FABRÍCIO DANIEL FREITAS

SMARTPHONES COMO PLATAFORMA DE INTELIGÊNCIA ARTIFICIAL DE BORDA PARA ASSISTÊNCIA AO **CONDUTOR:** ANÁLISE DE DESEMPENHO COMPUTACIONAL, TÉRMICO E ENERGÉTICO.

LAVRAS – MG
2025

FABRÍCIO DANIEL FREITAS

SMARTPHONES COMO PLATAFORMA DE INTELIGÊNCIA ARTIFICIAL DE BORDA PARA ASSISTÊNCIA AO CONDUTOR: ANÁLISE DE DESEMPENHO COMPUTACIONAL, TÉRMICO E ENERGÉTICO.

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Sistemas de Computação, para a obtenção do título de Mestre.

Prof. Dr. Arthur de Miranda Neto
Orientador

LAVRAS – MG
2025

Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Freitas, Fabrício Daniel

Smartphones como plataforma de inteligência artificial de borda para assistência ao condutor : Análise de desempenho computacional, térmico e energético. / . – Lavras : UFLA, 2025.

79p. : il.

Dissertação(Mestrado Acadêmico)–Universidade Federal de Lavras, 2025.

Orientador: Prof. Dr. Arthur de Miranda Neto.

Bibliografia.

1. Sistemas avançados de assistência ao condutor. 2. Inteligência artificial de borda. 3. Dispositivos móveis. 4. Aprendizado profundo. 5. Processamento em tempo real. I. de Miranda Neto, Arthur. II. Universidade Federal de Lavras. III. Título

FABRÍCIO DANIEL FREITAS

SMARTPHONES COMO PLATAFORMA DE INTELIGÊNCIA ARTIFICIAL DE BORDA PARA ASSISTÊNCIA AO CONDUTOR: ANÁLISE DE DESEMPENHO COMPUTACIONAL, TÉRMICO E ENERGÉTICO.

SMARTPHONES AS AN EDGE ARTIFICIAL INTELLIGENCE PLATFORM FOR DRIVER ASSISTANCE: COMPUTATIONAL, THERMAL, AND ENERGY PERFORMANCE ANALYSIS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, área de concentração em Sistemas de Computação, para a obtenção do título de Mestre.

APROVADA em 21 de julho de 2025.

Prof. DSc. Gustavo Lobato Campos	AUTI/IFMG-Campus Betim
Prof. DSc. Luiz Henrique Andrade Correa	DCC/UFLA

Prof. Dr. Arthur de Miranda Neto
Orientador

LAVRAS – MG
2025

Dedico este trabalho, com amor e gratidão, à memória dos meus tios Antônio e Aparecida, cuja generosidade e incentivo foram fundamentais para minha formação e educação. Ao meu pai, que já não está entre nós, mas cujo orgulho e confiança em meu potencial sempre me impulsionaram a seguir em frente, deixo minha eterna admiração e saudade. E, especialmente, à minha mãe, por sua presença constante, apoio incondicional e carinho incansável em todos os momentos da minha vida. Sem vocês, esta conquista não seria possível.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, pela dádiva da vida, pela saúde e pela força concedida em cada etapa desta caminhada acadêmica. Sem a fé e o amparo espiritual, seria impossível vencer tantos desafios.

Sou profundamente grato à minha família, base fundamental de todo o meu percurso. Desde a infância, sempre encontrei em vocês encorajamento inabalável e apoio irrestrito, que foram essenciais não apenas para os meus estudos, mas para a formação dos valores e do caráter que carrego comigo. Todo o conhecimento e todas as conquistas até aqui alcançadas pertencem também a vocês.

Ao Professor Arthur, registro minha gratidão pela orientação dedicada, pelos ensinamentos e pela disposição em compartilhar seu conhecimento ao longo de toda esta jornada. Sua orientação criteriosa foi fundamental para a realização deste trabalho.

Estendo meu reconhecimento ao Instituto Federal de Minas Gerais (IFMG), instituição que tem sido pilar em minha trajetória acadêmica desde a graduação. Em especial, agradeço à equipe do Setor de Laboratórios, pela constante colaboração e espírito de equipe, bem como à equipe do Polo de Inovação, pela compreensão e flexibilidade diante dos desafios impostos pela minha rotina entre Lavras e Formiga.

Manifesto ainda minha gratidão à Universidade Federal de Lavras (UFLA), por proporcionar não apenas a oportunidade de cursar este mestrado, mas também um ambiente fértil para o crescimento acadêmico e profissional.

À minha amiga Luísa Carolina Silva, reservo um agradecimento especial pela generosidade de, com tanto carinho, ter me acolhido em Lavras durante este período. Seu gesto foi imprescindível para tornar possível a realização desta pesquisa.

Aos amigos, minha sincera apreciação pela compreensão, paciência e apoio nos momentos em que precisei ausentar-me do convívio social para dedicar-me integralmente a esta etapa. A amizade e o incentivo de vocês foram essenciais para que eu encontrasse motivação e equilíbrio ao longo desta trajetória.

Por fim, a todos que, de alguma forma, contribuíram direta ou indiretamente para a concretização deste trabalho e para o meu amadurecimento, deixo o meu sincero agradecimento.

Ora, se algum de vós tem falta de sabedoria, peça-a a Deus, que a todos dá liberalmente e não censura, e ser-lhe-á dada.

(Tiago 1:5)

RESUMO

Os Sistemas Avançados de Assistência ao Condutor (ADAS) têm ganhado relevância crescente na indústria automotiva, com o objetivo de aprimorar a segurança e o conforto na condução. Paralelamente, o avanço da Inteligência Artificial (IA) e a evolução dos smartphones abriram novas oportunidades para implementar essas tecnologias diretamente em dispositivos móveis, promovendo uma solução acessível e economicamente viável. Esta dissertação analisa o desempenho computacional, térmico e energético de algoritmos de aprendizado profundo aplicados em smartphones para ADAS, levando em consideração as limitações de recursos e as exigências de processamento em tempo real. Além da revisão sistemática da literatura, realizou-se uma análise experimental utilizando benchmarks padronizados (AI Benchmark e Burnout Benchmark) e uma aplicação ADAS desenvolvida especificamente para este estudo. Os experimentos abrangeram dispositivos de diferentes categorias (*premium*, intermediário e básico), destaca-se diferenças significativas no desempenho sustentado, eficiência energética e capacidade de inferência em tempo real. Os resultados demonstraram que smartphones equipados com aceleradores de hardware dedicados, como GPUs e NPUs, apresentam desempenho superior e maior eficiência em cenários realistas de uso. Por fim, apresentou-se recomendações práticas para otimização técnica dos modelos visando melhorar sua viabilidade em aplicações reais, contribuindo diretamente para avanços na segurança veicular e computação móvel.

Palavras-chave: sistemas avançados de assistência ao condutor; inteligência artificial de borda; dispositivos móveis; aprendizado profundo; processamento em tempo real; análise de desempenho.

ABSTRACT

Advanced Driver Assistance Systems (ADAS) have gained increasing relevance in the automotive industry, aiming to improve driving safety and comfort. At the same time, advances in Artificial Intelligence (AI) and the evolution of smartphones have opened new opportunities to implement these technologies directly on mobile devices, promoting an accessible and economically viable solution. This dissertation analyzes the computational, thermal, and energy performance of deep learning algorithms applied to smartphones for ADAS, taking into account resource limitations and real-time processing requirements. In addition to a systematic literature review, an experimental analysis was conducted using standardized benchmarks (AI Benchmark and Burnout Benchmark) and an ADAS application developed specifically for this study. The experiments covered devices from different categories (premium, mid-range, and entry-level), highlighting significant differences in sustained performance, energy efficiency, and real-time inference capability. The results demonstrated that smartphones equipped with dedicated hardware accelerators, such as GPUs and NPUs, exhibit superior performance and greater efficiency in realistic usage scenarios. Finally, practical recommendations were presented for technical optimization of the models to improve their viability in real applications, directly contributing to advances in vehicle safety and mobile computing.

Keywords: advanced driver assistance systems; edge artificial intelligence; mobile devices; deep learning; real-time processing; performance analysis.

INDICADORES DE IMPACTO

Esta dissertação de mestrado investiga o potencial dos smartphones como plataforma de inteligência artificial de borda para assistência ao condutor, analisando o desempenho, a viabilidade e as limitações dessa abordagem para sistemas avançados de assistência ao condutor (ADAS). Ao propor o uso de dispositivos móveis amplamente acessíveis, a pesquisa destaca a democratização da tecnologia de assistência veicular, ao proporcionar acesso a recursos de segurança antes restritos a veículos de alto padrão. O impacto tecnológico reside na caracterização e avaliação de arquiteturas de redes neurais profundas otimizadas para smartphones, bem como na análise do desempenho de diferentes processadores dedicados presentes nesses dispositivos, evidenciando seu potencial para aplicações em tempo real no contexto da mobilidade urbana.

No âmbito social, este estudo oferece bases para o desenvolvimento de soluções de baixo custo com potencial de beneficiar condutores, passageiros e pedestres, especialmente em regiões onde a frota de veículos com sistemas embarcados ainda é limitada. O uso de smartphones para ADAS pode contribuir para a redução de acidentes, promove maior segurança no trânsito e alinha-se ao ODS 3 – Saúde e Bem-estar e ao ODS 11 – Cidades e Comunidades Sustentáveis. Do ponto de vista econômico, a viabilidade demonstrada para uso de dispositivos já presentes no cotidiano dos brasileiros representa uma oportunidade de redução de custos para implementação de sistemas de assistência, o que pode estimular o mercado nacional de aplicativos e fomentar o desenvolvimento de soluções inovadoras na cadeia automotiva e tecnológica.

Este trabalho também apresenta caráter extensionista, pois estabelece uma conexão entre a universidade, profissionais do setor de tecnologia e transporte, e a sociedade em geral, estimulando futuras parcerias com instituições de ensino, empresas e órgãos públicos para o desenvolvimento e disseminação dessas tecnologias. A pesquisa contribui para a área temática de Tecnologia e Produção com extensões possíveis à Educação, ao fomentar o debate sobre segurança veicular e transformação digital no contexto brasileiro. O público potencialmente beneficiado abrange motoristas, empresas de transporte, órgãos reguladores e a população urbana, que poderá dispor de maior segurança por meio do uso ampliado das tecnologias analisadas. Assim, o trabalho realizado oferece uma base para avanços futuros em pesquisas, políticas públicas e iniciativas inovadoras, contribuindo para o desenvolvimento sustentável e para o fortalecimento da indústria nacional de tecnologia aplicada à mobilidade segura.

IMPACT INDICATORS

This master's thesis investigates the potential of smartphones as an edge artificial intelligence platform for driver assistance, analyzing the performance, feasibility, and limitations of this approach for advanced driver assistance systems (ADAS). By proposing the use of widely accessible mobile devices, the research highlights the democratization of vehicle assistance technology by providing access to safety features previously restricted to high-end vehicles. The technological impact lies in the characterization and evaluation of deep neural network architectures optimized for smartphones, as well as in the analysis of the performance of different dedicated processors present in these devices, highlighting their potential for real-time applications in the context of urban mobility.

In the social sphere, this study provides a basis for the development of low-cost solutions with the potential to benefit drivers, passengers, and pedestrians, especially in regions where the fleet of vehicles with on-board systems is still limited. The use of smartphones for ADAS can contribute to the reduction of accidents, promotes greater road safety, and is aligned with SDG 3 – Good Health and Well-being and SDG 11 – Sustainable Cities and Communities. From an economic standpoint, the demonstrated viability of using devices that are already present in the daily lives of Brazilians represents an opportunity to reduce costs for implementing assistance systems, which can stimulate the national application market and foster the development of innovative solutions in the automotive and technology chains.

This work also has an extensionist nature, as it establishes a connection between universities, professionals in the technology and transportation sector, and society in general, encouraging future partnerships with educational institutions, companies, and public agencies for the development and dissemination of these technologies. The research contributes to the thematic area of Technology and Production with possible extensions to Education, by fostering the debate on vehicle safety and digital transformation in the Brazilian context. The potential beneficiaries include drivers, transportation companies, regulatory agencies, and the urban population, who will be able to enjoy greater safety through the expanded use of the technologies analyzed. Thus, the work carried out provides a basis for future advances in research, public policies, and innovative initiatives, contributing to sustainable development and to the strengthening of the national technology industry applied to safe mobility.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

ms milissegundo(s)

LISTA DE SIGLAS

ADAS	Sistemas Avançados de Assistência ao Condutor (Advanced Driver Assistance Systems)
ANTT	Agência Nacional de Transportes Terrestres
APUs	Unidades de Processamento Acelerado (Accelerated Processing Units)
CNN	Rede Neural Convolucional (Convolutional Neural Network)
CPU	Unidade Central de Processamento (Central Processing Unit)
cuDNN	CUDA Deep Neural Network library
DL	Aprendizado Profundo (Deep Learning)
DSP	Processador de Sinal Digital (Digital Signal Processor)
Edge AI	Inteligência Artificial de Borda (Edge Artificial Intelligence)
FPGAs	Arranjos de Portas Programáveis em Campo (Field-Programmable Gate Arrays)
FPS	Quadros por Segundo (Frames Per Second)
GPS	Sistema de Posicionamento Global (Global Positioning System)
GPU	Unidade de Processamento Gráfico (Graphics Processing Unit)
IA	Inteligência Artificial (Artificial Intelligence)
IoT	Internet das Coisas (Internet of Things)
IoV	Internet dos Veículos (Internet of Vehicles)
Ipea	Instituto de Pesquisa Econômica Aplicada
JAX	JAX (*Framework* para *Machine Learning*)
MKL-DNN	Math Kernel Library for Deep Neural Networks
ML	Aprendizado de Máquina (Machine Learning)
NLP	Processamento de Linguagem Natural (Natural Language Processing)
NN	Rede Neural (Neural Network)
NNAPI	Neural Networks API
NPU	Unidade de Processamento Neural (Neural Processing Unit)

OLED	Diodo Orgânico Emissor de Luz (do inglês, Organic Light Emitting Diode). Tecnologia de tela que permite que cada pixel emita sua própria luz, resultando em pretos mais profundos, contraste superior e cores mais vibrantes em comparação com tecnologias LCD tradicionais.
ONNX	Open Neural Network Exchange
ONU	Organização das Nações Unidas
PTQ	Quantização Pós-Treinamento (Post-Training Quantization)
QAT	Quantização Consciente de Treinamento (Quantization Aware Training)
QNN HTP	Qualcomm Neural Network Hexagon Tensor Processor
SAE	Sociedade dos Engenheiros Automotivos (Society of Automotive Engineers)
SBCs	Computadores de Placa Única (Single Board Computers)
SDK	Kit de Desenvolvimento de Software (Software Development Kit)
SoC	Sistema em um Chip (System on a Chip)
TPUs	Unidades de Processamento Tensorial (Tensor Processing Units)
TVM	Tensor Virtual Machine
VPUs	Unidades de Processamento Visual (Vision Processing Units)
YOLO	You Only Look Once

LISTA DE SÍMBOLOS

FPS	Quadros Por Segundo (do inglês, Frames Per Second). Métrica que indica a frequência com que imagens consecutivas são exibidas por um dispositivo, sendo crucial para a fluidez visual em vídeos e aplicações em tempo real, como sistemas de assistência ao condutor.
nm	nanômetro (unidade de comprimento)
GHz	Gigahertz (unidade de frequência)

SUMÁRIO

1 INTRODUÇÃO

A segurança no trânsito representa um dos desafios mais prementes da sociedade contemporânea, com impactos significativos na saúde pública, economia e bem-estar social. Segundo dados recentes do Instituto de Pesquisa Econômica Aplicada (Ipea), mais de um milhão de vidas são perdidas anualmente em todo o mundo devido a acidentes de trânsito (Carvalho, 2023). No Brasil, a situação é significativa, com mais de 50.000 fatalidades anuais, conforme relatado pela ANTT (Agência Nacional de Transportes Terrestres, 2023).

Uma pesquisa realizada pelo Ipea indica que, no período entre 2010 e 2019, o Brasil apresentou um incremento de 13,5% nas fatalidades decorrentes de acidentes de trânsito, o que vai em sentido contrário à meta estabelecida globalmente pela Organização das Nações Unidas (ONU), que era de diminuir em 50% as mortes no trânsito até o ano de 2020 (Carvalho, 2023). Em 2006, falhas humanas eram responsáveis por mais de 90% dos acidentes, e essa cifra continua quase inalterada. O uso excessivo de celulares durante a condução tornou-se um dos principais fatores de risco. Conforme os dados de 2024 evidenciados pelo Centro de Documentação e Memória do Mercado Segurador, "a falta de atenção causou 42% dos acidentes. Dentre eles, houve 28.522 incidentes atribuídos a reações tardias dos motoristas (10.912), falta total de reação (10.658), e entrada na via sem verificar a presença de outros veículos (6.952)"(CEDOM, 2025).

Outro fator relevante é a saúde mental dos motoristas, um tema cada vez mais discutido. O estresse, a fadiga e a rotina acelerada comprometem a atenção e a tomada de decisões ao volante, aumentando os riscos de colisões. Além das perdas humanas, os acidentes de trânsito geram custos superiores a 50 bilhões de reais por ano, impactando significativamente a economia em gastos com previdência, redução de renda familiar, custos hospitalares e danos patrimoniais. Em comparação com outros países, o Brasil apresenta uma taxa de mortalidade no trânsito de 20 por 100.000 habitantes, significativamente superior à média de países desenvolvidos, com menos de 3 mortes por 100.000 habitantes em países como Dinamarca, Noruega, Suécia, Reino Unido e Islândia (Mundo Logística, 2024).

Para mitigar essa problemática, tecnologias emergentes como os Sistemas Avançados de Assistência ao Condutor (Advanced Driver Assistance Systems - ADAS) têm sido desenvolvidas e implementadas. Os ADAS incorporam funcionalidades como frenagem de emergência automática, controle de cruzeiro adaptativo e assistência de manutenção de faixa, visando au-

mentar a segurança e a eficiência na condução (Renault, 2024). Contudo, a implementação generalizada desses sistemas enfrenta desafios significativos, principalmente relacionados aos custos elevados e à dependência de hardware especializado (Weig, 2016).

No presente trabalho, o escopo se limita ao Nível 1 da classificação SAE J3016 para sistemas ADAS, concentrando-se exclusivamente na detecção de objetos em tempo real por meio de smartphones, sem qualquer interferência ou automação sobre os comandos do veículo. Assim, a proposta visa apenas o apoio informativo ao condutor, oferecendo alertas ou monitoramento passivo, sem executar ações automáticas no automóvel. Essa delimitação é fundamental para deixar claro que o objetivo deste estudo é analisar o potencial dos smartphones como ferramenta de apoio passivo à direção, por meio de técnicas de inteligência artificial embarcada.

Nesse contexto, a Inteligência Artificial de Borda (Edge Artificial Intelligence - Edge AI) emerge como uma tecnologia promissora para superar esses obstáculos. A Edge AI permite a execução de algoritmos de inteligência artificial diretamente nos dispositivos, reduzindo a latência, melhorando a eficiência energética e aumentando a privacidade dos dados (Chen *et al.*, 2023). Particularmente, a aplicação da Edge AI em smartphones tem ganhado destaque devido à crescente capacidade de processamento desses aparelhos e à integração de sensores sofisticados (Ignatov *et al.*, 2018).

Os smartphones modernos estão equipados com uma variedade de sensores, incluindo acelerômetro, giroscópio, câmera e receptor do Sistema de Posicionamento Global (Global Positioning System - GPS), que permitem a coleta de dados em tempo real sobre o comportamento do usuário e as condições ambientais. O avanço nos modelos de Aprendizado Profundo (Deep Learning - DL) e a otimização das arquiteturas de redes neurais para dispositivos móveis ampliaram a capacidade desses aparelhos em executar tarefas complexas de reconhecimento de padrões, análise de imagens e tomada de decisões (Ignatov *et al.*, 2018). Por exemplo, smartphones podem ser utilizados para monitorar o comportamento do motorista, detectar sinais de fadiga ou distração e até mesmo complementar sistemas ADAS existentes por meio de alertas sonoros ou visuais (Theivadas; Ponnan, 2024).

Este estudo insere-se no contexto atual de rápida evolução tecnológica dos smartphones, que nos últimos anos têm apresentado capacidades de processamento e sensoriamento cada vez mais avançadas. Essa evolução abre novas possibilidades para aplicações complexas como os ADAS, tradicionalmente dependentes de hardware especializado e custoso. A utilização de smartphones como plataforma para ADAS pode ter um impacto significativo, especialmente

em países em desenvolvimento como o Brasil, onde a renovação da frota de veículos é mais lenta e o acesso a tecnologias avançadas de segurança veicular é limitado.

No contexto atual da pesquisa em Edge AI e ADAS, vários estudos têm explorado o potencial dos smartphones para aplicações de segurança veicular. Por exemplo, Noronha *et al.* (2019) investigaram o uso de dispositivos móveis para detecção de fadiga do motorista, enquanto Do *et al.* (2016) exploraram a integração destes com sistemas veiculares existentes para melhorar a percepção do ambiente. No entanto, ainda existem lacunas significativas na compreensão das capacidades e limitações dos smartphones modernos para executar algoritmos de DL em tempo real para aplicações ADAS. Esta pesquisa visa preencher essas lacunas, fornecendo análise do desempenho de smartphones em tarefas relevantes para ADAS.

Este trabalho propõe uma metodologia para avaliar a eficiência de smartphones na execução de algoritmos de DL, utilizando a AI Benchmark (Zurich, 2024a) como principal ferramenta de análise comparativa. A AI Benchmark destaca-se como uma plataforma especializada na avaliação do desempenho de dispositivos móveis em tarefas críticas de aprendizado profundo, incluindo reconhecimento de imagem (Sandler *et al.*, 2019; Szegedy *et al.*, 2015; Bochkovskiy; Wang; Liao, 2020), segmentação semântica (Chen *et al.*, 2018a), estimação de profundidade (Zhang *et al.*, 2021; Ignatov *et al.*, 2021) e reconhecimento facial (Howard *et al.*, 2019). A ferramenta executa uma série de testes diretamente no dispositivo, empregando redes neurais como MobileNet (Sandler *et al.*, 2019; Howard *et al.*, 2019), Inception-V3 (Szegedy *et al.*, 2015), YOLOv4-Tiny (Bochkovskiy; Wang; Liao, 2020), DeepLab-V3+ (Chen *et al.*, 2018a) e MV3-Depth (Zhang *et al.*, 2021; Ignatov *et al.*, 2021), permitindo uma análise da capacidade de processamento, gerenciamento de memória e eficiência energética.

Para garantir uma avaliação mais robusta e abrangente, a metodologia incorpora também a Burnout Benchmark (Zurich, 2024b), uma ferramenta especializada na análise do desempenho térmico e energético de dispositivos móveis. Essa ferramenta realiza uma avaliação sistemática dos principais componentes de Sistemas em Chip (Systems on Chip - SoCs), sendo: Unidade Central de Processamento (Central Processing Unit - CPU), Unidade de Processamento Gráfico (Graphics Processing Unit - GPU), Unidade de Processamento Neural (Neural Processing Unit - NPU) e Processador de Sinal Digital (Digital Signal Processor - DSP), sob diferentes condições de carga. Por meio de testes isolados e combinados desses componentes, a Burnout Benchmark simula cenários de uso intensivo, proporcionando percepções sobre o comportamento térmico do dispositivo e sua capacidade de manter desempenho sustentável sob demanda contínua.

A integração estratégica dessas duas ferramentas estabelece uma estrutura de avaliação holística, que permite não apenas quantificar o desempenho dos smartphones em tarefas de DL, mas também compreender as implicações práticas de sua implementação em aplicações ADAS. Essa abordagem dual possibilita uma análise mais precisa da viabilidade técnica de utilizar smartphones como plataforma para sistemas de assistência ao condutor, considerando tanto aspectos de desempenho computacional quanto limitações físicas do hardware.

1.1 Objetivos

Este trabalho tem como objetivo principal avaliar e analisar a eficiência de smartphones na execução de algoritmos de DL - Edge AI, com foco no processamento de dados em tempo real, visando estabelecer uma compreensão das capacidades e limitações desses dispositivos em tarefas computacionalmente intensivas, bem como investigar sua viabilidade como plataforma alternativa para futuras implementações de ADAS, contribuindo assim para a democratização dessas tecnologias de forma acessível e economicamente viável. Para tanto, os seguintes objetivos específicos foram definidos:

a) realizar uma análise comparativa do desempenho de diferentes smartphones Android na execução de algoritmos de DL com a AI Benchmark;
b) avaliar o desempenho térmico e energético dos dispositivos utilizando a Burnout Benchmark;
c) avaliar comparativamente o desempenho de SoCs móveis em tarefas de DL;
d) identificar e caracterizar os gargalos técnicos e limitações operacionais dos smartphones na execução de algoritmos de DL;
e) desenvolver recomendações técnicas para otimização de modelos de DL específicos para execução em smartphones, visando aplicações ADAS, considerando o equilíbrio entre precisão, velocidade e consumo de recursos.

1.2 Organização do trabalho

Este documento encontra-se organizado da seguinte maneira:

A Capítulo 1 apresenta a justificativa, o contexto e os objetivos deste estudo, destacando a importância da avaliação de smartphones para aplicações ADAS utilizando Edge AI.

A Capítulo 2 aborda os fundamentos teóricos necessários para a compreensão do trabalho. São discutidos conceitos de Edge AI, arquiteturas de redes neurais profundas, sistemas ADAS e as tecnologias presentes em smartphones modernos relevantes para o processamento de algoritmos de DL.

A Capítulo 3 apresenta uma revisão inicial da literatura sobre implementações de ADAS em dispositivos móveis e análises de desempenho de DL em smartphones, identificando lacunas e oportunidades de pesquisa.

A Capítulo 4 detalha o planejamento metodológico para a execução da pesquisa. São descritas as ferramentas de avaliação (AI Benchmark e Burnout Benchmark), os critérios preliminares para seleção de dispositivos e as métricas de avaliação a serem consideradas.

A Capítulo 5 apresenta os resultados obtidos, incluindo análises e discussões dos dados coletados durante a pesquisa.

Por fim, o Capítulo 6 elucida os resultados e possibilidades de trabalhos futuros baseados no que foi desenvolvido.

Apêndices e anexos são incluídos para fornecer informações complementares, como especificações técnicas detalhadas dos dispositivos testados e detalhes adicionais sobre os benchmarks utilizados.

2 REFERENCIAL TEÓRICO

Nesta seção, realizou-se uma revisão dos conceitos fundamentais que sustentam este trabalho e é apresentada a base teórica para o estudo da aplicabilidade de algoritmos de DL em conjunto com smartphones em ADAS.

2.1 Sistemas Avançados de Assistência ao Condutor

A Sociedade dos Engenheiros Automotivos (Society of Automotive Engineers - SAE), por meio da norma SAE J3016 (2021), estabelece uma classificação de seis níveis de automação na condução, variando desde a ausência de automação (Nível 0) até a automação plena (Nível 5). Nos níveis 1 e 2, os veículos são equipados com tecnologias que auxiliam o motorista em tarefas específicas, como controle de cruzeiro adaptativo, assistência de permanência em faixa e frenagem automática de emergência. Statista (2024) ressalta que esses sistemas, denominados ADAS, aprimoram significativamente a segurança e o conforto, embora não substituam a vigilância e o discernimento humano.

Masello *et al.* (2022) elucidam os múltiplos benefícios proporcionados pelos sistemas de assistência, que incluem a análise do comportamento do condutor, alertas de risco, personalização da experiência, feedback pós-condução e prevenção de acidentes. A eficácia desses sistemas está intrinsecamente ligada à sua capacidade de resposta em tempo real, presente em sistemas ADAS.

Nesse contexto, Katare *et al.* (2023) enfatizam que os modelos de IA para direção autônoma estão sujeitos a rigorosos requisitos de latência, demandando tempos de resposta na ordem de milissegundo(s) (ms). Essa exigência é particularmente crítica em funções como localização, frenagem de emergência e detecção de obstáculos.

Jiang *et al.* (2023) corroboram com essa perspectiva, destacando que os sistemas ADAS devem ser capazes de detectar objetos nas vias em tempo real, o que impõe severas restrições ao tempo de resposta. O desafio é amplificado pela limitação dos recursos computacionais disponíveis nos dispositivos de borda, onde esses sistemas frequentemente operam. Os autores também ressaltam a importância da velocidade de inferência de imagens, no caso de uso da câmera do smartphone, medida em Quadros por Segundo (Frames Por Second - FPS), como uma métrica crucial para a avaliação desses sistemas.

Adicionalmente, Theivadas e Ponnan (2024) salientam a necessidade de detecção em tempo real da fadiga do motorista como medida preventiva contra acidentes. Hina *et al.* (2021) complementam essa visão, apontando que a fusão de dados provenientes de múltiplos sensores e serviços é um processo computacionalmente intensivo que deve ser executado com celeridade para permitir que um veículo inteligente tome decisões informadas.

A convergência desses estudos evidencia a importância de analisar a eficiência do emprego de DL nos smartphones para atuação em sistemas ADAS, considerando as restrições de tempo real e recursos computacionais limitados.

2.2 Inteligência Artificial na Borda

Li *et al.* (2022) conceituam a Edge AI como a convergência entre computação de borda e inteligência artificial, na qual os serviços de IA são implementados próximos ao ponto de geração dos dados, na periferia da rede. Essa abordagem descentralizada permite que os dispositivos na borda da rede realizem análises de dados e tomadas de decisão de forma autônoma, reduzindo a dependência de servidores centralizados.

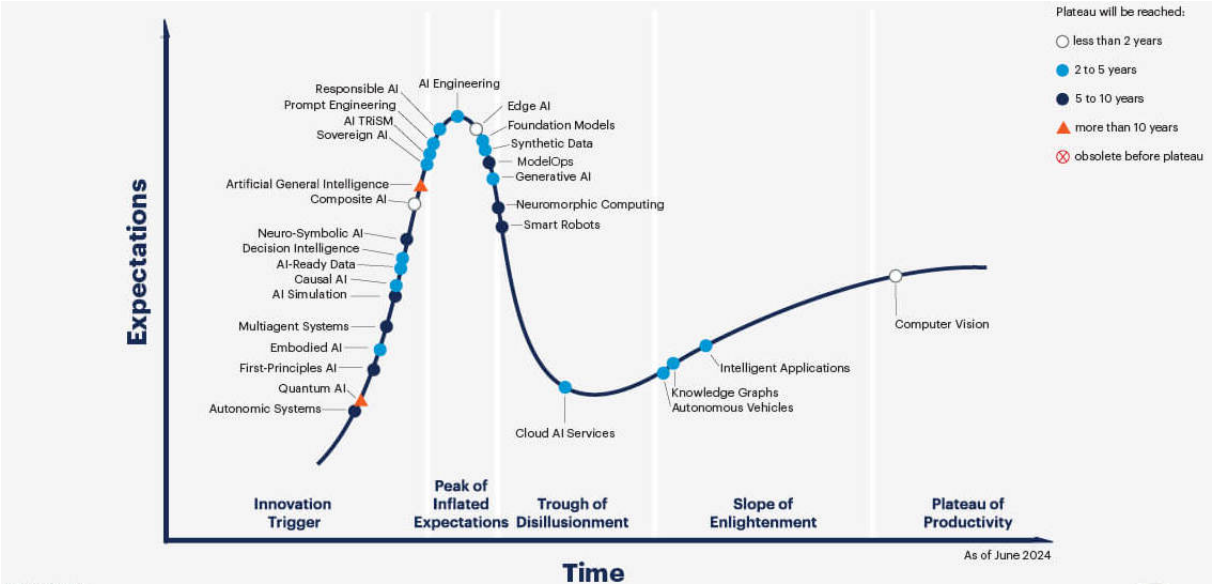
A Edge AI viabiliza o processamento local de dados em dispositivos como smartphones, relógios inteligentes e outros dispositivos de Internet das Coisas (Internet of Things - IoT), em contraposição ao envio de todos os dados para processamento em nuvem. Essa estratégia resulta em redução de latência, melhoria na eficiência energética e aumento da privacidade dos dados.

As perspectivas para a evolução e o posicionamento de mercado da Edge AI demonstram rápida transformação. Em 2024, segundo Afraz e Haritha (2024), a tecnologia estava posicionada no "pico das expectativas infladas", indicando entusiasmo elevado, mas também riscos de expectativas irrealistas, conforme mostra a Figura 2.1.

No entanto, dados mais recentes de 2025 (Khandabattu (2025) evidenciam um estágio de maior maturidade, com previsão de que 55% dos dados sejam processados em dispositivos de borda até o final do ano, como representado na Figura 2.2.

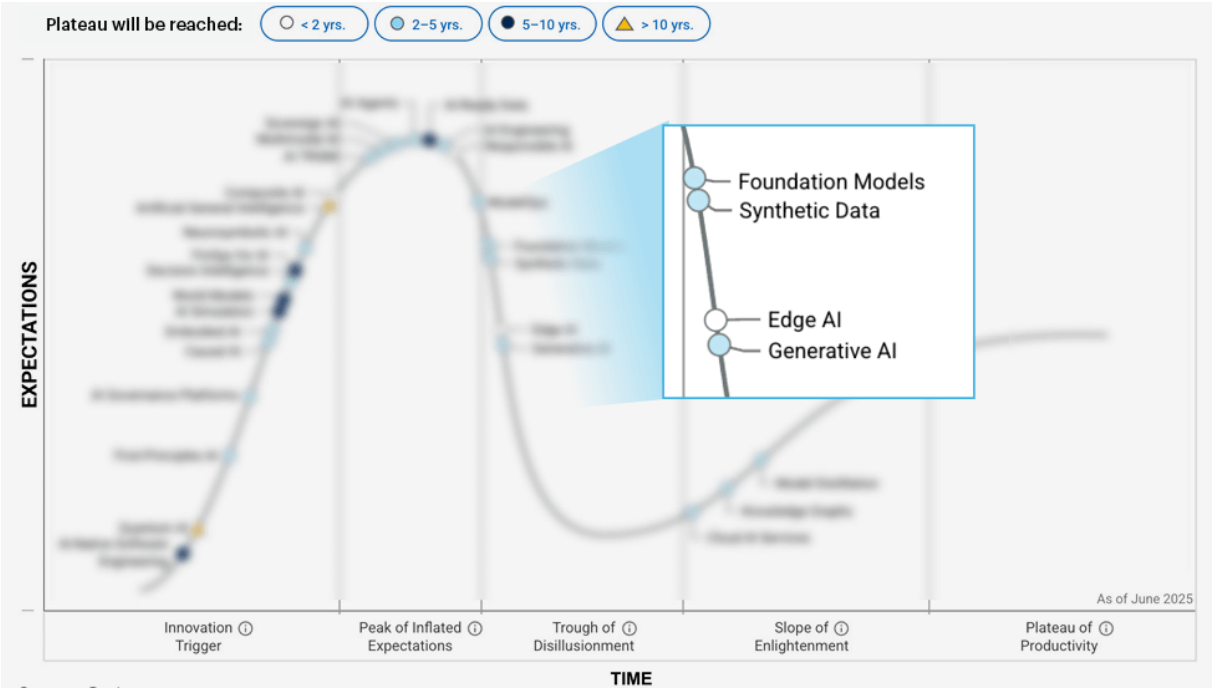
Essa evolução evidencia não apenas a consolidação da Edge AI, mas também torna o cenário mais propício para aplicações práticas, como sistemas ADAS embarcados em smartphones.

Figura 2.1 – Posição da Edge AI segundo a Gartner, 2024



Fonte: Gartner, Afraz e Haritha (2024)

Figura 2.2 – Posição da Edge AI segundo a Gartner, 2025



Fonte: Gartner, Khandabattu (2025)

2.3 Evolução dos SoCs para Aceleração de IA em Dispositivos Móveis

A integração de hardware dedicado para aceleração de IA em smartphones representa um marco significativo na evolução dos SoCs. Embora o uso de hardware especializado em dispositivos móveis não seja uma novidade, a última década testemunhou uma aceleração sem precedentes no desenvolvimento de soluções de hardware eficientes para IA. Ignatov *et al.* (2018) traçam essa evolução, destacando:

> "No início da década de 1990, os DSPs começaram a ser incorporados em telefones celulares, inicialmente para codificação e compressão de voz e processamento de sinais de rádio. Com a integração de câmeras e recursos multimídia, os DSPs expandiram seu papel para o processamento de imagem, vídeo e som. Ao contrário dos computadores de mesa, os DSPs não foram substituídos por CPU e GPU em dispositivos móveis, pois frequentemente ofereciam desempenho superior com menor consumo de energia, um fator crítico para dispositivos portáteis. Nos últimos anos, o poder computacional dos DSPs móveis e outros componentes SoCs cresceu drasticamente e, complementados por GPUs, NPUs e núcleos de IA dedicados, agora permitem computações baseadas em IA e aprendizado profundo" (Ignatov *et al.*, 2018).

No mercado de SoCs para dispositivos móveis com capacidade de aceleração por hardware, há uma alta diversidade de propostas tecnológicas. No entanto, conforme evidenciado por Ignatov *et al.* (2018) e Ignatov *et al.* (2021), existem sete empresas que se destacam na vanguarda dessa revolução tecnológica.

Primeiramente, a **Qualcomm** (2017), reconhecida como uma das pioneiras no desenvolvimento de SoCs móveis, destaca-se por oferecer o Snapdragon Neural Processing Engine (SNPE), uma solução projetada para acelerar o processamento de redes neurais artificiais em dispositivos embarcados. Os seus *chipsets* integram o Hexagon DSP, componente especializado no processamento em tempo real de sinais, o que aumenta a eficiência em aplicações de inteligência artificial. Além do pioneirismo tecnológico, a Qualcomm é amplamente conhecida por sua política de licenciamento e cobrança de *royalties* sobre patentes essenciais para padrões industriais, notadamente em áreas como comunicação móvel (5G, LTE, CDMA) e tecnologias de processadores (Qualcomm, 2017).

A **HiSilicon** (2021), subsidiária da Huawei, ganhou destaque global com sua linha de *chipsets* Kirin, especialmente graças ao desempenho do Kirin 970 em tarefas envolvendo redes neurais. Sua plataforma HiAI foi desenvolvida para permitir a execução eficiente de modelos avançados de inteligência artificial nos próprios SoCs da marca, tornando os dispositivos mais autônomos e potentes para aplicações de IA.

Outro grande protagonista desse mercado é a **MediaTek** (2022), que figura entre os principais fornecedores de SoCs ao oferecer o NeuroPilot SDK. Essa ferramenta suporta tanto o uso de GPU quanto de APUs (Unidades de Processamento de Acelerado), viabilizando maior aceleração em algoritmos de DL em seus dispositivos.

No mesmo contexto de inovação, a **Samsung** (2022), por meio da tradicional linha Exynos, tem ampliado investimentos em pesquisa e desenvolvimento para incorporar tecnologias de DL em dispositivos móveis. Recentemente, a empresa passou a integrar NPU em seus chips, elevando significativamente o desempenho em tarefas de inteligência artificial.

A **Unisoc** (2020), antes conhecida como Spreadtrum, consolida-se como empresa chinesa relevante no segmento de semicondutores. Ela disponibiliza o SDK UNIAI, dedicado à otimização das operações de IA executadas em suas plataformas SoCs, contribuindo para o avanço do ecossistema de IA móvel.

Apesar de sua tradição em processadores para computadores pessoais, a **Intel** (2024) também tem direcionado esforços para o mercado de Edge AI. Entre suas contribuições, destaca-se a biblioteca Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN), que proporciona aceleração substancial de operações de DL em sistemas embarcados e dispositivos inteligentes.

Por fim, a **NVIDIA** (2025), conhecida mundialmente por suas GPUs de alto desempenho, também se destaca na produção de SoCs por meio da plataforma Tegra. As bibliotecas CUDA e CUDA Deep Neural Network Library (cuDNN) da empresa tornaram-se referência para o desenvolvimento e a aceleração de redes neurais profundas, permitindo ganhos significativos em eficiência e desempenho em diversas aplicações embarcadas e móveis.

Essas empresas atuam na vanguarda tecnológica, promovendo a integração de capacidades avançadas de inteligência artificial em dispositivos móveis e transformando profundamente a forma na qual nos relacionamos com a tecnologia no cotidiano.

A competição acirrada entre os fabricantes de SoCs tem impulsionado avanços significativos na capacidade de processamento de IA em dispositivos móveis. A integração de NPU, DSP otimizados e GPU mais eficientes tem permitido a execução de modelos de aprendizado profundo cada vez mais complexos diretamente nos smartphones, abrindo caminho para aplicações avançadas como ADAS *Mobile Telematics*.

A rápida evolução desses SoCs não só melhora o desempenho em tarefas de IA, mas também otimiza o consumo de energia, um fator crítico em dispositivos móveis. Essa tendência

de Edge AI está alinhada com as projeções de mercado discutidas anteriormente, reforçando a importância de estudos que avaliem a eficácia desses sistemas em aplicações práticas como ADAS.

2.4 *Frameworks* e Ferramentas para Desenvolvimento de Edge AI em Dispositivos Móveis

O desenvolvimento de aplicações de Edge AI para dispositivos móveis requer *frameworks* especializados que otimizem o desempenho e a eficiência energética, considerando as limitações de recursos desses dispositivos. A análise comparativa dos principais *frameworks* disponíveis revela uma diversidade de abordagens e capacidades, cada uma com suas próprias características e limitações, como apresentadas no Quadro 2.1 e no Quadro 2.2.

Quadro 2.1 – *Frameworks* de Edge AI (parte 1)

Recurso	**TensorFlow Lite**	**Edge Impulse**	**OpenVINO**	**ONNX Runtime**
Desenvolvedor	Google	Edge Impulse	Intel	Microsoft
Código aberto	Sim	Parcialmente	Sim	Sim
Plataformas principais	iOS, Android, Linux SBCs, MCUs	MCUs, dispositivos restritos	Intel CPU, GPU, VPUs, FPGAs	Windows, Linux, macOS, Android, iOS, JavaScript
Formatos de modelo	.tflite, .pb	Edge Impulse, .tflite	.xml, .bin, .onnx	.onnx, .ort
Treinamento *on-device*	Limitado	Sim, via plataforma	Não	Não
Otimizações principais	Quantização, poda	Automatizadas, aumento de dados	Otimização de grafo	Otimização de grafo
Aceleração de hardware	GPU, TPUs, Edge TPUs	Via plataforma Edge Impulse	Intel VPUs, GPU, CPU	CPU, GPU, NPU, outros
Casos de uso principais	Visão computacional, NLP, IoT	TinyML, IoT restrito	Visão computacional, IoT	*Cloud* e *edge computing*

Fonte: Elaborado pelo autor

Quadro 2.2 – *Frameworks* de Edge AI (parte 2)

Recurso	**Apache TVM**	**Arm NN**	**Core ML**
Desenvolvedor	Apache Software Foundation	Arm	Apple
Código aberto	Sim	Sim	Não
Plataformas principais	Diversos *back-ends* de hardware	CPU, GPU e NPU Arm	iOS, watchOS, macOS, tvOS
Formatos de modelo	Múltiplos (TensorFlow, PyTorch, etc.)	ONNX, .tflite	.mlmodel
Treinamento *on-device*	Limitado	Não	Sim, via Create ML
Otimizações principais	Nível de grafo e operador, autoTVM	Otimizações de grafo	Específicas da plataforma
Aceleração de hardware	Diversos aceleradores	Arm CPU, GPU, NPU	Apple Neural Engine, GPU, CPU

Fonte: Elaborado pelo autor

Entre os *frameworks* mais relevantes para o desenvolvimento de Edge AI em smartphones, o TensorFlow Lite (Google, 2017b) destaca-se como uma das principais soluções. Desenvolvido pelo Google, esse *framework* oferece otimizações específicas para inferência móvel, suportando um conjunto otimizado de operações do TensorFlow e fornecendo um sistema de *delegates* para aceleração em hardware especializado. O sistema de *delegates* permite a aceleração via GPU e integração com aceleradores específicos de fabricantes, possibilitando otimização automática para diferentes arquiteturas de hardware e balanceamento entre desempenho e consumo energético.

O PyTorch Mobile (Meta AI, 2019) representa uma alternativa robusta, especialmente para desenvolvedores familiarizados com o ecossistema PyTorch. O *framework* oferece integração transparente com modelos PyTorch existentes, otimizações específicas para dispositivos móveis e suporte a técnicas de quantização e *pruning* de modelos. Além disso, disponibiliza ferramentas integradas para análise de desempenho, facilitando a otimização das aplicações.

No contexto do sistema operacional Android, a Android Neural Networks API (NNAPI), de acordo com (Google, 2017a), desempenha um papel fundamental como camada de abstração entre as aplicações e o hardware. Essa API (Interface de Programação de Aplicação, do inglês, Application Programming Interface) fornece uma interface unificada para acesso a diferentes tipos de aceleradores (NPU, DSP, GPU), oferecendo otimizações específicas para diferentes arquiteturas e gerenciamento automático de recursos de hardware. A NNAPI permite que *frameworks* como TensorFlow Lite e PyTorch Mobile aproveitem de forma facilitada os aceleradores de hardware disponíveis nos dispositivos.

Os principais fabricantes de SoCs também oferecem suas próprias soluções de desenvolvimento, como o Qualcomm Neural Processing SDK (Qualcomm Technologies, Inc., 2017) – otimizado para processadores Snapdragon; o MediaTek NeuroPilot SDK (MediaTek Inc., 2019) – focado em APUs MediaTek; e o Samsung Neural SDK (Samsung Electronics Co., Ltd., 2018), específico para processadores Exynos. Embora esses Kits de Desenvolvimento de Software (Software Development Kit - SDK) ofereçam otimizações específicas para seus respectivos hardwares, sua natureza proprietária pode limitar a portabilidade das aplicações entre diferentes plataformas.

Recentemente, o Google lançou o LiteRT, um *runtime* de alta performance projetado como evolução do TensorFlow Lite, com o objetivo de ampliar ainda mais a eficiência de inferências em dispositivos móveis e de borda. O LiteRT apresenta como principais diferenciais

o suporte expandido a modelos provenientes de múltiplos *frameworks* – como TensorFlow, PyTorch e JAX – e uma integração otimizada com diversos tipos de aceleradores de hardware (GPU, NPU, DSP), especialmente com o uso do NNAPI no Android (Google AI, 2024). Apesar desses atrativos e de ganhos em redução de binários e execução ainda mais ágil, seu ecossistema e ferramentas ainda estão em processo de consolidação, e há desafios quanto à compatibilidade de operações complexas, à curva de aprendizado para otimizações avançadas e ao processo de depuração de modelos, que pode ser menos intuitivo. Assim, o LiteRT surge como uma alternativa promissora para Edge AI, mas recomenda-se cautela em sua adoção imediata em projetos críticos, reservando sua utilização principalmente para testes e validações até o amadurecimento da tecnologia.

2.5 Algoritmos de Aprendizagem Profunda para Dispositivos Móveis

Os avanços recentes em aprendizagem profunda têm impulsionado significativamente as capacidades de processamento inteligente em smartphones. Essa seção explora os principais algoritmos utilizados nesse contexto, com ênfase naqueles avaliados pela AI Benchmark (Zurich, 2024a), uma ferramenta importante para a análise de desempenho de IA em dispositivos móveis.

As Redes Neurais Convolucionais (Convolutional Neural Networks - CNNs) formam a base de muitas aplicações de visão computacional em smartphones. A MobileNet (Howard *et al.*, 2017) destaca-se por sua arquitetura inovadora que utiliza convoluções separáveis em profundidade para reduzir significativamente a carga computacional. Essa abordagem a torna particularmente eficiente para dispositivos com recursos limitados, permitindo aplicações em tempo real de reconhecimento de imagem e detecção de objetos. A arquitetura também serve como base para variantes otimizadas, como a QF-MobileNet, especialmente projetada para quantização e inferência eficientes em dispositivos móveis.

A Inception (Szegedy *et al.*, 2017) e a EfficientNet (Tan; Le, 2020) representam abordagens distintas para equilibrar precisão e eficiência computacional. A Inception, com suas múltiplas versões, estabeleceu-se como uma arquitetura profunda e eficiente, especialmente quando combinada com conexões residuais na forma do Inception-ResNet. A EfficientNet, por sua vez, introduziu uma metodologia sistemática para dimensionamento de CNNs, otimizando

simultaneamente profundidade, largura e resolução para maximizar o desempenho dentro das restrições de recursos disponíveis.

Para detecção de objetos em tempo real, o YOLO-Tiny (Guan *et al.*, 2022; Shen; Liao; Zheng, 2024; Redmon; Farhadi, 2018) oferece uma solução otimizada por meio de sua arquitetura de estágio único. Essa versão compacta do YOLO prioriza a velocidade de inferência, embora com algum compromisso na precisão em comparação com modelos maiores. Variantes como o YOLOv4-tiny e YOLOv4-dense foram desenvolvidas especificamente para dispositivos de borda, oferecendo diferentes equilíbrios entre velocidade, precisão e eficiência computacional.

A segmentação semântica em dispositivos móveis é frequentemente realizada utilizando o DeepLab (Sandler *et al.*, 2019; Chen *et al.*, 2023). Essa arquitetura emprega convoluções dilatadas para aumentar a resolução do mapa de características sem aumentar significativamente o custo computacional. A versão DeepLabv3+ (Chen *et al.*, 2018b) incorpora o MobileNetV2 como rede de extração de recursos, resultando em uma arquitetura mais leve e eficiente, mantendo alta precisão na segmentação.

De acordo com Ignatov *et al.* (2023), Zhang *et al.* (2021) e Wang *et al.* (2021) para estimativa de profundidade em imagens, arquiteturas baseadas em MobileNetV3 têm demonstrado resultados promissores. Essas soluções permitem estimar profundidade a partir de imagens monoculares, recurso esse voltado para aplicações de fotografia computacional e realidade aumentada em smartphones. A eficiência dessas implementações é frequentemente aprimorada com técnicas de destilação do conhecimento, permitindo que modelos mais leves emulem o desempenho de arquiteturas mais complexas, como o EfficientNetV2.

2.6 Técnicas de Otimização para Edge AI em Dispositivos Móveis

A implementação eficiente de modelos de aprendizagem profunda em dispositivos móveis requer técnicas de otimização específicas para superar as limitações de recursos e atender aos requisitos de processamento em tempo real. Entre essas técnicas, a quantização e a poda de rede destacam-se como métodos fundamentais para reduzir o tamanho do modelo e acelerar a inferência, mantendo a precisão em níveis aceitáveis.

A quantização é uma técnica que reduz a precisão numérica dos pesos e ativações em uma rede neural. Jacob *et al.* (2018) introduziram um esquema de quantização que permite

a execução de redes neurais usando aritmética de inteiros de 8 bits, em contraste com os tradicionais 32 bits de ponto flutuante. O processo de quantização pode ser descrito em várias etapas:

a) análise do intervalo: Primeiramente, é realizada uma análise para determinar o intervalo de valores dos pesos e ativações em cada camada da rede;
b) escolha do esquema de quantização: com base nessa análise, é escolhido um esquema de quantização, que mapeia os valores de ponto flutuante para inteiros. Um esquema comum é a quantização linear uniforme, em que:

$$q = \text{round}\left(\frac{x}{\text{scale}} + zero\_point\right) \tag{2.1}$$

q é o valor quantizado, *x* é o valor original em ponto flutuante, *scale* é um fator de escala e *zero_point* é um valor de deslocamento.
c) calibração: Os parâmetros de quantização (*scale* e *zero_point*) são calibrados para cada camada, geralmente usando um conjunto de dados representativo;
d) requantização: As operações entre camadas podem requerer requantização, ajustando a escala dos resultados intermediários;
e) treinamento fino (opcional): Em alguns casos, um treinamento fino é realizado após a quantização para recuperar qualquer perda de precisão.

Krishnamoorthi (2018) expandiu esse trabalho, descrevendo técnicas de quantização pós-treinamento (Post Training Quantization - PTQ) e quantização consciente de treinamento (Quantization Aware Training - QAT).

PTQ aplica quantização a um modelo já treinado, enquanto QAT incorpora a quantização durante o processo de treinamento, geralmente resultando em melhor precisão.

A poda de rede, por outro lado, envolve a remoção sistemática de conexões ou neurônios menos importantes. Han *et al.* (2015) demonstraram que essa técnica pode reduzir significativamente o número de parâmetros do modelo, frequentemente em mais de 90%, com mínima perda de precisão. O processo de poda típico inclui as seguintes etapas:

a) treinamento: A rede é inicialmente treinada normalmente;
b) avaliação de importância: A importância de cada conexão ou neurônio é avaliada, geralmente com base na magnitude dos pesos ou em métricas mais sofisticadas como a sensibilidade da função de perda à remoção do parâmetro;

c) poda: As conexões ou neurônios menos importantes são removidos. Isso pode ser feito de uma vez ou iterativamente;
d) retreinamento: A rede podada é retreinada para recuperar a precisão perdida durante a poda;
e) iteração: Os passos 2-4 podem ser repetidos várias vezes para alcançar níveis mais altos de compressão.

Molchanov *et al.* (2017) propuseram um método de poda baseado em critérios de Taylor, que avalia a importância dos neurônios com base na sensibilidade da função de custo à sua remoção. Essa abordagem mostrou-se particularmente eficaz para CNNs.

A combinação de quantização e poda pode levar a reduções ainda mais significativas no tamanho do modelo e na complexidade computacional. Por exemplo, Han *et al.* (2015) introduziram a "Deep Compression", uma técnica que combina poda, quantização e codificação Huffman para comprimir redes neurais profundas por um fator de 35-49x sem perda de precisão.

Além disso, a eficácia dessas técnicas pode variar dependendo da arquitetura da rede e do domínio da aplicação. Por exemplo, Li *et al.* (2017) observaram que camadas de uma rede convolucional podem ter sensibilidades diferentes à quantização e à poda, sugerindo a necessidade de abordagens adaptativas que aplicam níveis variáveis de compressão em diferentes partes do modelo.

A pesquisa contínua nessa área explora técnicas ainda mais avançadas, como esquemas de quantização não uniforme (Zhou *et al.*, 2017) e métodos de poda estruturada que removem canais ou filtros inteiros (Liu *et al.*, 2019). Essas abordagens prometem melhorar ainda mais a eficiência dos modelos de aprendizagem profunda em dispositivos com recursos limitados, abrindo caminho para aplicações cada vez mais sofisticadas de Edge AI em dispositivos móveis.

3 TRABALHOS RELACIONADOS

Diversos estudos têm explorado o uso de algoritmos de aprendizagem profunda em smartphones para aplicações de ADAS, promovendo avanços em segurança e eficiência veicular. Este capítulo apresenta alguns dos principais trabalhos que abordam soluções práticas, benchmarks e desafios em ADAS móveis, fornecendo um panorama atualizado do campo.

Ignatov *et al.* (2019) realizaram um benchmark abrangente de IA para smartphones, avaliando diferentes *chipsets* e *frameworks* de aprendizagem profunda por meio da plataforma AI Benchmark. O estudo destacou a evolução da aceleração por hardware em tarefas como classificação de imagens e reconhecimento facial, comparando resultados entre CPUs/GPUs de smartphones e desktops, identificando principais tendências e limitações tecnológicas.

Expandindo esse trabalho, Ignatov *et al.* (2021) apresentaram a Burnout Benchmark, voltado à avaliação do desempenho sustentado de SoCs em cargas intensas de IA para smartphones. A Burnout Benchmark foca em aspectos críticos para aplicações contínuas, como o ADAS, incluindo eficiência energética, comportamento sob *throttling* térmico e desempenho detalhado de CPU, GPU e NPU.

No contexto de benchmarks mais sistêmicos, Tabani *et al.* (2021) desenvolveram o **ADBench**, dedicado à avaliação fim a fim de sistemas de condução autônoma. Diferentemente dos benchmarks centrados em componentes isolados, o ADBench propõe métricas para desempenho, robustez e confiabilidade em cenários realistas, promovendo avaliações padronizadas de sistemas completos de direção autônoma e complementando abordagens anteriores.

Hina *et al.* (2021) propuseram o Projeto CASA, um sistema ADAS alternativo baseado em smartphones, focado em acessibilidade e escalabilidade. Utilizando fusão de dados multifonte e modelagem semântica, o sistema foi validado em ambiente virtual e real, demonstrando a viabilidade de soluções ADAS baseadas em dispositivos móveis.

No aprimoramento de modelos para dispositivos móveis, Howard *et al.* (2019) apresentaram as variantes MobileNetV3, otimizadas para cenários de recursos restritos, alcançando altos índices de precisão e velocidade em classificação, detecção e segmentação de imagens em tempo real em smartphones.

Chen *et al.* (2023) realizaram avanços em algoritmos de detecção de veículos e segmentação de imagens para transporte inteligente, integrando melhorias em YOLOv4 e DeepLabv3+

com mecanismos de atenção, o que resultou em incrementos de precisão e viabilidade para tarefas em tempo real em dispositivos móveis.

Katare *et al.* (2023) realizaram uma revisão abrangente sobre Edge AI para direção autônoma eficiente em energia, identificando gargalos de consumo e propondo técnicas de otimização e uso inteligente da computação para reduzir a dependência da nuvem.

Musa *et al.* (2022) abordaram a convergência de redes centradas em informação e computação de borda para Internet dos Veículos (Internet of Vehicles - IoV), propondo soluções de IA para desafios de latência, altas taxas de dados e mobilidade, e destacando aplicações como condução cooperativa e percepção colaborativa.

Weig (2016) analisaram oportunidades e desafios para a indústria de semicondutores em ADAS, enfatizando a importância da entrada e padronização tecnológica neste mercado, além da integração entre hardware e software para diferenciação e segurança.

Os estudos revisados demonstram que a aplicação de aprendizagem profunda em smartphones para ADAS é um campo em rápida evolução, impulsionado por benchmarks inovadores, otimização de modelos e avanços em hardware dedicado. Destaca-se a necessidade de avaliações robustas (como AI Benchmark, Burnout Benchmark e ADBench) e a tendência de maior integração entre sistemas, promovendo eficiência, acessibilidade e precisão em ambientes restritos, como o dos dispositivos móveis.

Abaixo, os principais trabalhos relacionados são sintetizados em quadro, facilitando a comparação entre escopo, contribuições e eventuais lacunas identificadas. Ressalta-se que a última linha refere-se ao presente trabalho, permitindo situá-lo no contexto dos estudos analisados.

Quadro 3.1 – Resumo dos trabalhos relacionados

(Continua)

Autores	**Resumo**	**Resultados**	**Contribuições**	**Lacunas**
Ignatov et al. (2019)	Benchmark de IA em smartphones via AI Benchmark.	Métricas detalhadas de desempenho em DL móvel.	Evolução no entendimento do estado da arte do HW para IA.	Necessidade de avaliações sob uso contínuo.
Ignatov et al. (2021)	Burnout Benchmark para desempenho sustentado em IA móvel.	Análise de *throttling*, eficiência energética e durabilidade do SoC.	Avaliação crítica de limitações térmicas e energéticas em uso intensivo.	Cobertura restrita a tarefas padronizadas, não sistemas completos.

Fonte: Elaborado pelo autor

Quadro 3.2 – Resumo dos trabalhos relacionados

(Conclusão)

Autores	**Resumo**	**Resultados**	**Contribuições**	**Lacunas**
Tabani et al. (2021)	ADBench: benchmark de sistemas autônomos de direção de ponta a ponta.	Avaliação holística de desempenho, robustez e confiabilidade.	Padronização de métricas para sistemas completos de ADAS e direção autônoma.	Adaptação gradual do benchmark em larga escala; maior aplicação em pesquisa.
Chen et al. (2023)	Melhoria em detecção de veículos e segmentação em borda.	Precisão de detecção de 82,03% para 86,22%.	Otimização de YOLOv4 / DeepLabv3+ para dispositivos móveis.	Validar em cenários reais diversos.
Howard et al. (2019)	MobileNetV3 eficiente para DL em smartphones.	Avanços em velocidade e precisão vs. MobileNetV2.	Otimização DL para recursos restritos.	Maior análise em ADAS reais ainda necessária.
Hina et al. (2021)	Projeto CASA: ADAS acessível via smartphone.	Validação prática com fusão de dados e modelagem semântica.	Democratização de ADAS em dispositivos do usuário.	Escalabilidade a ser testada em diferentes contextos.
Katare et al. (2023)	Revisão sobre Edge AI e eficiência energética em autônomos.	Propostas de soluções para consumo inteligente em IA móvel.	Ênfase em redução de dependência da nuvem.	Adoção limitada por custo/infraestrutura.
Musa et al. (2022)	Convergência de ICN, Edge Computing e IA para IoV.	Soluções para latência e mobilidade dinâmica.	Integração de abordagens de rede e IA.	Aplicação comercial ainda emergente.
Florian et al. (2016)	Análise de desafios/op. para semicondutores no mercado ADAS.	Estratégias de diferenciação em HW/SW e segurança.	Perspectiva mercadológica para evolução do setor.	Desempenho prático em ambientes restritos pouco avaliado.
Freitas (2025)	Análises de para otimizações de IA em dispositivos móveis.	Propostas de soluções para uso ADAS *Mobile Telematics*.	Avaliação de limitações no uso de smartphone como ADAS	Necessidade de validação em mais dispositivos.

Fonte: Elaborado pelo autor

4 METODOLOGIA

Este capítulo apresenta a abordagem metodológica adotada neste trabalho, detalhando as etapas de fundamentação teórica, seleção de dispositivos e ferramentas, configuração experimental, procedimentos de teste e a coleta e tratamento dos dados para avaliar o desempenho de inferência de inteligência artificial embarcada em dispositivos móveis.

4.1 Revisão de Literatura

O embasamento teórico foi realizado por meio de revisão sistemática em bases como IEEE Xplore, ACM Digital Library, Science Direct, Google Scholar e Scopus, com ênfase em publicações de 2019 a 2024. Foram priorizados estudos sobre inferência de IA em dispositivos móveis, otimização de redes neurais profundas e, principalmente, tecnologias ADAS. Por tratar-se do foco central deste trabalho, a literatura revisada inclui soluções de assistência ao condutor baseadas em IA, com destaque para técnicas, desafios e tendências em detecção e resposta a eventos de trânsito por meio de smartphones.

4.2 Seleção de Dispositivos e Ferramentas

A seleção dos dispositivos teve como objetivo abranger diferentes faixas de mercado, incluindo modelos de topo de linha e intermediários, assim como diversas arquiteturas de processadores e aceleradores, como CPU, GPU e NPU. Os dispositivos testados foram escolhidos dentre as opções disponíveis. As especificações técnicas completas dos aparelhos estão apresentadas no Apêndice A.

4.2.1 Ferramentas de Avaliação

As ferramentas utilizadas incluem:

AI Benchmark: esta ferramenta, na sua versão 6.0.2, é projetada para avaliar o desempenho de inteligência artificial em dispositivos móveis. Ela abrange uma variedade de tarefas de DL , que incluem, mas não se limitam a, classificação de imagens (utilizando redes como MobileNet e Inception-V3), segmentação semântica (com DeepLab-V3+) e estimação de profundi-

dade (com MV3-Depth). As redes neurais testadas são representativas de diversas arquiteturas, proporcionando uma visão abrangente das capacidades e limitações de várias estratégias utilizadas na resolução de diferentes problemas de inteligência artificial (Zurich, 2024a). Assim, a AI Benchmark serve como uma ferramenta essencial para identificar não apenas o desempenho, mas também a eficiência das soluções implementadas em dispositivos móveis.

Burnout Benchmark: na sua versão 2.0.8, utilizada para a avaliação do desempenho computacional, comportamento térmico e eficiência energética de CPUs, GPUs e NPUs de smartphones. Ela consegue carregar diferentes componentes de SoC simultaneamente, permitindo uma análise detalhada da sustentabilidade de desempenho e do consumo de energia sob carga máxima (Ignatov, 2025).

Aplicação Android Personalizada: desenvolvida pelo autor especificamente para este estudo utilizando a plataforma TensorFlow Lite, a aplicação implementa funcionalidade para sistemas ADAS. A aplicação foi projetada para processar dados de vídeo em tempo real e coletar métricas de desempenho durante a inferência.

4.3 Configuração Experimental

Os experimentos foram executados em ambiente controlado quanto à temperatura e iluminação, contemplando vídeos com cenários diurno e noturno. Todos os dispositivos foram mantidos apenas com a carga da bateria e acima de 50%, sem conexão à rede elétrica.

Testes com AI Benchmark e Burnout Benchmark: os aparelhos foram avaliados sem capa de proteção para maximizar a dissipação térmica e obter medições mais precisas da temperatura do hardware.

Testes com Aplicação Personalizada: os dispositivos permaneceram protegidos por capas de proteção durante toda a execução, simulando condições realistas de uso em aplicações ADAS, onde smartphones normalmente são utilizados com proteção.

Fotografias dos ambientes experimentais constam no apêndice D. Para consulta aos percursos completos dos testes, foram utilizados vídeos de referência disponíveis em (OZ, 2023; Utah, 2023).

4.4 Procedimentos de Teste

Esta seção descreve os procedimentos experimentais adotados para avaliar o desempenho dos dispositivos móveis selecionados. Os testes foram organizados em três categorias principais: benchmarks padronizados para estabelecimento de métricas de referência, testes específicos com a aplicação ADAS *Mobile Telematics* desenvolvida, e coleta sistemática de métricas de desempenho. Todos os procedimentos foram executados de forma a garantir a reprodutibilidade e validade dos resultados.

4.4.1 Testes de Benchmarks Padronizados

Cada dispositivo foi submetido a sequências padronizadas de execução dos benchmarks AI Benchmark e Burnout Benchmark para estabelecer métricas de referência de desempenho e estabilidade térmica.

4.4.2 Testes com Aplicação ADAS

Os testes com a aplicação personalizada foram realizados em dois contextos experimentais distintos. Um vídeo noturno de 52:58 minutos (Utah, 2023) e, na sequência, um vídeo de 41:51 minutos representando o cenário diurno (OZ, 2023), totalizaram aproximadamente 94:49 minutos de gravação. Os testes foram realizados em ambiente escuro, com uma TV OLED 4K a 60 FPS, processando vídeos gravados com boa visibilidade e condições típicas de tráfego urbano. A temperatura do ar foi controlada em 21 °C. Vide imagens e descrição no apêndice D, nas imagens 1 e 3.

4.4.3 Métricas Coletadas

Durante todos os experimentos, foram monitorados:

a) tempo de inferência (ms);
b) taxa de quadros por segundo (FPS);
c) temperaturas do SoC (quando permitido pelo sistema) e da bateria (°C), incluindo a variação térmica ($\Delta T = T_{\text{final}} - T_{\text{inicial}}$);

d) consumo energético (diferença de % inicial e % final da bateria);
e) latência de processamento por quadro;
f) estabilidade de desempenho ao longo do tempo;
g) número de objetos detectados a cada período;
h) contagem individual por classe de objeto detectada relevante a ADAS;
i) distribuição de classes detectadas.

4.4.4 Implementação da Aplicação ADAS

A aplicação Android desenvolvida para este estudo implementa um *pipeline* de processamento que inclui:

a) captura e pré-processamento: aquisição de quadros de vídeo e redimensionamento para as dimensões de entrada do modelo SSD-MobileNet v1;
b) inferência: execução do modelo SSD-MobileNet v1 em formato TensorFlow Lite (.tflite) utilizando NNAPI como *delegate* de aceleração;
c) pós-processamento: análise dos resultados da inferência para detecção e localização de objetos relevantes aos sistemas ADAS, incluindo filtragem por confiança mínima (*confidence threshold*);
d) coleta de métricas: registro em tempo real de todas as métricas de desempenho especificadas.

4.4.5 Configuração Técnica do Modelo

Arquitetura: SSD-MobileNet v1, composta por um *backbone* MobileNet v1 para extração de características e uma cabeça SSD (*Single Shot Detector*) para detecção de objetos. Esta arquitetura foi escolhida por oferecer o equilíbrio ideal entre velocidade de inferência e capacidade de detecção em tempo real, características essenciais para aplicações ADAS *Mobile Telematics*. A MobileNet v1 atua como extrator de características visuais, processando a imagem de entrada e gerando mapas de características em diferentes escalas, enquanto a cabeça SSD utiliza essas características para realizar simultaneamente a classificação de objetos e a regressão das *bounding boxes* em uma única passagem pela rede.

***Dataset* de Treinamento**: o modelo utilizado e disponibilizado pelo TensorFlow Lite Google (2017b) foi treinado no *dataset* MS-COCO, que contém 80 classes de objetos, incluindo categorias fundamentais para sistemas ADAS, como veículos (carros, caminhões, motocicletas, bicicletas), pessoas, semáforos e outros elementos urbanos relevantes para a condução assistida (Lin *et al.*, 2015).

***Framework* de Inferência**: TensorFlow Lite e NNAPI como *delegate* primário para aceleração de hardware, permitindo o aproveitamento automático dos aceleradores de IA disponíveis em cada SoC testado.

Especificações do Modelo:

a) arquitetura: SSD-MobileNet v1;
b) dimensões de entrada: 300×300 pixels;
c) formato de entrada: RGB, normalizado [0,1];
d) classes detectáveis: 80 classes do *dataset* COCO;
e) formato de saída: detecções múltiplas com *bounding boxes*, *scores* de confiança e classes;
f) implementação: TensorFlow Lite (.tflite).

Processamento de Detecções: o modelo SSD-MobileNet v1 produz múltiplas detecções por imagem, cada uma contendo informações de localização (*bounding box*), classificação (classe do objeto) e confiança (*score*). O pós-processamento inclui filtragem por *threshold* de confiança configurável e aplicação de algoritmo de supressão de não-máximos (NMS) para eliminar detecções redundantes do mesmo objeto (Lin *et al.*, 2015).

Justificativa da Escolha: a seleção da arquitetura SSD-MobileNet v1 fundamenta-se na necessidade de sistemas ADAS *Mobile Telematics* que operam em tempo real. A natureza *single-shot* da arquitetura SSD permite detecção e localização de objetos em uma única inferência, resultando em latências reduzidas em comparação com abordagens de múltiplos estágios. A MobileNet v1 como *backbone* oferece extração eficiente de características com baixo custo computacional, utilizando convoluções separáveis em profundidade que reduzem significativamente o número de parâmetros e operações necessárias.

A escolha de utilizar a implementação oficial disponível no repositório TensorFlow Lite (LeViet, 2024), garante compatibilidade otimizada com o ecossistema Android e aproveita otimizações específicas para dispositivos móveis desenvolvidas pela equipe do TensorFlow.

4.5 Coleta e Tratamento dos Dados

Os resultados dos benchmarks foram exportados em formato digital (CSV), permitindo a consolidação e análise dos dados referentes à inferência, FPS, contagem de artefatos, temperatura e consumo dos dispositivos experimentais.

Para a aplicação personalizada, os dados foram coletados através de registros estruturados gerados durante a execução, incluindo *timestamps* para análise temporal do desempenho.

Os dados específicos da detecção de objetos incluem informações detalhadas sobre cada objeto detectado: classe identificada e *timestamp* da detecção. Adicionalmente, para cada intervalo de quadros, é registrada a contagem individual de objetos por classe (ex.: número de carros, pessoas, bicicletas detectadas), permitindo análises quantitativas específicas da densidade de diferentes tipos de objetos relevantes para cenários ADAS.

Para cada *timestamp* (1s) de quadros processados, número total de objetos detectados acima do *threshold* de confiança configurado (0,5) e contagem específica por classe de objeto (carros, pessoas, bicicletas, motocicletas, semáforos, etc.). Estes dados permitem análise granular do desempenho do *pipeline* de detecção, identificação de gargalos de processamento e avaliação quantitativa da capacidade de detecção em cenários típicos de trânsito urbano.

Exemplos de arquivos, capturas de tela das ferramentas e relatórios de execução estão disponíveis nos apêndices correspondentes. A análise dos dados foi feita com foco especial na avaliação da viabilidade da detecção de objetos em tempo real para aplicações ADAS *Mobile Telematics* utilizando a arquitetura SSD-MobileNet v1.

Com o objetivo de assegurar a confiabilidade e a validade dos resultados, realizou-se oito testes consecutivos para cada dispositivo em cada cenário avaliado. Os dados apresentados ao longo deste trabalho correspondem à média desses oito testes, sendo previamente analisados para identificação e remoção de possíveis *outliers*. Inicialmente, os experimentos utilizaram vídeos distintos, porém, convencionou-se a execução consecutiva dos mesmos vídeos em todas as repetições, o que permitiu avaliar de modo mais rigoroso o desempenho sustentado dos dispositivos. Observou-se que, nos testes de benchmarks padronizados (AI Benchmark e Burnout Benchmark), as variações entre as execuções foram mínimas, indicando alta estabilidade dos resultados. Já nos testes práticos com a aplicação ADAS, embora tenham sido identificadas pequenas diferenças entre as execuções, estas foram tratadas por meio de remoção de *outliers*

e posterior cálculo da média, conferindo maior confiabilidade às métricas reportadas para cada aparelho.

5 RESULTADOS

Neste capítulo, são detalhados os resultados da avaliação de desempenho em inferência de Edge AI, com ênfase nas diferenças impostas por hardware, *delegate*, sistema operacional e integração de software. As seções subsequentes apresentam as tabelas síntese, os rankings, as representações gráficas e uma discussão crítica dos principais achados do estudo.

5.1 Resultados da AI Benchmark

Esta seção apresenta, de forma sistemática, os resultados obtidos com a ferramenta AI Benchmark. Os testes englobaram diferentes arquiteturas de rede neural e múltiplos *delegates* de inferência, possibilitando uma análise quantitativa do impacto do hardware, da implementação dos *delegates* e do sistema operacional nos tempos de resposta e na eficiência computacional.

5.1.1 Gráficos Resumo dos Tempos de Inferência

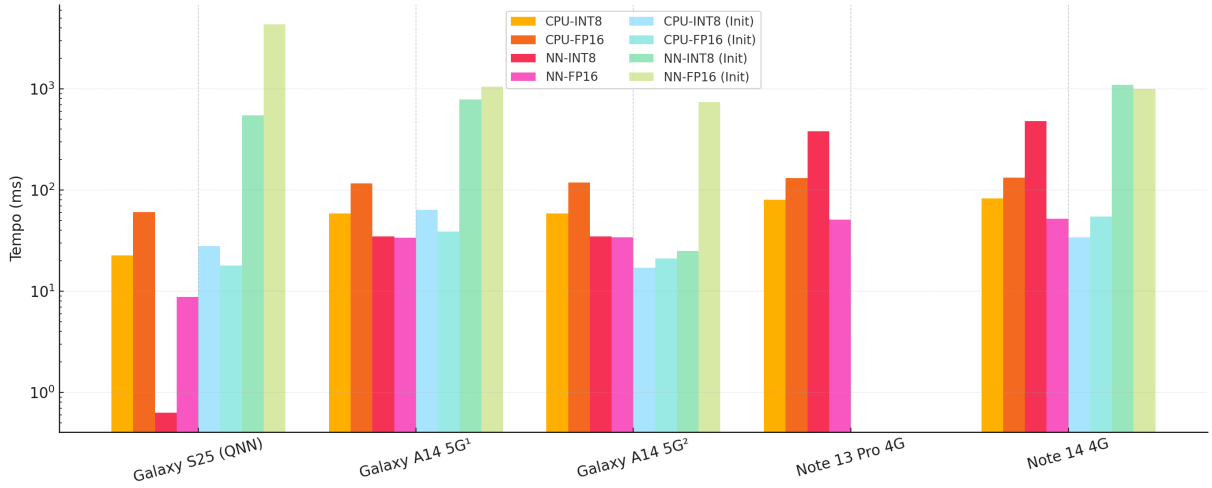
Os gráficos a seguir detalham os tempos de inferência em milissegundos para cada modelo de rede neural testado, onde valores menores indicam melhor desempenho. A análise foca na disparidade de performance entre o processamento via CPU, NNAPI e o *delegate* otimizado da Qualcomm (QNNHTP).

O Gráfico 5.1 apresenta os tempos de inferência em milissegundos para o modelo MobileNet-V3 em diferentes dispositivos e delegates, contemplando tanto a inferência sustentada quanto os tempos de inicialização. Valores menores indicam melhor desempenho, e a análise do gráfico destaca a disparidade de performance entre o processamento via CPU, NNAPI e o delegate otimizado da Qualcomm (QNNHTP).

Todos os dados apresentados foram obtidos por meio de testes locais realizados nos smartphones, exceto os resultados do Note 13 Pro 4G, que foram extraídos do site AI-Benchmark devido à indisponibilidade do dispositivo para medições diretas. Importante ressaltar que apenas os tempos de inicialização dos modelos não estão disponíveis para este aparelho, razão pela qual não aparecem no gráfico. A inclusão desses resultados se justifica pela similaridade das especificações técnicas e do processador em relação ao Note 14 4G, fornecendo assim uma base

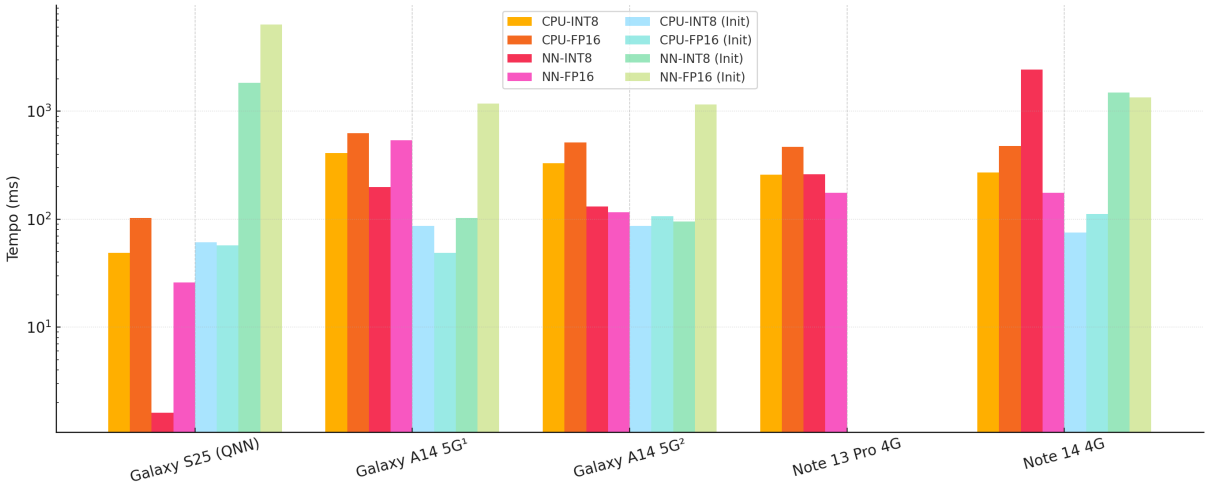
comparativa consistente. Para o Samsung Galaxy A14 5G[1], os testes foram conduzidos inicialmente com Android 14 e One UI 6, sendo posteriormente repetidos após a atualização para Android 15 e One UI 7[2].

Figura 5.1 – Tempos de Inferência e Inicialização - MobileNet-V3



Fonte: Elaborado pelo autor

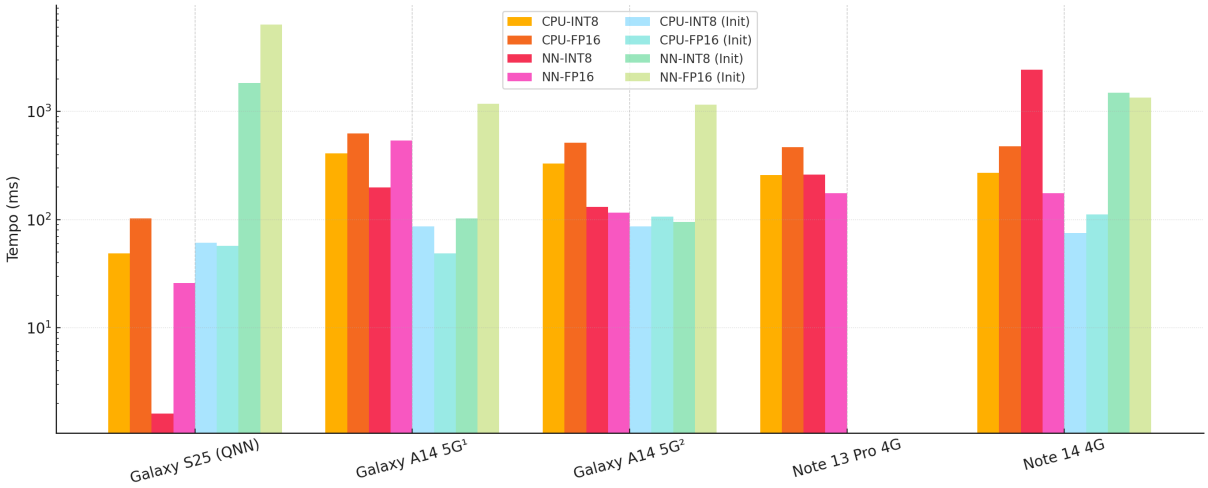
Figura 5.2 – Tempos de Inferência e Inicialização — EfficientNet-B4



Fonte: Elaborado pelo autor

A análise dos gráficos evidencia o abismo de desempenho entre o Galaxy S25 e os demais dispositivos testados. O S25, equipado com hardware dedicado e compatibilidade total com *delegates* otimizados como o QNN HTP, alcança tempos de inferência extremamente reduzidos para todos os modelos avaliados. Para MobileNet-V3 e EfficientNet-B4, é marcante a vantagem do uso do NNAPI aliado ao *delegate* Qualcomm: o tempo de processamento cai para

Figura 5.3 – Tempos de Inferência e Inicialização - DeepLab V3+



Fonte: Elaborado pelo autor

menos de 2 milissegundos em muitos cenários, tornando o dispositivo ideal para aplicações de IA em tempo real.

Nos smartphones intermediários (Note 13 Pro 4G e Note 14 4G) e de entrada (Galaxy A14 5G), os tempos de inferência são sensivelmente superiores, especialmente nas execuções via NNAPI, situação em que sempre ocorre *fallback* para os núcleos de CPU. Em alguns casos, observa-se até piora do desempenho ao tentar acionar supostos aceleradores, reflexo da ausência ou limitação de drivers e integração limitada das plataformas. Além disso, variações nas versões de sistema operacional e atualizações podem impactar os resultados, como verificado nas diferenças entre as versões de *firmware* testadas no Galaxy A14 5G.

Outro aspecto relevante é a diferença nos tempos de inicialização dos modelos, muito mais elevados quando se utiliza NNAPI em aparelhos intermediários e de entrada, chegando a superar, em certos casos, o próprio tempo de inferência em modelos mais simples. Essa limitação pode comprometer a experiência do usuário em cenários dinâmicos e reflete o grau de maturidade e otimização do ecossistema de IA em cada fabricante.

Todas as tabelas com os tempos de inferência detalhados para os modelos MobileNet-V3, EfficientNet-B4 e DeepLab V3+ estão reunidas no Apêndice B (Tabelas 1,2 e 3), servindo como referência para a análise dos gráficos apresentados ao longo desta subseção.

Por fim, vale destacar que, embora o foco desta análise esteja na performance bruta, resultados tão díspares ressaltam a importância não só do hardware, mas também do suporte de

software (sistema operacional, *delegates*, drivers e atualizações de plataforma) para a obtenção de desempenho adequado em tarefas complexas de IA em sistemas embarcados.

5.1.2 *Score* Global e Ranking

Para consolidar os múltiplos testes de inferência em uma única métrica comparativa, a AI Benchmark calcula um *score* global. Os gráficos a seguir apresentam essa pontuação para cada aparelho, oferecendo um ranking objetivo de desempenho em tarefas de IA. O *score* é composto principalmente pelas métricas de velocidade de processamento (AI Speed) e acurácia (AI Accuracy), proporcionando uma visão mais detalhada das capacidades de cada dispositivo.

Optou-se por dividir a visualização em dois gráficos devido à grande disparidade de desempenho observada entre o Galaxy S25 e os demais dispositivos testados. O Galaxy S25, equipado com hardware dedicado e suporte avançado a *delegates* otimizados, apresentou resultados significativamente superiores, o que "achata" a escala e dificulta a comparação visual quando todos os aparelhos são apresentados juntos. Dessa forma, um gráfico é dedicado exclusivamente ao S25 (Figura 5.4) e outro reúne os demais dispositivos (Figura 5.5), facilitando a análise individualizada e o destaque das diferenças relativas em cada grupo.

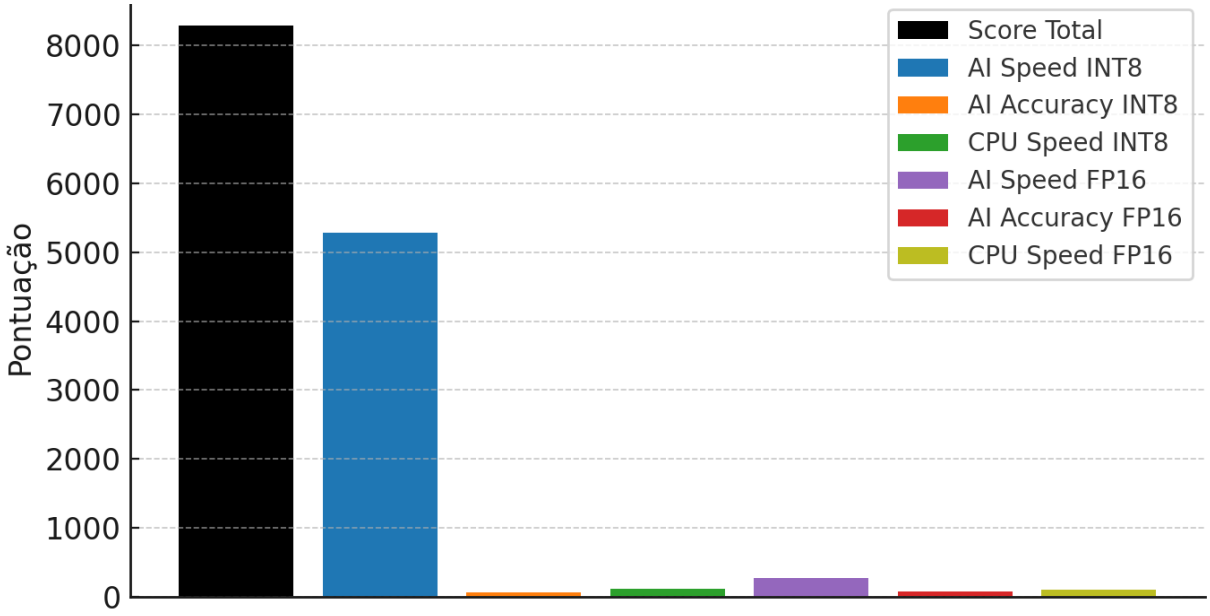
Os valores numéricos completos de cada métrica e aparelho encontram-se organizados na Tabela 4, disponível no Apêndice B.

5.1.3 Discussão dos Resultados da AI Benchmark

As análises quantitativas revelam que o desempenho de pico em inferência é ordens de magnitude superior no dispositivo *flagship* (Galaxy S25) quando comparado aos intermediários e de entrada. A utilização de hardware dedicado (NPU) e *delegates* otimizados (QNN HTP) constitui o fator determinante para essa disparidade, possibilitando latências na casa de poucos milissegundos, o que é essencial para aplicações em tempo real. Os demais, por sua vez, apresentam um desempenho consideravelmente mais baixo, limitado pelo poder de suas CPUs ou pela ausência de suporte otimizado para seus aceleradores. Contudo, como será detalhado a seguir, o desempenho de pico não se traduz diretamente em desempenho sustentado.

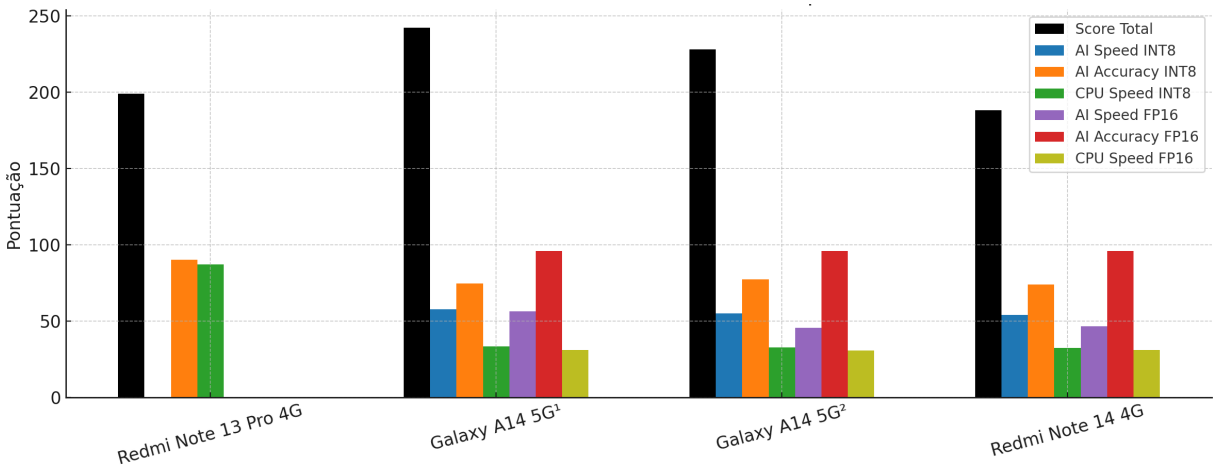
Mesmo entre os dispositivos de categorias diferentes, como o de entrada da Samsung Galaxy A14 5G (Exynos 1330) e o intermediário Redmi Note 14 4G (MediaTek Helio G99

Figura 5.4 – Score Total e Subscores — Galaxy S25 (QNN)



Fonte: Elaborado pelo autor

Figura 5.5 – Score Total e Subscores - Demais Dispositivos



Fonte: Elaborado pelo autor

Ultra), todos os testes de inferência foram processados exclusivamente na CPU. Durante os experimentos, foram feitas tentativas de execução por meio do NNAPI, com o objetivo de acionar a GPU do Exynos 1330 ou explorar a APU Helio G99 Ultra, além da solicitação do uso do *delegate* TFLite GPU Delegate. Em todos os casos, observou-se o comportamento de *fallback* para a CPU, ou seja, a execução permaneceu integralmente no processador central dos dispositivos, sem delegação das tarefas de inferência para aceleradores dedicados.

No caso do Redmi Note 14 4G (MediaTek Helio G99 Ultra), tal comportamento já era previsto com base nas informações disponíveis no site da AI Benchmark, principalmente em relação aos testes com o modelo Note 13 Pro. Conforme destacado nesses dados, mesmo quando configurado para utilização do NNAPI, o dispositivo adota o *backend* CPU como padrão, não aproveitando outros aceleradores existentes no hardware.

Outro achado relevante refere-se à comparação entre o Redmi Note 14 4G e o Redmi Note 13 Pro 4G, ambos equipados com o mesmo modelo de SoC. Apesar da similitude de hardware, os dados de desempenho obtidos em benchmarks comparativos disponíveis no site apresentaram diferenças significativas entre os dois modelos, principalmente na quantização para INT 8. Tal fenômeno pode ser explicado, em grande parte, pelas diferenças nos sistemas operacionais: o Note 14 4G executa o HyperOS, sistema que adota uma arquitetura baseada em microkernel, integrando parte do Android ao NuttX — um sistema operacional em tempo real com arquitetura de microkernel, utilizado por soluções como Vela OS — e ao Mina OS, um microkernel voltado à segurança (Xiaomi Corporation, 2024). O HyperOS também incorpora adaptações do kernel Vela, visando maior compatibilidade com dispositivos IoT da Xiaomi. Por outro lado, o Redmi Note 13 Pro utiliza a MIUI ROM baseada no Android 14, e preserva o kernel monolítico do Android.

Os resultados sugerem que o HyperOS, com Android 15, apresenta menor compatibilidade com quantização do TensorFlow Lite, promovendo perdas de desempenho em determinadas tarefas de inteligência artificial. Já a MIUI ROM com Android 14, por adotar uma estrutura mais próxima do Android tradicional, pode ampliar determinadas otimizações proporcionadas pelo uso de um kernel mais puro e maduro. Esses achados evidenciam que a escolha e o grau de maturidade do sistema operacional impactam sensivelmente o aproveitamento dos recursos de hardware em tarefas de IA, mesmo entre dispositivos que compartilham componentes idênticos.

No caso do Galaxy S25, observa-se um *score* global substancialmente acima dos demais dispositivos, refletindo o impacto direto das otimizações presentes na categoria *flagship*. Esse resultado pode ser atribuído ao uso eficiente de hardware dedicado para IA, especialmente a NPU integrada ao SoC de última geração, bem como ao suporte pleno a *frameworks* modernos de inferência, como o QNN HTP *delegate*. Essa combinação permite latências extremamente baixas e taxas de processamento elevadas mesmo sob cargas intensas de trabalho, características críticas para aplicações avançadas em tempo real, tais como sistemas automotivos, reconhecimento de imagens e assistentes pessoais inteligentes. Além disso, a presença de recursos de

software e *firmware* otimizados, aliados ao investimento em atualizações contínuas de drivers, posiciona o Galaxy S25 como um referencial de desempenho para inferência embarcada em dispositivos móveis de alto padrão.

Tais resultados reforçam a importância do hardware especializado e da maturidade do ecossistema de software para alcançar desempenho elevado em tarefas de inteligência artificial, diferenciando claramente os *flagships* dos modelos intermediários e de entrada.

5.2 Resultados da Burnout Benchmark

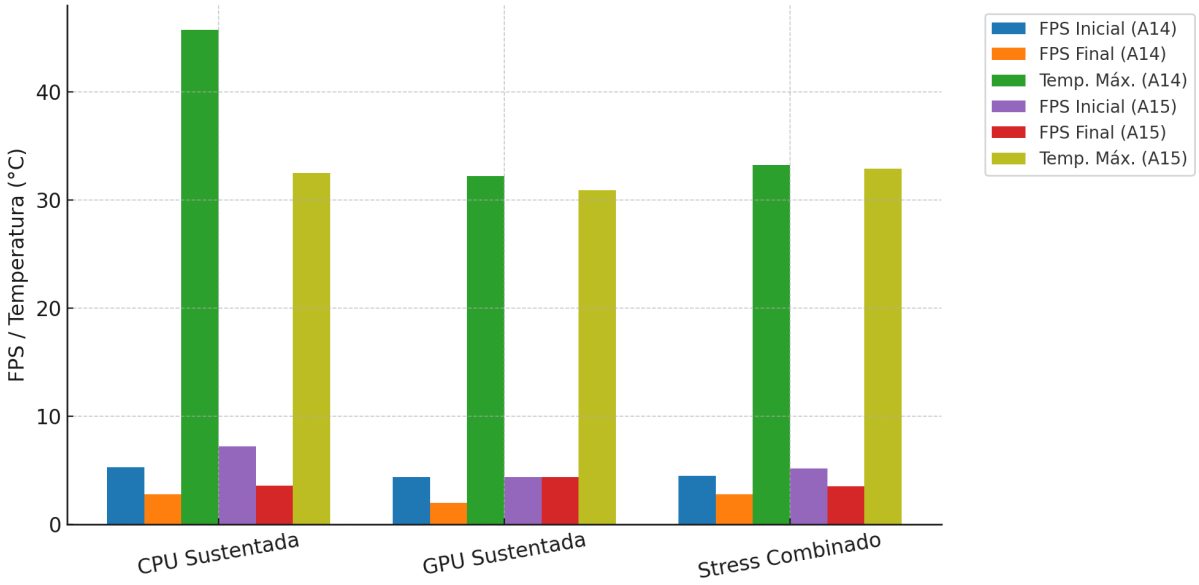
Nesta seção, a análise aprofunda-se no comportamento térmico, na degradação de desempenho e na eficiência energética dos dispositivos sob cargas prolongadas de inferência de IA. O objetivo é avaliar a resiliência de cada aparelho em cenários de uso intensivo e contínuo, como os exigidos por aplicações ADAS.

5.2.1 Resultados dos Testes de Estresse Térmico e Desempenho Sustentado

Nesta subseção, são apresentados os resultados detalhados de testes de estresse térmico e desempenho sustentado conduzidos com a Burnout Benchmark em cada dispositivo avaliado. As tabelas a seguir resumem as principais métricas coletadas: valores de FPS (quadros por segundo) inicial e final ao longo do estresse prolongado, bem como as temperaturas máximas registradas durante a execução dos testes. Os experimentos foram realizados com diferentes versões de sistema operacional, quando possível, e os dados são fundamentais para avaliar a resiliência térmica, a tendência de degradação de desempenho e o teto de performance sob uso intenso típico de aplicações ADAS.

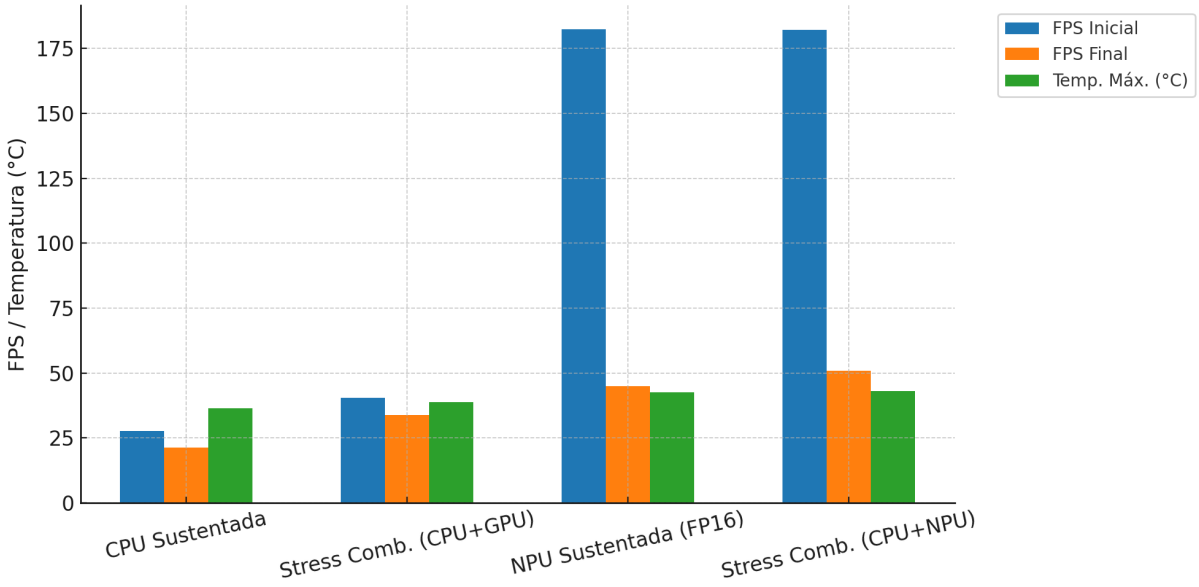
A análise dos dados evidencia padrões distintos de comportamento térmico e sustentação de desempenho entre os dispositivos avaliados. Galaxy A14 5G e Redmi Note 14 4G, revelam resiliência na sustentação do desempenho sob carga contínua, mas sempre em patamares inferiores aos do *flagship*. No caso do Galaxy A14 5G, a atualização do sistema operacional de Android 14 (A14) para 15 (A15), Figura 5.6 resultou em variações notáveis no perfil térmico, reduzindo significativamente a temperatura máxima registrada nos testes, especialmente na CPU, embora sem ganhos perceptíveis em desempenho de GPU.

Figura 5.6 – Testes de estresse térmico e desempenho sustentado — Galaxy A14 5G



Fonte: Elaborado pelo autor

Figura 5.7 – Resumo dos testes de estresse térmico e desempenho sustentado — Galaxy S25

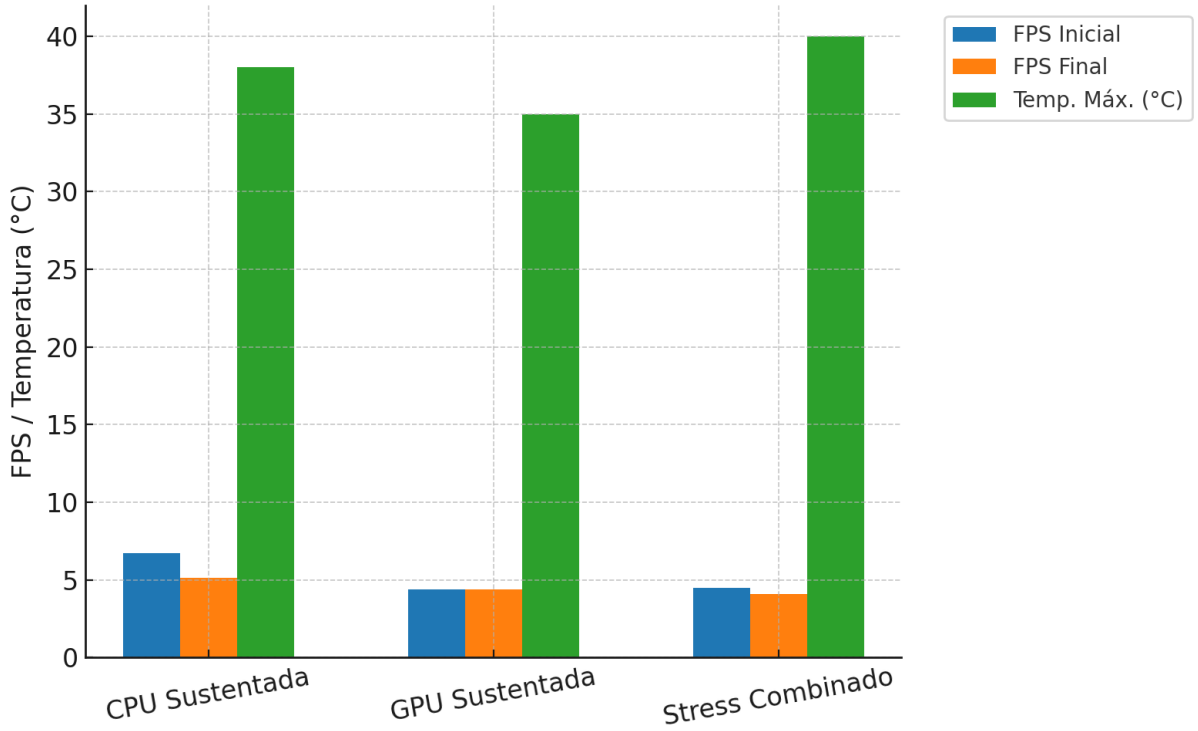


Fonte: Elaborado pelo autor

O Galaxy S25 (Figura 5.7) destaca-se com os maiores valores de FPS tanto em cenário de uso CPU+NPU quanto em testes combinados, demonstrando alto potencial para cargas intensas de inferência graças à presença de hardware dedicado e gerenciamento térmico mais eficiente. Em contrapartida, mesmo com este hardware avançado, observa-se uma queda acentuada no FPS ao longo do tempo e eventuais interrupções automáticas por superaquecimento

quando exigido ao extremo, reforçando a necessidade de soluções de controle térmico agressivas em sistemas embarcados de alto desempenho.

Figura 5.8 – Testes de estresse térmico e desempenho sustentado — Redmi Note 14 4G

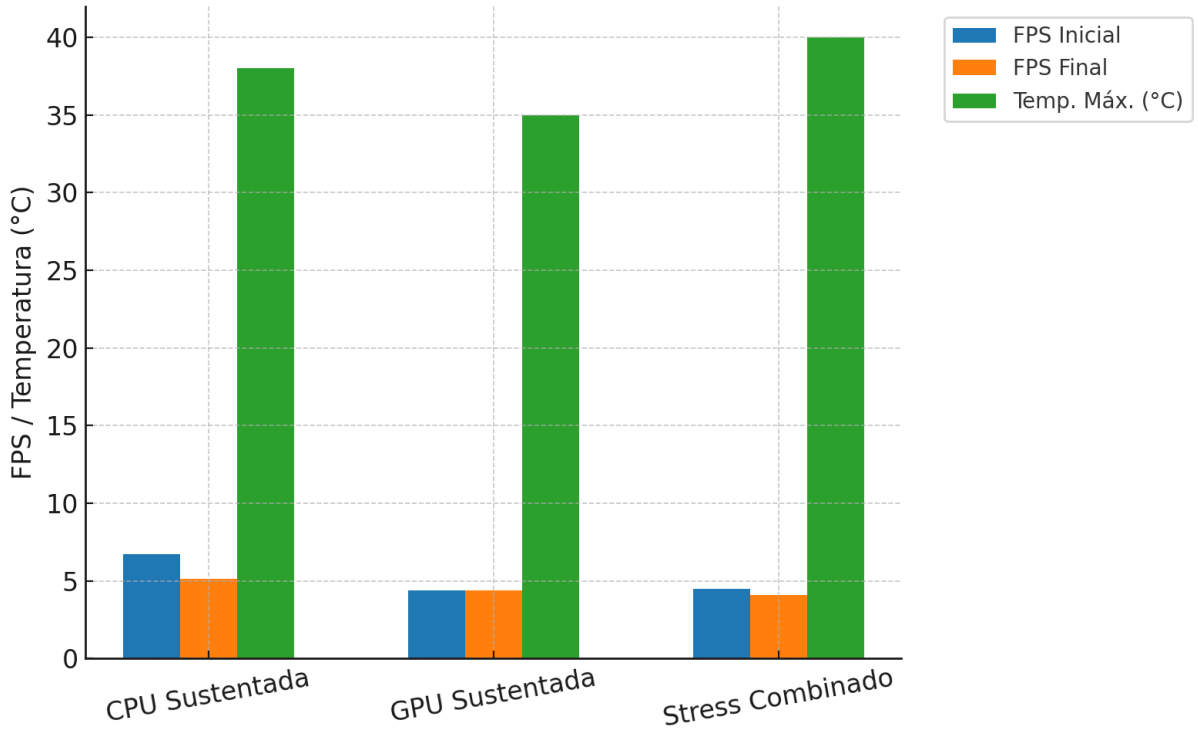


Fonte: Elaborado pelo autor

No Redmi Note 14 4G (Figura 5.9), o comportamento é mais estável, mas os valores de FPS permanecem limitados e o aquecimento é relativamente contido, compatível com o perfil de consumo energético do aparelho.

Esses resultados reforçam que tanto o potencial de processamento quanto a eficiência no controle térmico variam significativamente com a categoria do dispositivo e a maturidade do sistema operacional, impactando diretamente a viabilidade de aplicações embarcadas de inteligência artificial que exigem uso contínuo e resposta em tempo real. Os valores numéricos completos de cada métrica e aparelho encontram-se organizados nas Tabelas 5, 6 e 7, disponíveis no Apêndice C.

Figura 5.9 – Testes de estresse térmico e desempenho sustentado — Redmi Note 14 4G



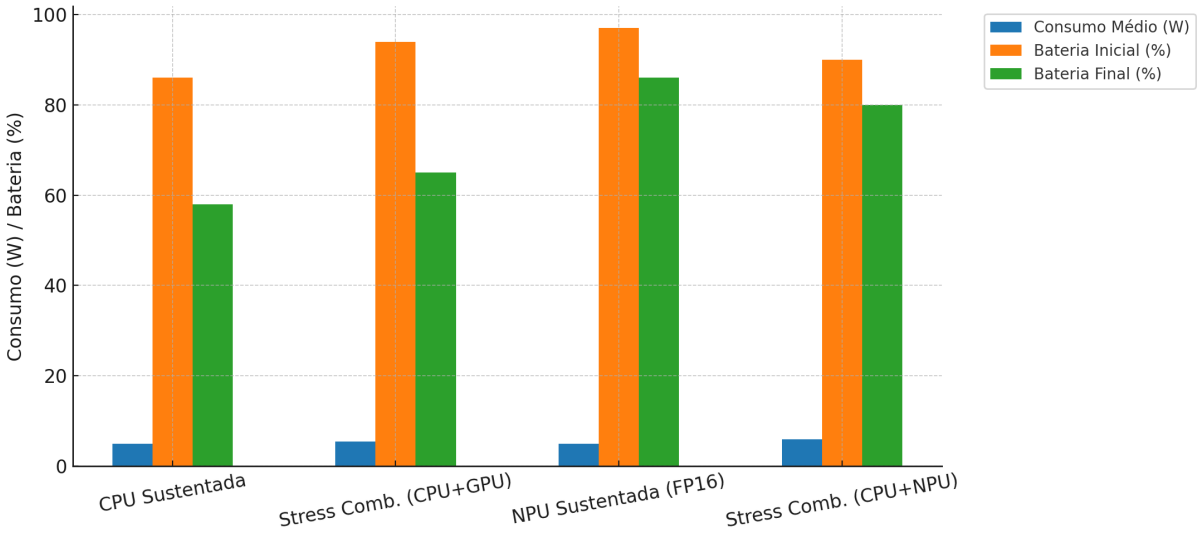
Fonte: Elaborado pelo autor

5.2.2 Resultados dos Testes de Consumo Energético Durante o Uso Prolongado

Nesta subseção, apresentam-se os resultados referentes ao comportamento energético dos dispositivos durante os testes prolongados de inferência de IA. As tabelas a seguir reúnem os dados de consumo instantâneo de potência (W) e a variação do estado de carga da bateria antes e após cada teste de estresse. Essa abordagem possibilita avaliar de forma comparativa a eficiência energética de cada aparelho, sua autonomia em cenários de uso intensivo e as diferenças entre versões de sistema operacional. Tais métricas são relevantes para aplicações embarcadas, especialmente em casos onde há restrição de fonte de alimentação.

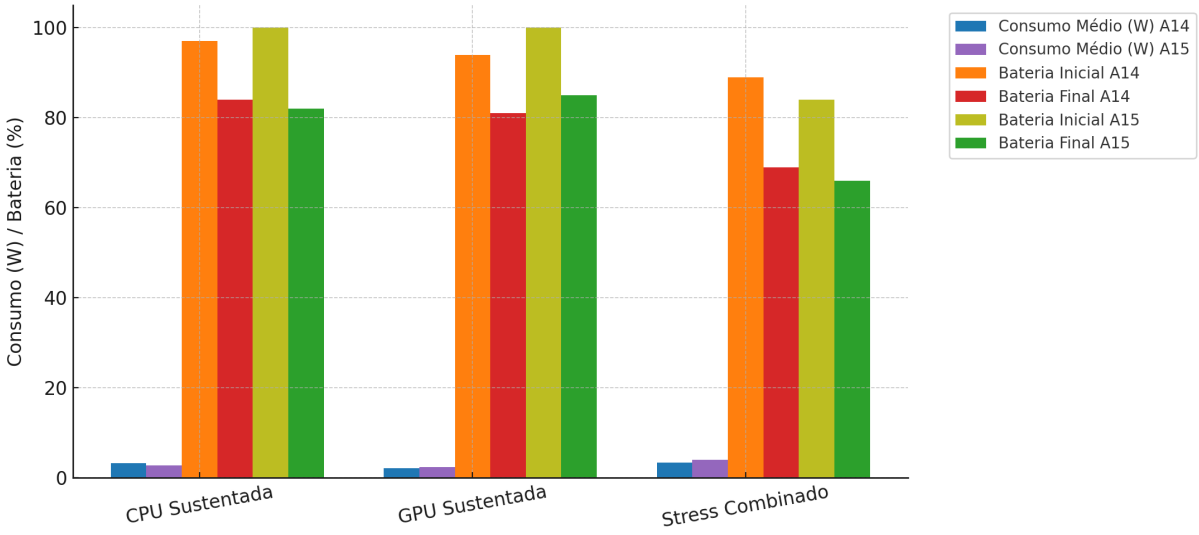
A análise dos dados revela diferenças marcantes no consumo energético e na autonomia entre os diferentes perfis de hardware dos aparelhos testados. O Galaxy S25 (Figura 5.10), embora ofereça o maior desempenho de inferência, também apresenta os maiores índices de consumo, especialmente em cenários que usam CPU, NPU ou GPU de forma combinada, resultando em quedas mais abruptas no nível de bateria. Esse resultado reflete o custo energético da alta capacidade computacional.

Figura 5.10 – Resumo dos testes de estresse térmico e desempenho sustentado — Galaxy S25



Fonte: Elaborado pelo autor

Figura 5.11 – Comportamento energético durante os testes prolongados — A14 5G

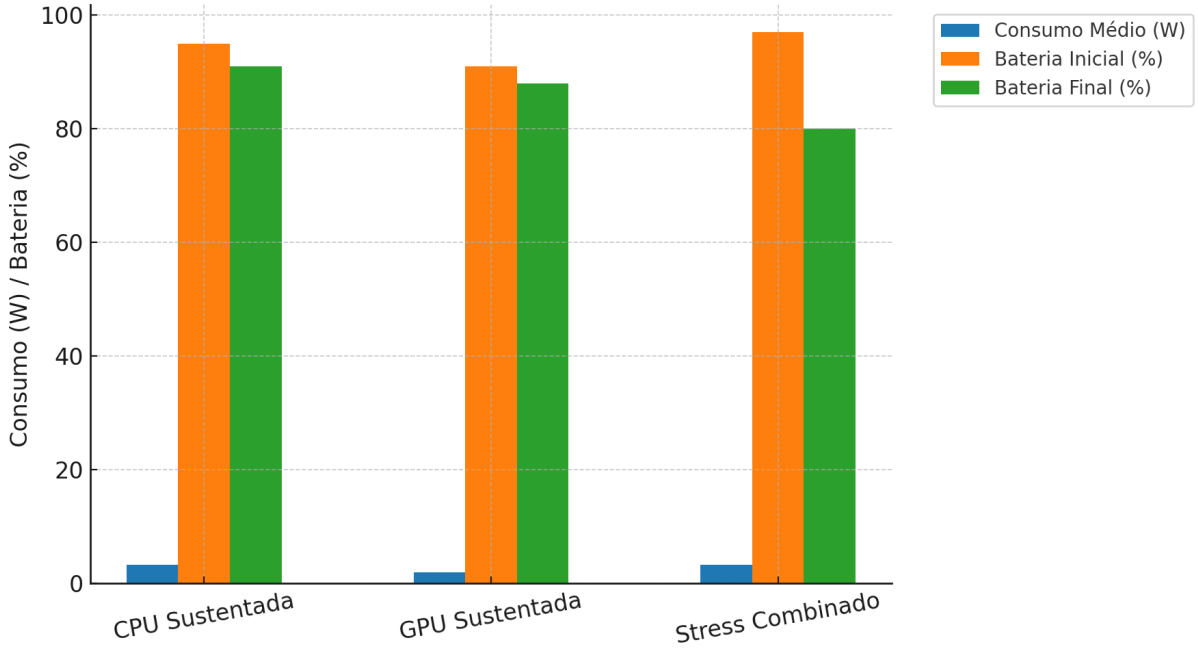


Fonte: Elaborado pelo autor

Já o smartphone intermediário mostra consumo mais moderado, ainda que com desgaste relevante de bateria em testes combinados. O Redmi Note 14 4G (Figura 5.12) se destaca pela eficiência em cargas via GPU, com consumo bem abaixo do S25. No Galaxy A14 5G (Figura 5.11), observa-se que a atualização do sistema operacional do Android 14 para o 15 trouxe leve redução do consumo e ganho de autonomia em alguns testes.

Esses achados reforçam a relação direta entre desempenho e demanda energética em dispositivos móveis: quanto maior a potência exigida, maior o consumo e o risco de aqueci-

Figura 5.12 – Testes de estresse térmico e desempenho sustentado — Redmi Note 14 4G



Fonte: Elaborado pelo autor

mento, o que reduz a autonomia. Para sistemas que dependem de operações de IA contínuas, o balanço entre autonomia e performance é fundamental na seleção do hardware e do perfil de uso. Os valores numéricos completos de cada métrica e aparelho encontram-se organizados nas Tabelas 5, 9 e 10, disponíveis no Apêndice C.

5.2.3 Discussão Geral dos Resultados da Burnout Benchmark

Os resultados evidenciam diversos fatores que afetam o desempenho sustentado e a viabilidade da inteligência artificial em smartphones Android em diferentes dispositivos móveis.

O equilíbrio entre potência computacional, controle térmico e eficiência energética varia conforme a arquitetura, sistema operacional e a estratégia de cada fabricante. Dispositivos de entrada, como o Galaxy A14 5G e intermediários como o Redmi Note 14 4G, ilustram abordagens opostas: o A14 5G entrega desempenho de pico, seguido de *throttling* para autoproteção, enquanto o Note 14 4G privilegia estabilidade e resiliência ao custo de menor desempenho máximo, favorecendo aplicações de longa duração.

Nos *flagships*, o uso de aceleradores dedicados como NPU impulsiona o desempenho, mas amplia desafios térmicos: a limitação do envelope térmico e o encerramento automático

em casos extremos indicam a importância da integração entre software, hardware e soluções eficientes de dissipação. Mesmo assim, a CPU e a GPU do S25 mantêm estabilidade sob carga mista.

A análise conjunta dos dados energéticos mostra que mais desempenho costuma significar mais consumo, afetando diretamente a autonomia. Em cenários onde a eficiência é crítica, casos de veículos autônomos, portáteis de saúde ou sistemas embarcados restritos por bateria, os intermediários como o Redmi Note 14 4G e de entrada como Galaxy A14 5G podem se encaixar melhor, ainda que enfrentem limitações de resposta em tempo real e *throughput*.

Atualizações de sistema operacional podem suavizar picos térmicos e melhorar autonomia, mas nem sempre trazem ganho real de performance. Destaca-se ainda a discrepância nas medições de temperatura entre sensores (bateria e SoC), evidenciando a cautela necessária ao interpretar benchmarks na prática.

Conclui-se que a escolha do dispositivo exige avaliação do perfil de carga, limites térmicos e energéticos, e capacidade real de sustentar desempenho ao longo do tempo. Buscar apenas desempenho máximo pode levar à instabilidade ou falha em cenários críticos, enquanto perfis conservadores favorecem resiliência e previsibilidade, mesmo diante de demandas exigentes.

5.3 Resultados da Aplicação ADAS

Esta seção apresenta a análise comparativa do desempenho dos três dispositivos em aplicação prática de sistemas ADAS, evidenciando as diferenças arquiteturais e suas implicações em cenários reais de uso.

A Tabela 5.1 sintetiza os principais indicadores de desempenho observados durante a execução prática da aplicação ADAS nos três dispositivos avaliados. São apresentados dados referentes ao tempo de teste, quantidade de *frames* processados, desempenho térmico, latências de inferência, taxas de quadros e consumo energético ao longo do experimento. Esses resultados permitem uma comparação direta da eficiência computacional, da robustez térmica e do impacto energético entre diferentes faixas de hardware, fornecendo subsídios objetivos para a análise das limitações e potencialidades de cada plataforma em cenários de uso real.

Já a Tabela 5.2 apresenta o total de detecções de objetos relevantes para sistemas ADAS realizadas por cada dispositivo durante os testes. Os dados estão organizados por classe de objeto, contemplando veículos, elementos urbanos e sinalizações presentes no ambiente de cir-

Tabela 5.1 – Desempenho comparativo da aplicação ADAS nos três dispositivos

Métrica	Galaxy S25	Redmi Note 14 4G	Galaxy A14 5G
Duração do teste	96,6 min	96,9 min	96,5 min
Frames processados	173.455	107.021	53.254
Temperatura do SoC (°C)			
Inicial	52,3	–	–
Final	62,6	–	–
Elevação térmica	10,3	–	–
Tempo de inferência			
Médio	24,3 ms	57,3 ms	119,7 ms
Mínimo	12 ms	37 ms	33 ms
Máximo	61 ms	151 ms	374 ms
Taxa de quadros (FPS)			
Média	29,9 fps	18,4 fps	9,2 fps
Mínima	8 fps	4 fps	2 fps
Máxima	48 fps	24 fps	24 fps
Consumo de bateria			
Total	39%	22%	33%
Por minuto	0,40%	0,23%	0,34%
Temperatura da bateria (°C)			
Inicial	31,9	23,0	25,2
Final	39,0	36,0	27,8
Elevação térmica	+7,1	+13,0	+2,6

Fonte: Elaborado pelo autor

culação. Esta comparação evidencia a capacidade de cada plataforma em identificar, em condições reais, os principais elementos que compõem o contexto viário, oferecendo subsídios para a avaliação da eficácia prática de cada solução embarcada.

Tabela 5.2 – Detecções de objetos relevantes para ADAS – Comparativo por dispositivo

Classe de Objeto	Galaxy S25	Redmi Note 14 4G	Galaxy A14 5G
Veículos			
Carros	14.301	14.846	17.234
Caminhões	52	34	39
Motocicletas	18	5	4
Elementos urbanos			
Pessoas	5.274	4.497	2.231
Semáforos	418	1.568	2.422
Placas de sinalização	0	0	0
Total de detecções	**20.063**	**20.950**	**21.930**

Fonte: Elaborado pelo autor

5.3.1 Análise Qualitativa das Detecções

A análise qualitativa dos resultados evidencia que o número absoluto de detecções reportado por cada dispositivo não pode ser interpretado como medida direta de performance prática ou fidelidade na identificação de objetos únicos ao longo dos testes. Isso se deve, principalmente, à ausência de algoritmos de *tracking* temporal na abordagem adotada, o que potencializa fenômenos de recontagem de um mesmo objeto em múltiplos quadros, sobretudo nos cenários em que a taxa de quadros é reduzida.

Além disso, a discrepância entre as quantidades de memória RAM disponíveis em cada dispositivo — 4 GB no Galaxy A14 5G, 8 GB no Redmi Note 14 4G e 12 GB no Galaxy S25 — impacta diretamente o potencial para a implementação de estratégias mais sofisticadas de associação temporal e rastreamento multiobjeto, restringindo especialmente os modelos de entrada a abordagens mais conservadoras de detecção. Como consequência, o Galaxy A14 5G apresentou menores taxas de processamento e a maior taxa contagem de objetos, exibiu maior número de detecções totais, influenciado tanto por sua configuração de memória quanto pela recontagem de objetos causada pela ausência de *tracking*.

No caso do Galaxy S25, apesar de sua maior capacidade de processamento e memória possibilitarem taxas elevadas de processamento e estabilidade térmica, a limitação metodológica imposta pela não utilização de validação *ground-truth* e de algoritmos de *tracking* temporal impede qualquer afirmação conclusiva sobre a aproximação de sua contagem ao número real de objetos distintos presentes na cena.

Dessa forma, as variações observadas entre os dispositivos refletem fundamentalmente a interação entre taxa de quadros, capacidade computacional, memória disponível e ausência de rastreamento temporal, não podendo ser interpretadas isoladamente como métricas de acurácia absoluta. Esta limitação será discutida em maiores detalhes na seção específica sobre limitações metodológicas e perspectivas futuras.

5.3.2 Limitações da Metodologia de Contagem

A metodologia empregada, baseada em processamento quadro independente sem algoritmos de *tracking* temporal, representa uma simplificação que evidencia fenômenos fundamentais, mas não reflete implementações comerciais otimizadas. Sistemas ADAS práticos implementam algoritmos sofisticados de *multi-object tracking* que mitigam artefatos de contagem temporal observados.

A utilização de *threshold* único de 50% para todos os dispositivos, embora garanta comparabilidade direta, pode não otimizar individualmente a relação precisão-*recall* para cada arquitetura específica. Dispositivos com diferentes características de inferência podem beneficiar-se de ajustes específicos de *threshold* que maximizem a qualidade das detecções.

A ausência de validação *ground-truth* impede a quantificação precisa de taxas de falsos positivos e falsos negativos, limitando a análise qualitativa das detecções. Esta limitação afeta particularmente a interpretação de diferenças no volume total de detecções entre dispositivos.

5.4 Síntese dos Resultados e Contribuições

A avaliação dos três dispositivos operando exclusivamente via CPU trouxe avanços para a compreensão dos limites e potencialidades de sistemas ADAS em plataformas móveis. Os experimentos demonstraram que o volume bruto de detecções não garante, por si só, melhor performance prática, especialmente na ausência de técnicas de *tracking* temporal. Essa obser-

vação desafia a tradição de se adotar métricas quantitativas simples como critério exclusivo de avaliação em visão computacional em sistemas embarcados.

5.4.1 Precisão Temporal e Robustez Prática

Os resultados apontam que a continuidade temporal do processamento é determinante para a confiabilidade em contagem de objetos únicos, sendo um critério mais relevante que a soma total de detecções. Dispositivos capazes de manter intervalos de processamento inferiores a 50 ms, como o Galaxy S25, tendem a preservar maior fidelidade temporal, reduzindo a incidência de recontagens. Já dispositivos intermediários e de entrada, com intervalos superiores a 100 ms, apresentaram aumento notável em artefatos de recontagem, afetando a utilidade dos dados para aplicações críticas.

5.4.2 Limitações de Aceleração por Hardware

A impossibilidade de acionar aceleradores dedicados em todos os dispositivos testados reforça a importância de validação prática, além das especificações técnicas. A dependência exclusiva de CPU evidenciou disparidades de desempenho entre arquiteturas ARM de diferentes segmentos, ressaltando a necessidade de análises com cargas de trabalho reais. Essa limitação impacta diretamente a seleção de hardware para projetos comerciais, pois demonstra que disponibilidade nominal de aceleradores não assegura sua utilização efetiva na prática.

5.4.3 Categorias de Aplicação e Diretrizes

***Premium*:** O Galaxy S25 demonstrou condições para aplicações ADAS críticas, com latências baixas e continuidade temporal robusta.

Intermediário: O Redmi Note 14 4G mostrou-se viável para cenários urbanos e monitoramento prolongado, ainda que com precisão temporal limitada.

Entrada: O Galaxy A14 5G, devido à descontinuidade acentuada e restrições de memória, restringe-se a aplicações básicas, *offline* ou de caráter experimental.

5.4.4 Considerações Finais

Os achados apresentados servem como base para a escolha de plataformas móveis em projetos ADAS, enfatizando que testes realísticos e análise do comportamento temporal são essenciais. Contudo, limitações metodológicas — em especial, a ausência de validação *ground-truth* e a não utilização de algoritmos de *tracking* temporal — indicam que os resultados não devem ser interpretados como medida definitiva de acurácia ou desempenho absoluto.

Recomenda-se, para trabalhos futuros, a integração de *pipelines* completos com *tracking* multiobjeto, validação quantitativa com *ground-truth* e testes sob diferentes cenários operacionais, a fim de ampliar a robustez das conclusões e a aplicabilidade prática das soluções desenvolvidas.

6 CONCLUSÃO

Este estudo apresentou uma análise da eficiência de smartphones na execução de algoritmos de DL embarcados para aplicações de ADAS, com ênfase no desempenho computacional, térmico e energético. Utilizando benchmarks padronizados (AI Benchmark e Burnout Benchmark) e experimentos práticos com uma aplicação baseada em SSD-MobileNet v1, foi possível estabelecer um panorama claro das capacidades e limitações dos dispositivos avaliados em cenários reais de uso.

Os resultados evidenciaram disparidades marcantes entre os aparelhos de diferentes categorias de hardware. O Galaxy S25 destacou-se, consolidando-se como a única opção viável para aplicações ADAS críticas em tempo real, com latências médias de inferência inferiores a 30 ms e excelente robustez térmica ao longo de todo o teste. O Redmi Note 14 4G revelou desempenho intermediário: embora apresente eficiência energética superior e desempenho razoável, sua taxa de quadros mais baixa impactou negativamente a precisão temporal, aumentando a incidência de artefatos de recontagem de objetos. O Galaxy A14 5G, por sua vez, confirmou as limitações severas dos modelos de entrada: com taxas de quadros muito baixas e processamento exclusivamente via CPU, demonstrou ser inadequado para ADAS em tempo real, sendo aceitável apenas em aplicações de monitoramento básico ou análise offline.

Uma limitação importante identificada foi a impossibilidade de ativação dos *delegates* de aceleração por hardware (NNAPI ou GPU) em todos os dispositivos testados, restringindo a avaliação ao uso exclusivo da CPU. Esta limitação ressalta a importância de investigações futuras com *firmwares* otimizados e maior acesso a recursos nativos dos SoCs, seja por Android puro ou SDKs dos fabricantes, para explorar o potencial dos aceleradores dedicados.

A análise dos resultados de detecção de objetos demonstrou que o maior volume de detecções não se traduz necessariamente em melhor desempenho prático, devido à ausência de algoritmos de *tracking* temporal e à influência da taxa de quadros sobre a contagem efetiva de objetos únicos. Sem validação por *ground-truth*, não é possível afirmar com rigor qual dispositivo apresentou a contagem mais precisa – o que reforça a necessidade de adoção de metodologias mais robustas, incluindo anotação manual e algoritmos de rastreamento em experimentos futuros.

Destaca-se, ainda, a ausência de inovação disruptiva em dissipação térmica e baterias no mercado de smartphones, apesar do avanço incremental de hardware. Produtos como smartpho-

nes *gamers* com ventoinhas e sistemas de refrigeração ativa como Loop LiquidCool - Xiaomi (ROG Phone, Mi Mix 4) são exceções que não chegaram ao *mainstream*, mantendo o segmento estagnado em termos de eficiência energética e gerenciamento térmico – fatores cruciais para a viabilidade de Edge AI embarcada.

Portanto, recomenda-se para pesquisas e implementações futuras:

a) adoção de modelos otimizados e algoritmos de *multi-object tracking* para melhor correlação temporal;
b) ajustes finos nos limiares de confiança conforme o perfil do hardware;
c) utilização de *frameworks* e SDKs nativos para ativação efetiva de aceleradores;
d) validação dos resultados com *ground-truth* manual e análise qualitativa detalhada.

Por fim, a acentuada heterogeneidade do ecossistema Android, tanto em hardware quanto em software, ainda representa um desafio para a padronização e replicabilidade de resultados. A utilização futura de dispositivos Apple, seja via LiteRT ou Neural Engine, pode oferecer um ambiente mais homogêneo e facilitar comparações sistemáticas.

Este estudo contribui para o entendimento das limitações e possibilidades dos smartphones como plataformas para ADAS baseados em *Mobile Telematics*, estabelecendo parâmetros para avaliação de eficiência, robustez e impacto prático. Ao evidenciar os gargalos existentes, reforça-se a necessidade de inovação real em hardware, especialmente em soluções térmicas e energéticas, para a democratização desses sistemas.

REFERÊNCIAS

AFRAZ, J.; HARITHA, K. **Hype Cycle for Artificial Intelligence, 2024**. [*S.l.*], 2024. Accessed on: 24-11-2024. Disponível em: https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence.

Agência Nacional de Transportes Terrestres. **Boletim Anual de Sinistros de Trânsito em Rodovias 2023**. Brasília, 2023. Superintendência de Infraestrutura Rodoviária: Roger da Silva Pêgas. Gerência de Fiscalização de Infraestrutura e Operação Rodoviária: José Luis Vianna Ferreira.

BOCHKOVSKIY, A.; WANG, C.-Y.; LIAO, H.-Y. M. **YOLOv4: Optimal Speed and Accuracy of Object Detection**. 2020. Disponível em: https://arxiv.org/abs/2004.10934.

CARVALHO, C. H. R. de. **Balanço da 1ª década de ação pela segurança no trânsito no Brasil e perspectivas para a 2ª década**. [*S.l.*], 2023.

CEDOM, C. de Documentação e Memória do M. S. Acidentes de trânsito no brasil: de 2006 a 2025, muitas mudanças. **Notícias do Seguro**, mar. 2025. Disponível em: https://www.editoraroncarati.com.br/v2/Artigos-e-Noticias/Artigos-e-Noticias/Acidentes-de-transito-no-Brasil-de-2006-a-2025-muitas-mudancas.html. Acesso em: 01 jun. 2025.

CHEN, C. *et al.* Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. **IEEE Transactions on Intelligent Transportation Systems**, v. 24, n. 11, p. 13023–13034, 2023.

CHEN, L.-C. *et al.* Encoder-decoder with atrous separable convolution for semantic image segmentation. In: FERRARI, V. *et al.* (Ed.). **Computer Vision – ECCV 2018**. Cham: Springer International Publishing, 2018. p. 833–851. ISBN 978-3-030-01234-2.

CHEN, L.-C. *et al.* Encoder-decoder with atrous separable convolution for semantic image segmentation. In: FERRARI, V. *et al.* (Ed.). **Computer Vision – ECCV 2018**. Cham: Springer International Publishing, 2018. p. 833–851. ISBN 978-3-030-01234-2.

DO, H. N. *et al.* Automatic license plate recognition using mobile device. In: **2016 International Conference on Advanced Technologies for Communications (ATC)**. [*S.l.: s.n.*], 2016. p. 268–271.

Google. **Android Neural Networks API (NNAPI)**. 2017. https://developer.android.com/ndk/guides/neuralnetworks. Acessado em: Acessado em: [23 ago. 2024].

Google. **TensorFlow Lite**. 2017. https://www.tensorflow.org/lite. Acessado em: [23 de ago. 2024].

Google AI. **Compatibilidade do operador LiteRT e do TensorFlow**. 2024. Acesso em: 1 jun. 2025. Disponível em: https://ai.google.dev/edge/litert/models/ops_compatibility?hl=pt-br.

GUAN, L. *et al.* A lightweight framework for obstacle detection in the railway image based on fast region proposal and improved yolo-tiny network. **IEEE Transactions on Instrumentation and Measurement**, v. 71, p. 1–16, 2022.

HAN, S. *et al.* Learning both weights and connections for efficient neural networks. In: **Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1**. Cambridge, MA, USA: MIT Press, 2015. (NIPS'15), p. 1135–1143.

HINA, M. D. *et al.* CASA: An Alternative Smartphone-Based ADAS. **International Journal of Information Technology and Decision Making**, World Scientific Publishing, v. 21, n. 01, p. 273–313, set. 2021. Disponível em: https://hal.science/hal-04489867.

HiSilicon. **Kirin 970**. 2021. Online. Acesso em: 14 jun. 2025. Disponível em: https://www.hisilicon.com/en/products/kirin/kirin-flagship-chips/kirin-970.

HOWARD, A. *et al.* **Searching for MobileNetV3**. 2019. Disponível em: https://arxiv.org/abs/1905.02244.

HOWARD, A. G. *et al.* **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications**. 2017. Disponível em: https://arxiv.org/abs/1704.04861.

IGNATOV, A. *et al.* Fast and accurate single-image depth estimation on mobile devices, mobile ai 2021 challenge: Report. In: **2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [*S.l.: s.n.*], 2021. p. 2545–2557.

IGNATOV, A. *et al.* Microisp: Processing 32mp photos on mobile devices with deep learning. In: KARLINSKY, L.; MICHAELI, T.; NISHINO, K. (Ed.). **Computer Vision – ECCV 2022 Workshops**. Cham: Springer Nature Switzerland, 2023. p. 729–746. ISBN 978-3-031-25063-7.

IGNATOV, A. *et al.* Ai benchmark: All about deep learning on smartphones in 2019. In: **2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)**. [*S.l.: s.n.*], 2019. p. 3617–3635.

IGNATOV, A. D. **Burnout Benchmark**. 2025. https://burnout-benchmark.com/download.html. Versão 2.0.8; Última atualização: 17 de março de 2025.

IGNATOV, A. D. *et al.* Ai benchmark: Running deep neural networks on android smartphones. In: **ECCV Workshops**. [*s.n.*], 2018. Disponível em: https://api.semanticscholar.org/CorpusID:52910608.

Intel. **Transitioning from Intel MKL-DNN to oneDNN**. 2024. Online. Acesso em: 14 jun. 2025. Disponível em: https://www.intel.com/content/www/us/en/docs/onednn/developer-guide-reference/2024-1/transitioning-from-intel-mkl-dnn-to-onednn.html.

JACOB, B. *et al.* Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [*S.l.: s.n.*], 2018. p. 2704–2713.

JIANG, Y. *et al.* Yolov4-dense: A smaller and faster yolov4 for real-time edge-device based object detection in traffic scene. **IET Image Processing**, v. 17, n. 2, p. 570–580, 2023. Disponível em: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12656.

KATARE, D. *et al.* A survey on approximate edge ai for energy efficient autonomous driving services. **Commun. Surveys Tuts.**, IEEE Press, v. 25, n. 4, p. 2714–2754, ago. 2023. ISSN 1553-877X. Disponível em: https://doi.org/10.1109/COMST.2023.3302474.

KHANDABATTU, H. **The 2025 Hype Cycle for Artificial Intelligence Goes Beyond GenAI**. 2025. Accessed: 2025-07-12. Disponível em: https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence.

KRISHNAMOORTHI, R. Quantizing deep convolutional networks for efficient inference: A whitepaper. **CoRR**, abs/1806.08342, 2018.

LEVIET, K. **Build and deploy a custom object detection model with TensorFlow Lite (Android)**. 2024. Google Developers Codelab. Last updated September 18, 2024. Interactive tutorial on training and deploying custom object detection models using TensorFlow Lite Model Maker and Task Library. Disponível em: https://developers.google.com/codelabs/tflite-object-detection-android.

LI, H. *et al.* **Pruning Filters for Efficient ConvNets**. 2017. Disponível em: https://arxiv.org/abs/1608.08710.

LI, P. *et al.* Multi-model running latency optimization in an edge computing paradigm. **Sensors**, v. 22, n. 16, 2022. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/22/16/6097.

LIN, T.-Y. *et al.* **Microsoft COCO: Common Objects in Context**. 2015. Dataset contains 330K images with 200K annotated for object detection, segmentation and captioning tasks across 80 object categories. Disponível em: https://cocodataset.org/.

LIU, Z. *et al.* Rethinking the value of network pruning. In: **International Conference on Learning Representations**. [*s.n.*], 2019. Disponível em: https://openreview.net/forum?id=rJlnB3C5Ym.

MASELLO, L. *et al.* On the road safety benefits of advanced driver assistance systems in different driving contexts. **Transportation Research Interdisciplinary Perspectives**, v. 15, p. 100670, 09 2022.

MediaTek. **NeuroPilot: MediaTek's Ecosystem for AI Development**. 2022. Online. Acesso em: 14 jun. 2025. Disponível em: https://neuropilot.mediatek.com/.

MediaTek Inc. **MediaTek NeuroPilot SDK**. 2019. https://neuropilot.mediatek.com. Acessado em: [23 de ago. 2024].

Meta AI. **PyTorch Mobile**. 2019. https://pytorch.org/mobile. Acessado em: [23 de ago. 2024].

MOLCHANOV, P. *et al.* **Pruning Convolutional Neural Networks for Resource Efficient Inference**. 2017. Disponível em: https://arxiv.org/abs/1611.06440.

Mundo Logística. **Pesquisa revela os 7 países do mundo mais seguros para dirigir**. MundoLogística, 2024. https://mundologistica.com.br/noticias/pesquisa-revela-sete-paises-do-mundo-mais-seguros-para-dirigir. Acessado em: [2024-12-07]. Disponível em: https://mundologistica.com.br/noticias/pesquisa-revela-sete-paises-do-mundo-mais-seguros-para-dirigir.

MUSA, S. S. *et al.* Convergence of information-centric networks and edge intelligence for iov: Challenges and future directions. **Future Internet**, v. 14, n. 7, 2022. ISSN 1999-5903. Disponível em: https://www.mdpi.com/1999-5903/14/7/192.

NORONHA, F. de A. *et al.* Detecção de fadiga a partir da análise de imagens faciais. In: **Colloquium Exactarum. ISSN: 2178-8332**. [*S.l.: s.n.*], 2019. v. 11, n. 2, p. 34–45.

NVIDIA. **cuDNN: NVIDIA CUDA Deep Neural Network Library**. 2025. Online. Acesso em: 14 jun. 2025. Disponível em: https://developer.nvidia.com/cudnn.

OZ, W. **BERLIN - Germany 4K Driving Tour 2023 | Potsdamer Platz, Charlottenburg, Wilmersdorf, Spandau**. 2023. https://www.youtube.com/watch?v=ehsiTo3v3rQ. Vídeo, 40 minutos. Disponível em: YouTube. Acesso em: 28 jun. 2025.

Qualcomm. **Qualcomm 5G NR Royalty Terms Statement**. 2017. https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/qualcomm-5g-nr-royalty-terms-statement.pdf. Qualcomm Incorporated disclosed a framework for industry participants to access Qualcomm's patented inventions used in 3GPP 5G New Radio standards. It includes licensing terms for patents up to release 15 of 3GPP specifications.

Qualcomm Technologies, Inc. **Qualcomm Neural Processing SDK for AI**. 2017. https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk. Acessado em: [23 de ago. 2024].

REDMON, J.; FARHADI, A. **YOLOv3: An Incremental Improvement**. 2018. Disponível em: https://arxiv.org/abs/1804.02767.

RENAULT. **Sistema de Assistência ao condutor (ADAS)**. 2024. Disponível em: https://www.renault.com.br/programahumanfirst/adas.html.

SAE International. **Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles**. SAE International, 2021. 41 p. Revised 2021-04-30. Disponível em: https://www.sae.org/standards/content/j3016_202104/.

Samsung. **Processor: The core that redefines your device**. 2022. Online. Acesso em: 14 jun. 2025. Disponível em: https://semiconductor.samsung.com/processor/.

Samsung Electronics Co., Ltd. **Samsung Neural SDK**. 2018. https://developer.samsung.com/neural/overview.html. Acessado em: [23 de ago. 2024].

SANDLER, M. *et al.* **MobileNetV2: Inverted Residuals and Linear Bottlenecks**. 2019. Disponível em: https://arxiv.org/abs/1801.04381.

SHEN, J.; LIAO, H.; ZHENG, L. A lightweight method for small scale traffic sign detection based on YOLOv4-Tiny. **Multimedia Tools and Applications**, v. 83, n. 40, p. 88387–88409, 12 2024. ISSN 1573-7721. Disponível em: https://doi.org/10.1007/s11042-023-17146-3.

STATISTA. **Newly Registered Cars by Autonomous Driving Level**. 2024. Accessed: 2024-05-27. Disponível em: https://www.statista.com/chart/25754/newly-registered-cars-by-autonomous-driving-level/.

SZEGEDY, C. *et al.* Inception-v4, inception-resnet and the impact of residual connections on learning. In: **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**. [*S.l.*]: AAAI Press, 2017. (AAAI'17), p. 4278–4284.

SZEGEDY, C. *et al.* Rethinking the inception architecture for computer vision. **CoRR**, abs/1512.00567, 2015.

TABANI, H. *et al.* Adbench: Benchmarking autonomous driving systems. **Computing**, Springer, v. 104, p. 1–32, 2021. Disponível em: https://upcommons.upc.edu/bitstream/handle/2117/351444/benchmark-springer.pdf.

TAN, M.; LE, Q. V. **EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks**. 2020. Disponível em: https://arxiv.org/abs/1905.11946.

THEIVADAS, J. R.; PONNAN, S. Vigileye: Machine learning-powered driver fatigue recognition for safer roads. **Measurement: Sensors**, v. 33, p. 101186, 2024. ISSN 2665-9174. Disponível em: https://www.sciencedirect.com/science/article/pii/S2665917424001624.

Unisoc. **UNISOC TAIoT: AIoT Development Platform**. 2020. Online. Acesso em: 14 jun. 2025. Disponível em: https://www.unisoc.com/en_us/sch/TAIoT.

UTAH, J. **Chicago 4K - Night Drive - Driving Downtown**. 2023. https://www.youtube.com/watch?v=9Prtp_eNUiI. Vídeo, 52 minutos. Disponível em: YouTube. Acesso em: 28 jun. 2025.

WANG, Y. *et al.* Knowledge distillation for fast and accurate monocular depth estimation on mobile devices. In: **2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [*S.l.: s.n.*], 2021. p. 2457–2465.

WEIG, F. W. Advanced driver-assistance systems: Challenges and opportunities ahead. In: **McKinsey's**. [*s.n.*], 2016. Disponível em: https://api.semanticscholar.org/CorpusID:214778522.

Xiaomi Corporation. **Xiaomi HyperOS: Overview and Features**. 2024. https://www.mi.com/us/hyperos. Acesso em: 6 jun. 2025.

ZHANG, Z. *et al.* A simple baseline for fast and accurate depth estimation on mobile devices. In: **2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [*S.l.: s.n.*], 2021. p. 2466–2471.

ZHOU, A. *et al.* **Incremental Network Quantization: Towards Lossless CNNs with Low-Precision Weights**. 2017. Disponível em: https://arxiv.org/abs/1702.03044.

ZURICH, E. **Is Your Smartphone Ready For AI?** 2024. Accessed: 2024-12-08. Disponível em: https://ai-benchmark.com\}.

ZURICH, E. **Is Your Smartphone Ready For AI?** 2024. Accessed: 2024-12-08. Disponível em: https://burnout-benchmark.com.

APÊNDICES

APÊNDICE A – ESPECIFICAÇÕES DOS DISPOSITIVOS

Os dispositivos analisados diferem de forma significativa em suas arquiteturas de hardware e *stacks* de software. Abaixo, as principais características técnicas de cada aparelho empregado nos testes, servindo como base para a interpretação dos resultados de desempenho subsequentes.

Quadro A.1 – Especificações técnicas dos dispositivos testados

Especificação	**Galaxy S25**	**Galaxy A14 5G**	**Redmi Note 13 Pro 4G / 14 4G**
SoC	Snapdragon 8 Elite for Galaxy	Exynos 1330	MediaTek Helio G99 Ultra
Node	3 nm (TSMC N3E)	5 nm (Samsung)	6 nm (TSMC)
CPU	2×Oryon V2 L (4,47 GHz) 6×Oryon V2 M (3,53 GHz)	2×Cortex-A78 (2,4 GHz) 6×Cortex-A55 (2,0 GHz)	2×Cortex-A76 (2,2 GHz) 6×Cortex-A55 (2,0 GHz)
GPU	Adreno 830	Mali-G68 MP2	Mali-G57 MC2
NPU/APU	Hexagon/QNN HTP	Não possui	Sim (APU)
RAM	12 GB LPDDR5X	4 GB LPDDR4X	8 GB LPDDR4X
Android/ROM	15 (One UI 7)	14 (One UI 6)	14 (MIUI 14) / 15 (HyperOS)

Fonte: HW Info e documentação dos fabricantes.

APÊNDICE B – TABELAS RESUMO DOS TEMPOS DE INFERÊNCIA

Este apêndice apresenta as tabelas nas quais detalham os tempos de inferência em milissegundos para cada modelo de rede neural testado, onde valores menores indicam melhor desempenho. A análise foca na disparidade de performance entre o processamento via CPU, NNAPI e o delegate otimizado da Qualcomm (QNNHTP).

Tabela 1 – Tempos de inferência (ms) — MobileNet-V3

Dispositivo/*Delegate*	**CPU-INT8**	**CPU-FP16**	**NN-INT8**	**NN-FP16**
Galaxy S25 (QNN)	**22,5**	**60,4**	**0,63**	**8,76**
Inicialização	28	18	548	4362
Galaxy A14 5G[1]	58,6	117	34,8	33,7
Inicialização	64	39	789	1049
Galaxy A14 5G[2]	58,9	119	35,0	34,0
Inicialização	17	21	25	739
Note 13 Pro 4G	80	131	383	51
Note 14 4G	**82,4**	**133**	**483**	**51,9**
Inicialização	34	55	1099	999

Fonte: Elaborado pelo autor

Tabela 2 – Tempos de inferência (ms) — EfficientNet-B4

Dispositivo/*Delegate*	**CPU-INT8**	**CPU-FP16**	**NN-INT8**	**NN-FP16**
Galaxy S25 (QNN)	**48,6**	**103**	**1,61**	**26,0**
Inicialização	61	57	1833	6384
Galaxy A14 5G[1]	**411**	**625**	198	538
Inicialização	87	49	103	1175
Galaxy A14 5G[2]	331	513	131	116
Inicialização	87	107	95	1155
Note 13 Pro 4G	258	469	260	175
Note 14 4G	271	479	**2434**	175
Inicialização	75	112	1492	**1341**

Fonte: Elaborado pelo autor

Tabela 3 – Tempos de inferência (ms) — DeepLab V3+

Dispositivo/*Delegate*	**CPU-INT8**	**CPU-FP16**	**NN-INT8**	**NN-FP16**
Galaxy S25 (QNN)	**106**	**232**	**2,0**	**31,5**
Inicialização	48	91	1963	6193
Galaxy A14 5G[1]	**617**	**1239**	670	**1728**
Inicialização	22	63	721	640
Galaxy A14 5G[2]	437	1127	404	391
Inicialização	<10	18	78	599
Note 13 Pro 4G	551	1014	553	499
Inicialização	68	126	1001	885
Note 14 4G	557	1021	**4960**	499
Inicialização	68	126	1001	885

Fonte: Elaborado pelo autor

Tabela 4 – Ranking dos AI *Scores* globais segundo AI Benchmark

Dispositivo	**INT8**			**FP16**		
	AI Speed	**AI Accuracy (%)**	**CPU Speed**	**AI Speed**	**AI Accuracy (%)**	**CPU Speed**
Galaxy S25 (QNN) *Score* Total: **8293**	**5286**	**69,9**	**119,0**	**278**	**72,3**	**109,2**
Redmi Note 13 Pro *Score* Total: 199	–	**90,0**	87,0	–	–	–
Galaxy A14 5G[1] *Score* Total: 242	57,7	74,7	33,4	56,5	95,9	30,9
Redmi Note 14 4G *Score* Total: **188**	**54,0**	74,0	**32,3**	46,6	95,9	30,9
Galaxy A14 5G[2] *Score* Total: 228	55,1	77,4	32,8	**45,4**	95,9	**30,8**

Fonte: Elaborado pelo autor

APÊNDICE C – RESULTADOS DA BURNOUT BENCHMARK

Este apêndice apresenta os resultados consolidados dos testes de benchmarking de IA realizados nos diferentes dispositivos avaliados ao longo deste trabalho. São reunidos aqui os AI Scores globais, calculados a partir do conjunto de métricas fornecidas pelo AI Benchmark, incluindo velocidade de inferência (*AI Speed*), acurácia (*AI Accuracy*) e desempenho do processador (*CPU Speed*) em operações INT8 e FP16.

Tabela 5 – Testes de estresse térmico e desempenho sustentado — Galaxy A14 5G

	Android 14			**Android 15**		
	FPS_{ini}	**FPS_{fin}**	**T_{max} (°C)**	**FPS_{ini}**	**FPS_{fin}**	**T_{max} (°C)**
CPU Sustentada	5,3	2,8	45,7	7,2	3,6	32,5
GPU Sustentada	4,4	2,0	32,2	4,4	4,4	30,9
Stress Combinado	4,5	2,8	33,2	5,2	3,5	32,9

Fonte: Elaborado pelo autor

Tabela 6 – Resumo dos testes de estresse térmico e desempenho sustentado — Galaxy S25

	FPS_{ini}	**FPS_{fin}**	**T_{max} (°C)***
CPU Sustentada	27,6	21,3	36,5
Stress Combinado (CPU+GPU)	40,5	33,7	38,8
NPU Sustentada (FP16)	182,5	44,9	42,6
Stress Combinado (CPU+NPU)	182,2	50,8†	43,1

Fonte: Elaborado pelo autor

* Temperatura registrada nos arquivos de registros, referente à bateria. Ver discussão no texto.
† O teste foi interrompido pelo sistema por superaquecimento.

Tabela 7 – Testes de estresse térmico e desempenho sustentado — Redmi Note 14 4G

	FPS_{ini}	**FPS_{fin}**	**T_{max} (°C)**
CPU Sustentada	6,7	5,1	38,0
GPU Sustentada	4,4	4,4	35,0
Stress Combinado	4,5	4,1	40,0

Fonte: Elaborado pelo autor

Tabela 8 – Comportamento energético durante os testes prolongados — A14 5G

	Consumo (W)	**Bateria (%)**		**Consumo (W)**	**Bateria (%)**	
		ini	**fin**		**ini**	**fin**
CPU Sustentada	3,0–3,6	97	84	2,1–3,3	100	82
GPU Sustentada	1,9–2,3	94	81	1,6–3,2	100	85
Stress Combinado	3,2–3,6	89	69	1,3–6,7	84	66

Fonte: Elaborado pelo autor

Tabela 9 – Comportamento energético durante os testes prolongados — Galaxy S25

	Consumo	**Bateria$_{ini}$**	**Bateria$_{fin}$**
CPU Sustentada	4,5–5,5 W	86%	58%
Stress Combinado (CPU+GPU)	5,0–6,0 W	94%	65%
NPU Sustentada (FP16)	4,0–6,0 W	97%	86%
Stress Combinado (CPU+NPU)	5,0–7,0 W	90%	80%

Fonte: Elaborado pelo autor

Tabela 10 – Comportamento energético durante os testes prolongados — Redmi Note 14 4G

	Consumo	**Bateria$_{ini}$**	**Bateria$_{fin}$**
CPU Sustentada	2,4–4,1 W	95%	91%
GPU Sustentada	1,7–2,1 W	91%	88%
Stress Combinado	3,1–3,5 W	97%	80%

Fonte: Elaborado pelo autor

APÊNDICE D – CENÁRIO EXPERIMENTAL

Este apêndice apresenta fotografias do ambiente experimental e dos dispositivos utilizados nos testes do cenário diurno e noturno. Os percursos completos podem ser consultados nos vídeos de referências disponíveis em (OZ, 2023) e (Utah, 2023).

Figura 1 – Ambiente experimental durante o período diurno.

Fonte: Do Autor

Figura 2 – Ambiente experimental durante o período noturno.

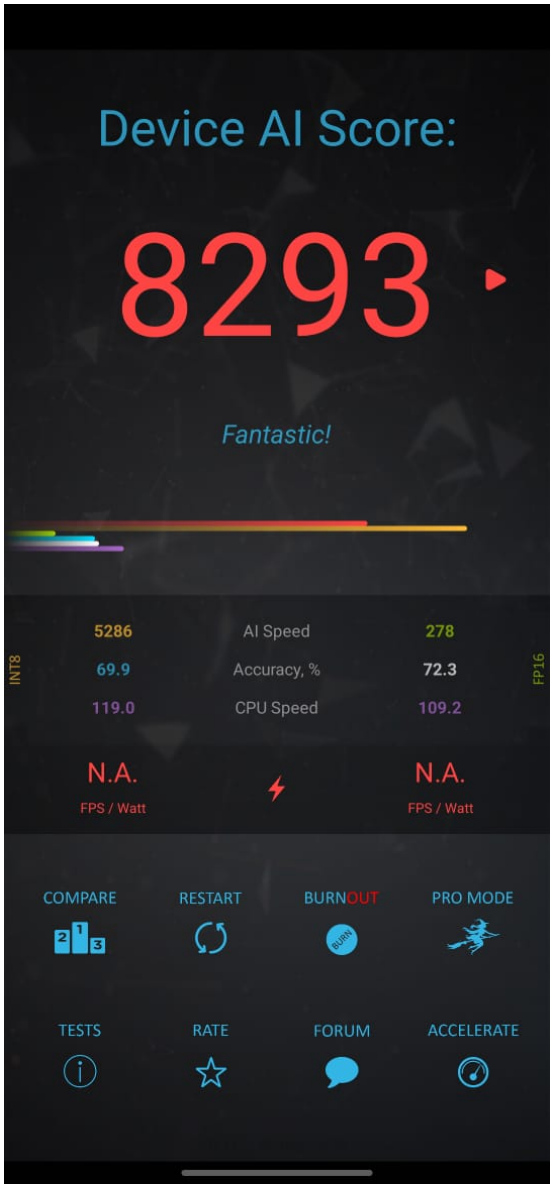
Fonte: Do Autor

APÊNDICE E – EXEMPLOS DOS RESULTADOS E *PRINTS* DAS FERRAMENTAS

Este apêndice contém capturas de tela das ferramentas AI Benchmark, Burnout Benchmark, da aplicação própria Android, bem como exemplos dos arquivos CSV gerados nos experimentos.

AI Benchmark

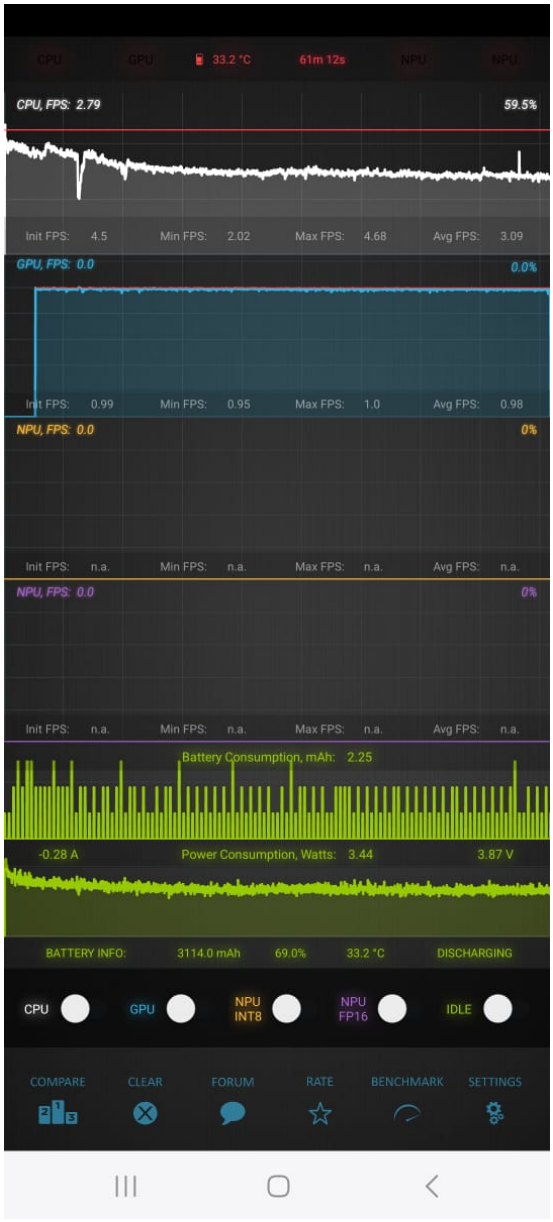
Figura 3 – *Print* da tela de resultados do AI Benchmark - S25.



Fonte: Do Autor

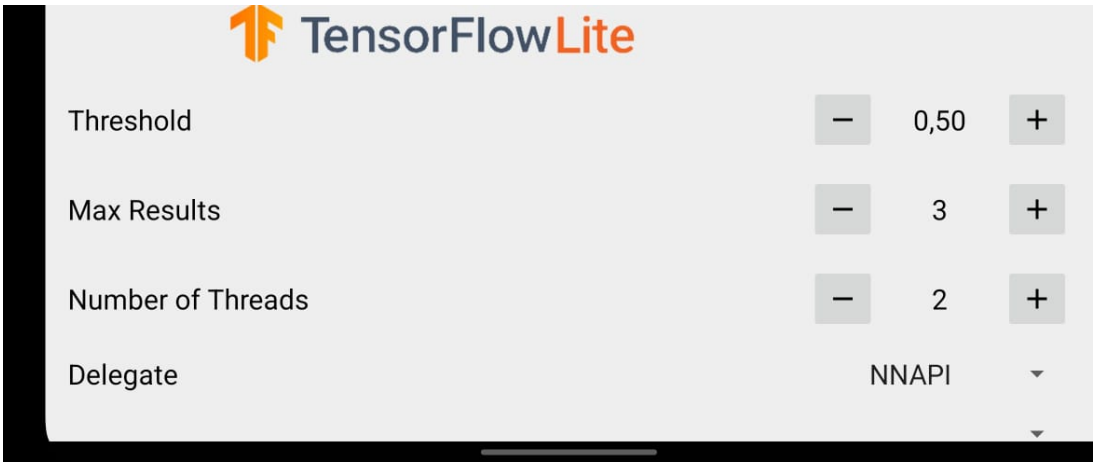
Burnout Benchmark

Figura 4 – *Print* da tela do Burnout Benchmark - A14.



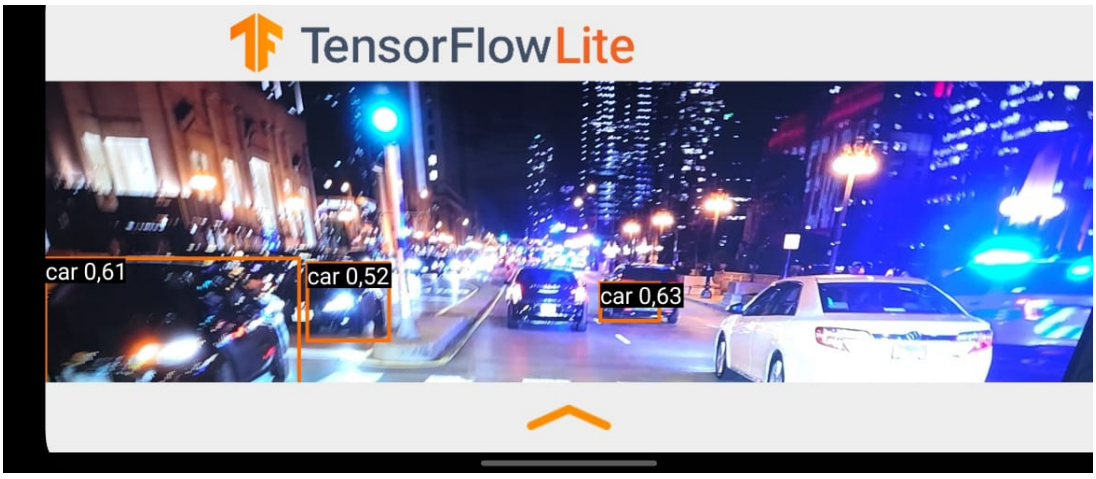
Fonte: Do Autor

Figura 5 – Tela de início da aplicação Android.



Fonte: Do autor

Figura 6 – Tela de execução da aplicação Android.



Fonte: Do autor

Aplicação Android Própria

Exemplo de arquivo CSV exportado

Abaixo alguns dos dados do arquivo CSV criado pelo aplicativo. De forma similar, são gerados CSV tanto pela AI Benchmark quanto pela Burnout Benchmark.

Figura 7 – *Print* da tela do Burnout Benchmark - A14.

timestamp	inference_time_ms	fps	battery_level_percent	battery_temp_c	count_person	count_bicycle	count_car	count_motorcycle	count_bus	count_train	count_truck	count_traffic_light
2025-06-29 20:44:55.029	24.0	29.94012	66	40.5	1	0	0	0	0	0	0	0
2025-06-29 20:45:00.030	23.0	29.92278	66	40.5	0	0	5	0	0	0	0	0
2025-06-29 20:45:05.032	28.0	30.0	66	40.5	0	0	3	0	0	0	0	0
2025-06-29 20:45:10.030	25.0	30.185007	66	40.5	0	0	3	0	0	0	0	0
2025-06-29 20:45:15.028	22.0	30.155642	66	40.5	0	0	1	0	0	0	0	0
2025-06-29 20:45:20.029	27.0	29.951693	66	40.5	0	0	5	0	0	0	0	0
2025-06-29 20:45:25.031	24.0	30.30303	66	40.5	3	0	2	0	0	0	0	0
2025-06-29 20:45:30.034	24.0	29.970028	66	40.5	0	0	1	0	0	0	0	0
2025-06-29 20:45:35.032	23.0	29.70297	66	40.5	1	0	0	0	0	0	0	0
2025-06-29 20:45:40.036	25.0	29.807693	66	40.5	0	0	1	0	0	0	0	0
2025-06-29 20:45:45.028	24.0	29.970028	66	40.5	2	0	1	0	0	0	0	0
2025-06-29 20:45:50.027	22.0	29.88048	66	40.5	2	0	2	0	0	0	0	0
2025-06-29 20:45:55.028	24.0	30.155642	66	40.5	1	0	3	0	0	0	0	0
2025-06-29 20:46:00.033	27.0	29.615005	66	40.5	1	0	4	0	0	0	0	0
2025-06-29 20:46:05.030	21.0	30.185007	66	40.5	0	0	5	0	0	0	0	0
2025-06-29 20:46:10.029	24.0	29.721956	66	40.5	0	0	5	0	0	0	0	0
2025-06-29 20:46:15.028	20.0	29.970028	66	40.5	0	0	3	0	0	0	0	0
2025-06-29 20:46:20.035	24.0	29.79146	66	40.5	0	0	5	0	0	0	0	0
2025-06-29 20:46:25.027	24.0	30.185007	66	40.5	0	0	5	0	0	0	0	0
2025-06-29 20:46:30.033	25.0	29.980656	66	40.5	0	0	5	0	0	0	0	0

Fonte: Do autor