *early-exit* com *pruning* e aceleração via FPGA. A estratégia proposta nesse artigo não modifica internamente os modelos com saídas intermediárias. A adaptação ocorre no nível da seleção entre modelos distintos, organizados em uma cadeia de complexidade crescente, o que permite avaliar o custo prático de acionar diferentes modelos em um microcontrolador restrito.

[Khan et al. 2025] investigam a seleção dinâmica do melhor modelo em um cenário de IoT voltado à detecção de incêndios, com foco em aumentar a adaptabilidade e a acurácia em diferentes condições de sensores. Em contraste, a proposta apresentada neste trabalho avalia uma estratégia progressiva de inferência para classificação de imagens, considerando explicitamente métricas de consumo energético, tempo de inferência, modo de baixo consumo e impacto da implantação dos modelos.

3. Metodologia

A proposta consiste em utilizar múltiplos modelos de redes neurais convolucionais com diferentes níveis de complexidade e selecionar dinamicamente qual modelo deve ser executado a partir da confiança associada à predição corrente. O experimento realizado possui o objetivo de avaliar a estratégia proposta de inferência adaptativa multimodelo em um dispositivo IoT alimentado por bateria. O desempenho da abordagem é analisado por meio de métricas de acurácia e consumo energético, sendo comparado ao de estratégias baseadas na execução individual de modelos fixos.

3.1. Estratégia adaptativa de seleção de modelos para inferência

A estratégia adaptativa baseia-se em um encadeamento progressivo de modelos classificadores, organizados em ordem crescente de complexidade, denominados M0, M1, M2 e

M3. Como detalhado na Figura 1, a inferência se inicia com a execução do modelo mais simples e, a partir de sua saída, calcula-se uma margem de confiança definida como a diferença entre os dois maiores valores produzidos pelo modelo. Essa margem é utilizada como critério de decisão para determinar se a classificação pode ser encerrada naquele estágio ou se é necessário acionar um modelo subsequente, de maior capacidade representacional.

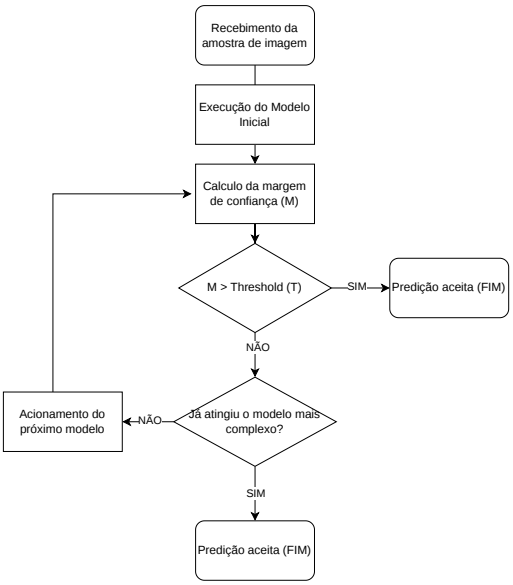

Figura 1. Fluxograma de execução da estratégia multimodelos

Quando a margem calculada é superior ao limiar adotado, a predição é considerada suficientemente confiável e o processo de inferência é finalizado. Caso contrário, o próximo modelo da sequência é executado. Esse procedimento é repetido até que a margem atenda ao critério de decisão ou até que o modelo mais complexo, M3, seja utilizado. O objetivo dessa estratégia é reduzir o número de ativações de modelos mais custosos, preservando a acurácia sempre que possível.

Embora a escolha do modelo inicial possa, conceitualmente, ser adaptada ao nível de energia disponível, essa seleção não foi utilizada no *setup* experimental avaliado, pois a lógica de decisão precisou ser executada externamente ao dispositivo. Assim, os experimentos com a abordagem adaptativa foram configurados para iniciar sempre a partir do modelo mais simples.

A decisão de troca entre modelos foi controlada por três valores de *threshold*: T100, T95 e T90. Essa definição seguiu a estratégia adotada por [Wang et al. 2023], na qual os limiares são determinados a partir dos casos em que o modelo corrente erra e o modelo subsequente acerta, permitindo calibrar o grau de permissividade da escalada entre modelos. O valor T100 corresponde ao maior valor de margem observado nos casos em que o modelo corrente errou e o seguinte acertou, enquanto T95 e T90 são versões mais permissivas, definidas para corrigir, respectivamente, 95% e 90% dessas divergências. Assim, valores menores de *threshold* tendem a manter a inferência em modelos mais simples por mais tempo.

3.2. Experimentos

Para a realização dos testes, foram utilizados dois módulos baseados no microcontrolador ESP32: (i) um TTGO LoRa com display OLED, responsável pela execução dos modelos de inferência; (ii) um ESP32 DevKit V1, utilizada como instrumento de monitoramento da tensão da bateria ao longo do experimento. Ambos os módulos foram alimentados por uma bateria de íon-lítio do tipo 18650, com capacidade nominal de 2600 mAh.

O monitoramento energético foi realizado por meio do conversor analógico-digital da ESP32 DevKit V1, utilizando um *firmware* específico para realizar leituras da tensão da bateria a cada 30 segundos e transmitir os valores via interface serial para um computador. Como a tensão nominal da bateria pode variar entre 0 e 4,2 V, ultrapassando a faixa operacional do ADC da ESP32, foi utilizado um divisor resistivo para adequar a tensão de entrada à faixa de 0 a 2,5 V, região em que o conversor apresenta comportamento mais linear. Além disso, testes de calibração com um voltímetro digital indicaram a necessidade de aplicar uma correção aditiva de 0,17 V aos valores lidos pela ESP32. Após essa correção, as medições mostraram compatibilidade com os valores do voltímetro até a primeira casa decimal, resultando em uma precisão estimada de 0,1 V.

Cabe ressaltar que o objetivo do experimento não é medir o consumo absoluto de energia com alta exatidão, mas sim comparar, em termos relativos, diferentes estratégias de inferência sob as mesmas condições físicas e operacionais. Por esse motivo, a resolução obtida mostrou-se suficiente para a análise proposta. Para garantir a consistência dos resultados, os mesmos dispositivos foram utilizados em todas as execuções do experimento, tanto na coleta das medições quanto na execução dos algoritmos.

3.3. Base de Dados

O conjunto de dados utilizado nos experimentos foi o CIFAR-10, um conjunto de imagens coloridas, amplamente usado em visão computacional, composto por 10 classes de objetos. Neste trabalho, foram selecionadas apenas três classes (*Dog, Truck e Shi*p) com o objetivo de simplificar a tarefa de classificação e viabilizar o uso de modelos menores, mais adequados às restrições de memória e de processamento da ESP32. Para compor o conjunto reduzido, foram utilizadas 5.000 imagens por classe para treinamento e 1.000 por classe para teste, totalizando 15.000 e 3.000 amostras, respectivamente. As imagens foram mantidas com resolução de 32×32 pixels e normalizadas para o intervalo [0,1].

3.4. Modelos Utilizados

Os modelos empregados nos experimentos são redes neurais convolucionais, escolhidas por permitirem arquiteturas com diferentes níveis de complexidade. Foram desenvolvidos quatro modelos distintos, denominados M0, M1, M2 e M3, com complexidade progressiva. As configurações dos modelos foram definidas da seguinte forma:

- M0: 1 camada convolucional e 1 camada de *Global Average Pooling*;
- M1: 2 camadas convolucionais e 1 camada totalmente conectada (*fully connected*);
- M2: 3 camadas convolucionais e 1 camada totalmente conectada;
- M3: 3 camadas convolucionais, 1 *batch* normalization e 1 *dropout*.

A Tabela 1 apresenta os dados de cada arquitetura, evidenciando o crescimento no número de parâmetros ao longo dos modelos: M1 possui aproximadamente 10 vezes

mais parâmetros que M0, M2 possui 4,5 vezes mais que M1 e M3, por sua vez, 1,1 vezes mais que M2. Também foi observado um aumento progressivo no uso de memória ROM entre os modelos.

Tabela 1. Tamanho de cada modelo

Modelo	Parâmetros	ROM (KB)
M0	499	3,3
M1	5187	23,08
M2	23747	95,96
M3	24131	98,14

3.5. Cálculo das Margens de Decisão

Como etapa preliminar à execução do algoritmo adaptativo, foram calculadas as margens de transição entre modelos consecutivos. Para isso, os modelos foram executados sobre o conjunto de teste, identificando-se os casos em que um modelo errava a classificação e o modelo seguinte acertava. Para cada uma dessas amostras divergentes, calculou-se a respectiva margem de confiança.

Em seguida, as margens obtidas foram organizadas em ordem decrescente, permitindo a extração dos valores mínimos necessários para atender aos *thresholds* definidos. Esse procedimento foi repetido para cada transição entre modelos consecutivos, resultando nos limiares utilizados durante a execução da estratégia adaptativa.

3.6. Avaliação do Consumo Energético

A avaliação do consumo energético foi realizada em duas etapas, uma para a avaliação do consumo de cada modelo sendo executado individualmente, e outra para a execução da abordagem adaptativa multimodelos. Na primeira etapa, o modelo foi executado em *loop* realizando a inferência de uma amostra por segundo, com e sem entrar no modo *deep sleep* entre cada inferência. O modo *deep sleep*, em que o microcontrolador suspende as atividades de todos os seus periféricos, permitiu evidenciar melhor a influência do tempo de inferência no consumo total do dispositivo.

Na segunda etapa, foi avaliada a abordagem adaptativa. Como a memória da ESP32 não permitia carregar simultaneamente as 3.000 amostras do conjunto de teste, a lógica adaptativa foi simulada externamente por um *script* em Python, que determinava, para cada amostra, a sequência de modelos acionada. Com base nessa sequência, a execução correspondente foi reproduzida no hardware, classificando uma amostrsera por vez e registrando o comportamento energético. Além disso, foram coletadas métricas complementares, como número de ativações de cada modelo, tempos de inferência e acurácia final, mantendo também o *deep sleep* de 500 ms entre inferências.

Para garantir reprodutibilidade e facilitar a medição energética, uma mesma amostra foi repetidamente usada durante a execução física dos caminhos previamente calculados, já que o foco dessa etapa era avaliar a sequência de ativações dos modelos, e não a classificação em si. A lógica adaptativa completa não pôde ser executada diretamente no ESP32 para todas as amostras, o que impediu a seleção dinâmica do modelo inicial com base no nível da bateria. Além disso, a implementação multimodelo precisou usar

a biblioteca padrão do TensorFlow Lite, pois a versão otimizada do Edge Impulse não permitia o embarque simultâneo de múltiplos modelos, aspecto que deve ser considerado na interpretação dos resultados energéticos.

Devido à limitação de memória da ESP32 para armazenar e percorrer em lote as 3.000 amostras do conjunto de teste, a lógica adaptativa foi executada previamente em um *script* Python, responsável por determinar a sequência de modelos acionada para cada amostra. Em seguida, essas sequências foram reproduzidas no hardware para medição energética. Essa escolha caracteriza uma avaliação baseada em rastros de execução, permitindo medir o impacto das ativações dos modelos sobre o consumo do dispositivo. Ressalta-se, entretanto, que essa limitação está associada ao protocolo experimental em lote, e não à impossibilidade de embarcar a lógica de decisão na ESP32

4. Resultados

Nesta seção, são apresentados os resultados da avaliação da estratégia de inferência adaptativa multimodelo proposta neste trabalho. A análise é organizada em quatro etapas. Inicialmente, são discutidas as margens de decisão calculadas para as transições entre modelos. Em seguida, são apresentados os resultados de consumo energético dos modelos individuais, considerando dois cenários experimentais distintos. Posteriormente, analisa-se o comportamento do algoritmo adaptativo em termos de frequência de ativações, consumo e tempo total de execução. Por fim, discute-se a acurácia final obtida pela estratégia adaptativa sob diferentes valores de *threshold*.

4.1. Margem de Decisão

Como etapa inicial, foram calculadas as margens de transição entre modelos consecutivos com base nos *thresholds* T100, T95 e T90. Para isso, os modelos foram executados sobre as 3.000 amostras do conjunto de teste, identificando-se os casos em que um modelo produzia uma predição incorreta e o modelo seguinte acertava a classificação. A partir dessas divergências, foram extraídos os valores mínimos de margem necessários para atender a cada um dos *thresholds* definidos. Os resultados são apresentados na Tabela 2.

Observa-se que as margens e o número de divergências variam entre as transições, indicando diferenças relevantes de comportamento entre os modelos da cadeia. De forma geral, os resultados confirmam que a escolha do threshold influencia diretamente a frequência com que a inferência é encerrada em modelos mais simples ou encaminhada para modelos mais complexos.

Tabela 2. Margens de transição encontradas

Transição	T_{100}	T_{95}	T_{90}	**Divergências**
M0 - M1	0.826690	0.471011	0.352986	694
M1 - M2	0.804802	0.486160	0.367679	370
M2 - M3	0.970064	0.842356	0.782334	712

4.2. Consumo de energia dos modelos individuais

Inicialmente, foi medida a descarga da bateria para cada modelo executado isoladamente. No cenário sem *deep sleep*, não foi observada diferença significativa de consumo entre modelos de complexidades distintas, como mostra a Figura 2, sugerindo que o custo

energético da plataforma como um todo exerce maior influência do que o custo computacional da inferência. Em seguida, foi adotado um segundo cenário com entrada em modo *deep sleep* entre inferências. Nessa configuração, observou-se uma redução clara no consumo ao utilizar modelos mais simples, indicando que, na ESP32, o tempo em que o dispositivo permanece ativo fora de *deep sleep* é um fator central para explicar o comportamento energético.

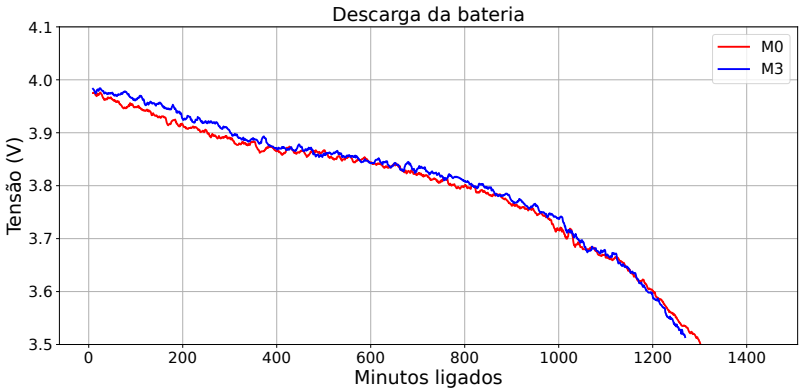

Figura 2. Descarga da Bateria para M0 e M3 sem deep-sleep

Esse resultado é particularmente importante porque mostra que, em microcontroladores como a ESP32, a variável mais relevante para o consumo energético não é apenas a complexidade abstrata do modelo, mas, sobretudo, o tempo durante o qual o dispositivo permanece ativo fora do modo *deep sleep*. Em outras palavras, quando o sistema alterna entre períodos de atividade e suspensão, o tempo de inferência passa a representar uma medida prática mais informativa para explicar o comportamento energético observado.

4.3. Algoritmo Adaptativo

A Tabela 3 apresenta a frequência de ativações de cada modelo para os diferentes valores de *threshold*. Observa-se que o M0 foi executado em todas as inferências e que o M1 foi acionado praticamente sempre, o que indica que o M0 raramente produz predições suficientemente confiáveis para encerrar a inferência na primeira etapa, acrescentando custo sem contribuir proporcionalmente para a decisão final. Dessa forma, sua permanência no pipeline adicionava tempo e consumo energético sem trazer redução proporcional no número de ativações posteriores. Com base nessa observação, o M0 foi descartado da configuração final da estratégia adaptativa, que passou a iniciar a inferência pelo M1. Observa-se, ainda, que *thresholds* mais permissivos reduzem o número de ativações dos modelos mais complexos, mantendo a inferência em modelos mais simples por mais tempo.

Tabela 3. Execuções por T

Modelo	T_{100}	T_{95}	T_{90}
M0	3000	3000	3000
M1	2999	2995	2993
M2	2487	2059	1791
M3	2013	1286	1004

Apesar disso, a Figura 3 mostra que, quando comparado à execução individual dos modelos, o algoritmo adaptativo apresentou consumo energético superior ao do modelo mais complexo executado isoladamente, isto é, o M3. Esse resultado contraria a expectativa inicial de que a seleção progressiva de modelos mais simples poderia reduzir o custo energético total da inferência.

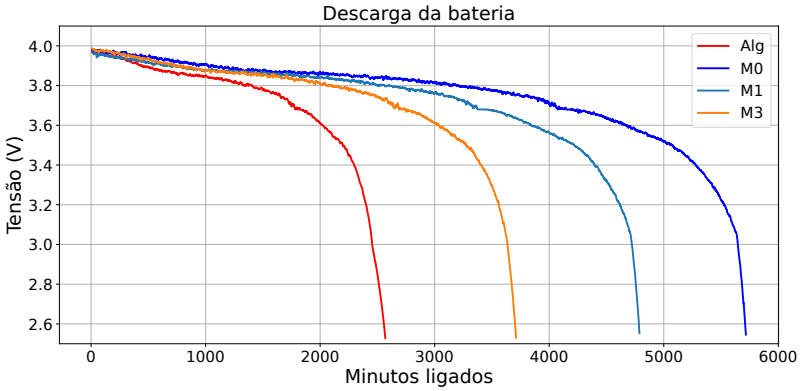

Figura 3. Descarga da Bateria do Algoritmo Adaptativo vs. Modelos individuais

Uma primeira interpretação para esse comportamento está associada ao processo de implantação dos modelos. Nos testes iniciais com modelos individuais, foi utilizada a plataforma Edge Impulse, cuja quantização produziu tempos de inferência menores. Entretanto, como essa biblioteca não permitia o embarque simultâneo de múltiplos modelos, a implementação do algoritmo adaptativo precisou utilizar a biblioteca padrão do TensorFlow Lite. Para evitar que a análise energética fosse atribuída exclusivamente à diferença de *toolchains*, foi incluído um experimento de controle no qual o modelo M3 foi embarcado utilizando a mesma biblioteca padrão do TensorFlow Lite empregada na abordagem adaptativa. Dessa forma, além da comparação com os modelos individuais otimizados pelo Edge Impulse, avaliou-se também o comportamento do modelo mais complexo sob a mesma infraestrutura de execução do algoritmo adaptativo. Os resultados apresentados na Figura 4 indicam que a ausência das otimizações do Edge Impulse aumenta o consumo energético, mas essa diferença não explica integralmente o consumo superior da abordagem adaptativa.

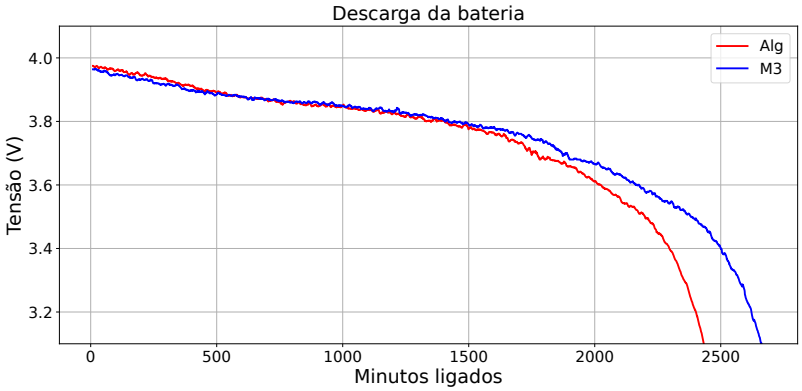

Figura 4. Descarga da Bateria do Algoritmo Adaptativo vs. M3

4.4. Tempo de Inferência e Consumo

Com o objetivo de investigar com maior precisão a relação entre consumo e tempo ativo, foi desenvolvido um procedimento específico para medir diretamente, no dispositivo, o tempo de inferência de cada modelo. Os resultados obtidos são apresentados na Tabela 4. O modelo M1 apresentou tempo médio de 283,1 ms por inferência, totalizando 849,3 segundos de atividade ao longo das 3.000 amostras. O modelo M2 apresentou tempo de 1023,8 ms por inferência, com total de 3071,4 segundos. Já o modelo M3 apresentou 1756,5 ms por inferência, totalizando 5269,5 segundos de tempo ativo.

Tabela 4. Tempo de inferência de cada modelo

Modelo	Tempo (ms)	Amostras	Total ligado (s)
M1	283,1	3000	849,3
M2	1023,8	3000	3071,4
M3	1756,5	3000	5269,5

A partir desses valores, e considerando ainda um tempo médio de 18,4 ms para alternância entre modelos, foi estimado o tempo total de execução do algoritmo adaptativo. Com base na quantidade de ativações e no número de transições previamente registradas, obteve-se um tempo total de 4539,4 segundos. Esse valor é inferior ao tempo total em que o dispositivo permanece ativo ao executar exclusivamente o modelo M3.

Embora o tempo ativo continue sendo uma variável central para explicar o consumo em microcontroladores com *clock* fixo, os resultados mostram que ele não foi suficiente, sozinho, para justificar o pior desempenho energético da abordagem adaptativa. Isso sugere a existência de outros custos associados à alternância entre modelos, possivelmente relacionados ao uso de memória RAM e flash, hipótese que permanece como investigação futura.

4.5. Acurácia Final do Algoritmo

Ao final da execução das 3.000 amostras do conjunto de teste, foi medida a acurácia do algoritmo adaptativo para os diferentes valores de threshold. Os resultados são apresentados na Tabela 5.

De forma geral, observa-se que o algoritmo adaptativo apresentou desempenho superior ao dos modelos individuais avaliados. Esse resultado sugere que a combinação entre modelos com diferentes capacidades de generalização permitiu explorar complementaridades entre eles, contribuindo para uma melhora global no desempenho classificatório. Trata-se de um achado importante do trabalho, pois indica que a estratégia

Tabela 5. Acurácia do algoritmo adaptativo e dos modelos individuais

Tipo de avaliação	Configuração	Acurácia (%)
Algoritmo adaptativo	T100	85,58
	T95	86,87
	T90	86,97
Modelo individual	M1	83,47
	M2	85,77
	M3	85,13

multimodelo pode produzir ganhos de acurácia mesmo em um ambiente embarcado com recursos restritos.

Além disso, observou-se um comportamento contraintuitivo: o melhor resultado de acurácia foi obtido com o *threshold* mais permissivo, T90. Uma explicação plausível para esse fenômeno é que valores menores de *threshold* reduzem a frequência de ativação do modelo mais complexo, M3. Como esse modelo apresentou degradação de desempenho após o processo de quantização, seu uso menos frequente pode ter contribuído, de forma paradoxal, para uma acurácia final mais elevada. Assim, neste experimento, acionar mais vezes o modelo mais complexo não significou necessariamente melhorar o desempenho final do sistema.

5. Conclusão

Este trabalho avaliou uma estratégia de inferência adaptativa multimodelo embarcada em uma ESP32, analisando os *trade-offs* entre acurácia e consumo energético em um cenário típico de IoT. Os resultados mostraram que, em plataformas com *clock* fixo, o consumo energético depende fortemente do tempo em que o sistema permanece ativo fora do modo *deep sleep*, e não apenas da complexidade do modelo. Também se observou que modelos muito simples, como o M0 no cenário avaliado, podem contribuir pouco para a eficiência, ao acrescentarem etapas sem reduzir de forma relevante o custo final da inferência.

Do ponto de vista da classificação, a estratégia adaptativa apresentou desempenho relevante, alcançando acurácia superior à dos modelos individuais utilizados como referência. O melhor resultado foi obtido com o *threshold* mais permissivo, T90, sugerindo que, neste *setup* experimental, acionar o modelo mais complexo com menor frequência pode ter sido vantajoso, possivelmente devido a efeitos introduzidos pela quantização.

Por outro lado, a abordagem adaptativa não superou, energeticamente, a execução isolada do modelo M3, o que indica que seus ganhos de eficiência não podem ser presumidos. Esses resultados indicam que a relação entre acurácia, tempo de execução e consumo energético depende fortemente das características de implementação. No caso avaliado, o M0 mostrou-se excessivamente simples para atuar como primeiro estágio efetivo da cadeia, sendo posteriormente descartado da configuração final do pipeline. Esse resultado reforça que a composição da cadeia adaptativa deve ser calibrada experimentalmente, considerando não apenas a complexidade individual dos modelos, mas também sua capacidade de encerrar inferências com confiança suficiente. Como trabalhos futuros, destacam-se a execução integral da lógica adaptativa no próprio dispositivo e a investigação mais detalhada do impacto das operações de memória no consumo energético. Pretende-se também realizar uma análise de *ablation* para medir separadamente o impacto da quantização, da escolha das arquiteturas e do conjunto de dados sobre o desempenho da cadeia adaptativa.

Referências

Gill, S. S., Golec, M., Hu, J., Xu, M., Du, J., Wu, H., Walia, G. K., Murugesan, S. S., Ali, B., Kumar, M., et al. (2024). Edge ai: A taxonomy, systematic review and future directions. *Cluster Computing*, 28(1):18.

Khan, M. A., Song, W., Khan, A., Ali, M., Karim, R., and Zhang, J. (2025). Machine learning hybrid dynamic best model selection algorithm for real-time fire prediction using

iot-enabled multi-sensor data in buildings. *Journal of Safety Science and Resilience*, page 100236.

Korol, G. and Beck, A. C. S. (2025). Iot–edge splitting with pruned early-exit cnns for adaptive inference. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

Lonkar, L. K. and Malarkodi, B. (2025). Energy-aware deep neural networks with intelligent adaptive inference for sustainable iot deployments. In *2025 IEEE 6th India Council International Subsections Conference (INDISCON)*, pages 1–6.

Wang, Z., Reviriego, P., Niknia, F., Conde, J., Liu, S., and Lombardi, F. (2023). Adaptive resolution inference (ari): Energy-efficient machine learning for internet of things. *IEEE Internet of Things Journal*, 11(8):14076–14087.

Zawish, M., Ashraf, N., Ansari, R. I., and Davy, S. (2022). Energy-aware ai-driven framework for edge-computing-based iot applications. *IEEE Internet of Things Journal*, 10(6):5013–5023.