www.revistaft.com.br

Revista FT | ISSN 1678-0817 | v. 30, n. 158, 2026 | DOI: 10.69849/ye94ex82

Inteligência artificial de borda para detecção de armas brancas: arquitetura de alta disponibilidade com supervisão humana na PMPA

Edge artificial intelligence for detecting bladed weapons: high-availability architecture with human supervision in the pmpa

Rogério Fernandes Oliveira[1]
Eder Bruno Bezerra Barros[2]
Daniel Andrade da Silva[3]

RESUMO

A realidade operacional da Polícia Militar do Pará (PMPA) envolvendo armas brancas, marcada pela apreensão de 7.679 unidades em 2024, exige sistemas de vigilância mais eficientes para apoiar a atuação policial em tempo real. Este artigo propõe o desenvolvimento de um sistema de visão computacional baseado em inteligência artificial para detecção automática de tais artefatos, arquitetado para operar em servidores de borda (Edge Server) dentro da estrutura da PMPA, permitindo identificação com baixa latência e maior confiabilidade. O sistema é concebido como ferramenta de apoio sob supervisão humana, em que a atuação do policial é solicitada apenas quando identificadas potenciais ameaças, reduzindo a necessidade de monitoramento constante e, consequentemente, a fadiga cognitiva dos operadores, fator que poderia tornar a detecção demorada e sujeita a falhas. Assim, o sistema preserva a confiabilidade da ocorrência por meio da filtragem dos falsos alarmes. A metodologia envolveu o treinamento de um modelo YOLO26l com um dataset de 2.356 imagens pós-sanitização, focado em cenários reais de baixa luminosidade e oclusão parcial, empregando estratégias de balanceamento de dados e o uso intensivo de imagens negativas

[1]Especialista em Internet das Coisas - IOT. Bacharel em Engenharia de Computação pelo Centro Universitário Internacional (UNINTER). Castanhal, Pará, Brasil. E-mail: grupocastanhal@hotmail.com. ORCID: 0009-0001-9455-7269.
[2]Especialista em Pedagogia do Esporte. Graduado em Educação Física pela Universidade Norte do Paraná (UNOPAR). Castanhal, Pará, Brasil. E-mail: edersd17@gmail.com
[3]Graduado em Licenciatura Plena em Educação Física pela (UNOPAR). Belém, Pará, Brasil. E-mail: dansilva81@hotmail

(Hard Negative Mining), representando 28,4% do total, para reduzir falsos alarmes. Os resultados demonstraram uma precisão média de 97,19% (mAP50) com processamento de 28,9 FPS. Tais índices comprovam a eficiência da ferramenta como um multiplicador de força, permitindo expandir a capacidade de vigilância sem a necessidade de aumento do efetivo.

Palavras-chave: Inteligência Artificial. PMPA. Detecção de Armas. Edge Server.Segurança Pública.

ABSTRACT

The operational reality of the Military Police of Pará (PMPA) involving bladed weapons, marked by the seizure of 7,679 units in 2024, demands more efficient surveillance systems to support real-time police action. This article proposes the development of a computer vision system based on artificial intelligence for the automatic detection of such artifacts, architected to operate on edge servers within the PMPA structure, allowing identification with low latency and greater reliability. The system is conceived as a support tool under human supervision, where police action is only requested when potential threats are identified, reducing the need for constant monitoring and, consequently, the cognitive fatigue of operators, a factor that could make detection time-consuming and prone to errors, preserving the reliability of the incident by filtering out false alarms. The methodology involved training a YOLO26l model with a dataset of 2,356 post-sanitization images, focused on real-world scenarios of low light and partial occlusion, employing data balancing strategies and the intensive use of negative images (Hard Negative Mining), representing 28.4% of the total, to reduce false alarms. The results demonstrated an average accuracy of 97.19% (mAP50) with a processing speed of 28.9 FPS. These figures prove the tool's efficiency as a force multiplier, allowing for expanded surveillance capacity without the need to increase personnel.

Keywords: Artificial Intelligence. PMPA. Weapon Detection. Edge Server. Public Security.

1 INTRODUÇÃO

A atuação da PMPA na preservação da ordem pública e da incolumidade das pessoas enfrenta desafios que extrapolam o Policiamento Ostensivo Geral (POG), demandando o uso de tecnologias modernas para potencializar o emprego tático da tropa no terreno. A circulação de armas brancas (objetos cortantes, perfurantes ou contundentes) no Estado do Pará é corroborada por dados robustos presentes no Relatório de Produtividade de 2024, emitido pelo Departamento-Geral de Operações (DGO), que registra a retirada de circulação, pela corporação, de 7.679 unidades de armas brancas no referido ano. Do ponto de vista legal, o Art. 19 do Decreto-Lei nº 3.688, de 3 de outubro de 1941 (Lei de Contravenções Penais) (BRASIL, 1941) esclarece a conduta esperada da tropa: embora o porte de arma branca não seja totalmente proibido, constitui infração quando a potencialidade lesiva se evidencia nas circunstâncias do caso concreto, associada à intenção do agente. Diante disso, a detecção antecipada desses objetos é um fator crítico para a segurança do cidadão e, sobretudo, para a integridade do policial militar que atua na linha de frente.

Concomitantemente ao enfrentamento da circulação ilícita de armas brancas, o emprego do efetivo em centros de monitoramento torna-se um desafio relevante, especialmente diante das limitações cognitivas inerentes ao ser humano. Na prática operacional, a fadiga cognitiva em ambientes de videomonitoramento compromete a eficiência da tropa; após longos períodos de vigilância, a acuidade visual do operador diminui consideravelmente (WARM; PARASURAMAN; MATTHEWS, 2008), elevando o risco de omissões na identificação de ocorrências críticas. Nas abordagens de rua, poucos segundos de atraso na detecção de uma arma branca em uma situação de risco podem resultar em consequências letais, comprometendo a resposta tática da guarnição.

Nesse cenário operacional, a Inteligência Artificial (IA) não se apresenta como substituta do policial militar, mas como ferramenta de apoio à atuação humana, operando sob supervisão em tempo real (AMERSHI et al., 2019). No entanto, devido à vasta extensão territorial e aos desafios de acesso no interior paraense, a utilização de soluções de processamento remoto em nuvem mostra-se limitada na prática (SHI et al., 2016), ao passo que a substituição massiva de câmeras convencionais por modelos embarcados inteligentes representa custo elevado para a corporação (NVIDIA, 2022).

Fugindo dessa dependência, destaca-se a *Edge Computing* (computação de borda), que utiliza servidores de borda (*Edge Servers*). Essa arquitetura consiste na implantação de servidores locais robustos, responsáveis pelo processamento das imagens das câmeras dentro da infraestrutura da PMPA, o que elimina a necessidade de envio contínuo de dados à nuvem, garante a continuidade da análise de vídeo mesmo diante de instabilidades na conexão centralizada e permite o aproveitamento do ecossistema de câmeras IP já existente. Diante desse contexto da urgência das ruas, do desgaste cognitivo nas salas de controle e das limitações da nuvem, formula-se o problema central desta pesquisa: como implantar uma arquitetura de visão computacional em servidores de borda para antecipar a detecção de armas brancas e mitigar a fadiga dos operadores da PMPA, garantindo que a automação não comprometa a confiabilidade das ocorrências frente aos falsos alarmes e à detecção de objetos distantes.

A partir desse contexto, o objetivo geral deste trabalho é desenvolver, avaliar e demonstrar a viabilidade de implantação de uma arquitetura de visão computacional em servidores de borda para antecipar a detecção de armas brancas e mitigar a fadiga dos operadores da PMPA, garantindo a confiabilidade das ocorrências perante falsos alarmes e limitações de detecção à distância. Para tanto, os objetivos específicos são: (I) desenvolver e treinar um modelo de detecção baseado em YOLO26l voltado a cenários operacionais locais; (II) avaliar o desempenho e a confiabilidade do modelo em testes de validação, com foco na precisão, velocidade e na confiabilidade frente a falsos alarmes e à detecção de objetos distantes; e (III) demonstrar a viabilidade da arquitetura de Edge Server como alternativa de baixo custo e alta disponibilidade para a corporação.

2 FUNDAMENTAÇÃO TEÓRICA

A busca por soluções tecnológicas avançadas na segurança pública paraense não é apenas uma inovação técnica, mas um direcionamento institucional claro. O Plano Estadual de Segurança Pública e Defesa Social do Pará (2022-2031), por meio do seu Programa de Modernização Institucional, estabelece como um dos seus elementos constitutivos o aperfeiçoamento das ações de inteligência (PARÁ, 2022). Para materializar essa diretriz no campo operacional, o plano traça ações estratégicas específicas, como a atualização do parque tecnológico de inteligência da SEGUP e a composição do Sistema de Inteligência do SIEDS

pela Polícia Militar. É justamente para atender a essa demanda de modernização e integração institucional que a adoção de novas arquiteturas computacionais se faz necessária, deslocando o poder de processamento e análise para mais perto do cenário de risco.

Na rotina do NIOp (Núcleo Integrado de Operações), responsável por mediar a comunicação entre o cidadão e os órgãos de Segurança Pública do Pará, a distinção entre processar dados na nuvem e fazê-lo na borda da rede, via *Edge Computing* (computação de borda) e *Edge Servers*, vai além da arquitetura computacional; trata-se de uma questão de sobrevivência operacional. Processar a informação ali mesmo, na origem, corta o atraso e viabiliza a resposta imediata (SHI et al., 2016). No cenário operacional da PMPA, isso se traduz em uma vantagem tática clara, atuando como um multiplicador de força, conceito doutrinário que permite ampliar a capacidade de vigilância sem acréscimo de efetivo. O alerta de uma ameaça chega ao comandante da guarnição em frações de segundo, e não em segundos. Em um confronto real, esse lapso temporal pode mudar o resultado da ocorrência: é a diferença entre neutralizar a ameaça a tempo ou sofrer baixas.

Para atender a essa demanda de agilidade, as redes neurais do tipo YOLO (*You Only Look Once*) tornaram-se referência no setor para detecção em tempo real. Diferente das arquiteturas antigas, que varrem a imagem várias vezes, elas localizam e classificam o objeto numa passada só, garantindo agilidade sem perder precisão (REDMON; FARHADI, 2018). A versão empregada neste estudo, a YOLO26l (ULTRALYTICS, 2026), traz refinamentos específicos para a extração de características em objetos de pequenas dimensões, uma melhoria crucial para o cenário policial, onde lâminas e armas brancas frequentemente ocupam apenas uma fração mínima do quadro visual, tornando sua identificação um desafio técnico complexo.

O problema é que a literatura técnica nem sempre considera o fator humano ou as variáveis imprevisíveis do terreno. Um modelo com alta precisão estatística em laboratório pode falhar nas ruas se não considerar as nuances do ambiente real, como a iluminação precária de um beco, a chuva torrencial do interior paraense ou o ângulo oblíquo de uma câmera mal posicionada. Por isso, o conceito de *Human-in-the-Loop* (Humano no Ciclo) torna-se um imperativo doutrinário (CAO et al., 2019; HOLZINGER, 2016): o sistema proposto atua apenas como um sensor avançado, sugerindo a ameaça, mas a decisão final permanece com o operador na sala de controle. Essa interação não apenas mitiga o risco de

alarmes falsos, uma caneta refletindo a luz do sol ou uma chave de fenda no cinto de um trabalhador, mas também alimenta o ciclo de aprendizado contínuo, permitindo o retreinamento do modelo para as especificidades de cada bairro ou até do turno em que a guarnição atua.

3 METODOLOGIA

A pesquisa aqui desenvolvida seguiu uma abordagem de natureza aplicada e com viés experimental, estruturada como estudo de caso (YIN, 2018) sobre a implantação de um sistema de visão computacional voltado à detecção de armas brancas em tempo real no contexto operacional da PMPA. O sistema opera via servidor de borda (*Edge Server*), uma estação de trabalho local com GPU dedicada, instalada fisicamente no NIOp da PMPA, processando os *feeds* de câmeras IP sem depender de infraestrutura de nuvem. Essa arquitetura garante baixa latência (a inferência ocorre em milissegundos, não em segundos de ida e volta para *datacenters* remotos), soberania dos dados (imagens sensíveis não saem do perímetro físico da corporação) e resiliência (o sistema continua operando mesmo com falha total de conectividade externa) (SHI et al., 2016).

No desenvolvimento do sistema, empregou-se a linguagem Python integrada a um *stack* tecnológico voltado à eficiência local: PySide6 para as interfaces de interação com o operador; OpenCV para a captura e tratamento de vídeo; o ecossistema PyTorch/Ultralytics para o treinamento e inferência da rede neural; e SQLite para a persistência e auditoria dos eventos. O trabalho foi estruturado em seis etapas, da sanitização dos dados brutos à validação operacional, detalhadas a seguir.

3.1 Etapa 1 – Sanitização e Deduplicação do Dataset

Na rotina de videomonitoramento, a extração contínua de quadros gera volumes massivos de imagens quase idênticas. Se não filtradas, essas redundâncias induzem o modelo ao *overfitting*, levando-o à memorização de cenários específicos em detrimento da generalização (GOODFELLOW et al., 2016). O dataset inicial, composto por imagens de bases públicas e cenas internas de videomonitoramento, foi submetido a esse processo de limpeza, resultando em um total de 2.356 imagens válidas. Para eliminar duplicatas sem

descartar variações visuais úteis, adotou-se uma abordagem híbrida em dupla camada: *Perceptual Hash* (pHash) para agrupamento inicial por similaridade estrutural (ZAUNER; RIPPERGER; INNERHOFER-OBERPERFLER, 2011), seguida do erro quadrático médio (MSE) em escala de cinza (64×64 pixels) como verificação de segurança contra falsos positivos. Nos limiares mais restritos, a confirmação de duplicata exige MSE < 5,0 para *hashes* idênticos (distância 0) e MSE < 30,0 para distância ≤ 2. A interface permite ao operador ajustar esse limiar em três zonas operacionais, Rigorosa (0–2, com validação pixel a pixel via MSE), Equilibrada (3–9, apenas pHash) e Ampla (10–15), conforme o perfil do *dataset*. Esse protocolo mitiga o *data leakage*, situação em que cópias da mesma fotografia migram simultaneamente para os conjuntos de treino e validação, comprometendo a capacidade preditiva do modelo em dados inéditos (CAWLEY; TALBOT, 2010).

3.2 Etapa 2 – Rotulagem Semiautomática e Balanceamento de Classes

A rotulagem foi concebida de forma semiautomática para contornar a fadiga visual inerente ao monitoramento contínuo e ao traço manual prolongado (WARM; PARASURAMAN; MATTHEWS, 2008). O *canvas* interativo conta com *Auto-Label* Contextual: um modelo YOLO pré-treinado sugere as caixas delimitadoras, mas apenas as da classe ativamente selecionada pelo operador são incorporadas. Essa dinâmica alinha-se aos princípios de aprendizado de máquina interativo, nos quais a intervenção humana refina e valida as sugestões do algoritmo de forma eficiente (HOLZINGER, 2016). Antes do *split*, exige-se a utilização de imagens de fundo (*backgrounds*), vital para cortar falsos alarmes via *Hard Negative Mining,* representando 28,4% do total do *dataset* final. O corte obedece à proporção 80/20 (treino/validação) com embaralhamento aleatório, e formatos exóticos (PNG, BMP, WEBP, AVIF) são convertidos automaticamente para JPG, padronizando o input e reduzindo a latência de leitura em disco.

3.3 Etapa 3 – Treinamento e Otimização do Modelo

Utilizou-se o YOLO26l (ULTRALYTICS, 2026) com *transfer learning*, congelando as 10 primeiras camadas do *backbone*. Essa escolha mantém o paradigma de detecção em passada única (*single pass*), que revolucionou a identificação de objetos em tempo real ao unificar as tarefas de localização e classificação em uma única rede neural (REDMON;

FARHADI, 2018). Os hiperparâmetros, detalhados na Tabela 1, visaram mitigar o desbalanceamento, com destaque para *Box Loss* (5,0) e *Copy-Paste* (0,3). O treinamento ocorreu em ambiente de nuvem (Kaggle, multi-GPU), opção viável pois a fase de treino não exige baixa latência, diferentemente da inferência operacional, com resolução de entrada fixada em 768×768 pixels. A validação e a inferência em tempo real foram executadas localmente no Edge Server (GPU única), garantindo que as métricas de desempenho refletissem as condições operacionais da PMPA. Concomitantemente, o script executa uma auditoria de viés interclasses: se o mAP50 da classe *knife* superar o de *scissors* em mais de 0,15, o modelo é rejeitado por falta de representatividade da classe minoritária.

Tabela 1 – Hiperparâmetros de treinamento do modelo YOLO26l

Hiperparâmetro	Valor	Justificativa Operacional
Otimizador	AdamW	Melhor generalização em datasets médios
Learning Rate inicial	0,0002	Taxa conservadora para evitar esquecimento catastrófico
Weight Decay	0,05	Regularização contra *overfitting*
Box Loss	5,0	Priorização da localização espacial
Cls Loss	1,2	Foco na classificação correta
DFL	1,0	Precisão nas bordas das caixas delimitadoras
Copy-Paste	0,3	Aumento sintético de instâncias raras

Mosaic	0,5	Variedade de contextos de fundo
Close Mosaic	20 épocas	Estabilização no final do treinamento

Fonte: Dados da pesquisa (2026).

3.4 Etapa 4 – Validação Estatística e Promoção para Produção

A validação é uma auditoria comparativa *hold-out* com métricas desagregadas por classe (Precision, Recall, mAP50, mAP50-95 e FPS), essencial para atestar a capacidade de generalização do modelo diante de dados inéditos (GOODFELLOW et al., 2016). Na prática, o sistema classifica o desempenho em faixas operacionais: mAP50 acima de 0,90 é considerado excelente para implantação; entre 0,80 e 0,89, bom; de 0,70 a 0,79, apenas aceitável (exigindo cautela); e abaixo de 0,70, insuficiente, o que bloqueia automaticamente o modelo. O protocolo de promoção compara essa nova versão contra o *baseline* vigente em produção. A substituição só ocorre se houver ganho de mAP50 > 0,01 e nenhuma queda de Recall superior a 0,05. Essa salvaguarda evita que melhorias marginais na precisão sejam conquistadas à custa da sensibilidade, cenário inaceitável em segurança pública, onde um falso negativo significa a não detecção de uma arma branca. Estabeleceu-se como patamar mínimo operacional para aprovação: mAP50 ≥ 0,75, Recall ≥ 0,80 e *throughput* ≥ 15 FPS no servidor de borda.

3.5 Etapa 5 – Inferência em Tempo Real e Captura de Evidências

A inferência usa arquitetura *multithread* otimizada para o processamento na borda da rede, garantindo que a análise ocorra próxima à fonte geradora dos dados, eliminando a latência de transmissão para a nuvem (SHI et al., 2016). Para assegurar a resiliência do sistema, implementou-se um mecanismo de *fallback* automático: em caso de falha crítica do driver da GPU, o processamento é migrado dinamicamente para a CPU. O sistema opera com aquisição contínua, processamento YOLO em buffer circular (filtrado por janela temporal de 3s) e gravação assíncrona. Opera-se com tamanho de inferência de 480px para maximizar

FPS, limiar de confiança baixo (0,25) para minimizar falsos negativos, delegando ao operador a filtragem de falsos positivos, e *cooldown* de 6s entre alertas, evitando saturação da interface. Ao detectar uma ameaça, o mecanismo de *Slots* (janelas de memória buffer) captura os 3 segundos anteriores e posteriores ao disparo, gerando uma janela de 6s de contexto temporal essencial para a cadeia de custódia.

3.6 Etapa 6 – Validação Operacional, Feedback Humano e Rastreabilidade

Em segurança pública, a decisão final não cabe à máquina. Estudos sobre detecção de itens proibidos demonstram que a integração do operador é indispensável para mitigar o risco de alarmes indevidos em cenários de alta complexidade visual (CAO et al., 2019). O conceito de *Human-in-the-Loop* (HITL) materializa-se em Cartões Flutuantes de Alerta, onde o operador confirma a ameaça ou a descarta como falso alarme, seguindo diretrizes de interação em que o humano mantém o controle da ação decisória (AMERSHI et al., 2019). O sistema sugere; o policial decide. Alertas simultâneos vão para uma Fila de Triagem em lote. Toda evidência é indexada em banco relacional local (SQLite) por *timestamp*, status e confiança. O processamento ocorre *on-premise*, sem nuvem, em conformidade com a Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018) e diretrizes internas da PMPA, restringindo o acesso ao efetivo de operação e auditoria.

4 RESULTADOS E DISCUSSÃO

Os testes de validação, executados sobre o conjunto de *hold-out*, atestaram que a arquitetura proposta possui maturidade técnica para implementação no ambiente operacional da PMPA. A apresentação dos dados divide-se entre as métricas estatísticas brutas e a interpretação de seu impacto na rotina de policiamento.

4.1 Desempenho Estatístico e Correção de Viés

O modelo treinado alcançou um mAP50 (mean Average Precision, métrica que calcula a média das precisões médias para cada classe de objeto detectado, considerando a sobreposição mínima de 50% entre a caixa prevista e a real) de 97,19%, índice excelente para

ambientes não controlados e de baixa luminosidade. As métricas gerais de desempenho foram: Precision de 95,21%, Recall de 95,63% e *throughput* de 28,9 FPS no *Edge Server*.

Analisando o *dataset*, chama atenção o desbalanceamento inerente: a classe 'knife' representava 84,9% das instâncias, enquanto 'scissors' somava apenas 15,1%. Em arquiteturas tradicionais, essa disparidade induz o modelo ao enviesamento, fazendo com que a rede simplesmente ignore a classe minoritária (GOODFELLOW et al., 2016). Porém, as estratégias de aumento sintético de dados (*Copy-Paste* e *Mosaic*) e a calibração da função de perda (*Box Loss* e *Cls Loss*) aplicadas no treinamento forçaram a rede a extrair características morfológicas do tipo 'lâmina cortante', generalizando o aprendizado. O resultado foi a manutenção de um mAP50 de 97,3% isoladamente para a classe minoritária (tesouras), diferença que cumpriu o critério de auditoria interclasses (abaixo de 0,15 de distância), comprovando que o viés estatístico foi mitigado com êxito.

4.2 Viabilidade Operacional, Limitações Táticas e Resiliência de Hardware

Do ponto de vista tático, as métricas brutas ganham outro contorno. O índice de *Precision* de 95,21% garante que, a cada 100 alertas emitidos, 95 apontam para ameaças reais. Na prática, isso evita a desconfiança do operador, um problema crítico documentado em sistemas que geram excessivos falsos alarmes, acelerando a fadiga cognitiva em vez de mitigála (WARM; PARASURAMAN; MATTHEWS, 2008). Já o *Recall* de 95,63% garante que o sistema identifica a vasta maioria das armas que cruzam o campo de visão, e os 28,9 FPS asseguram que o vídeo processado não sofre degradação visual perceptível ao olho humano.

Apesar dos índices expressivos, é fundamental reconhecer as limitações inerentes ao cenário real. A precisão de 95,21%, embora elevada, ainda produz falsos positivos em cenários de complexidade visual, como reflexos de luz em objetos metálicos cotidianos ou ferramentas de trabalho. Ademais, a eficácia do modelo cai consideravelmente quando a arma branca está muito distante da câmera, ocupando poucos pixels na imagem, ou em casos de oclusão severa. É justamente nessas falhas operacionais que a filosofia do Human-in-the-Loop se justifica plenamente: o sistema atua como um sensor de triagem que sinaliza a suspeita,

mas a prerrogativa de confirmar a ameaça em condições ambíguas deve permanecer no julgamento do operador, evitando ações letais baseadas em inferências incertas.

Na prática, o grande diferencial da arquitetura está na lógica de resiliência e interação. Em testes simulados de falha crítica de *hardware* (desligamento forçado do driver da GPU), o sistema migrou automaticamente o processamento para a CPU em menos de 6 segundos, mantendo o monitoramento ativo, embora com queda proporcional de FPS. Essa capacidade de tolerância a falhas valida a premissa de que o processamento em borda elimina a dependência de infraestrutura centralizada como ponto único de falha, entregando resiliência onde a computação em nuvem falharia por dependência de enlace (SHI et al., 2016).

Por sua vez, o processo de validação manual, feita pelo operador, mostrou-se o elo entre a estatística e a rotina operacional. Ao clicar em "Confirmar" ou "Falso Alarme", o policial não apenas valida a ocorrência imediata, mas também alimenta o banco de dados local. Esse fluxo viabiliza o conceito de aprendizado interativo (HOLZINGER, 2016). Embora a implantação atual se concentre no NIOp das principais cidades, a arquitetura é escalável: à medida que o sistema for instalado em BPMs (Batalhões), CIPMs (Companhias Independentes da PM), pelotões ou PPDs (Postos Policiais Destacados), ele poderá ser retreinado periodicamente com as nuances e o contexto operacional específicos de cada unidade.

5 CONCLUSÃO

O desenvolvimento e a validação do sistema demonstram que a computação de borda constitui o caminho viável e estruturante para a modernização da vigilância na PMPA. Diante do expressivo número de apreensões de armas brancas no estado, 7.679 unidades registradas apenas em 2024 pelo Departamento-Geral de Operações (DGO) (PARÁ, 2024), ferramentas que antecipem a identificação dessas ameaças deixam de ser um diferencial tecnológico para se tornarem uma necessidade operacional, uma vez que o Art. 19 da Lei de Contravenções Penais
(BRASIL, 1941) subordina a tipificação do porte de arma branca à conjugação de dois requisitos: a periculosidade do objeto, avaliada à luz da situação fática específica, e a

finalidade agressiva de quem o conduz. A detecção prévia pelo sistema viabiliza ao operador do NIOp a constatação dessa conduta com maior segurança e menor risco operacional.

A implantação se mostrou viável mediante o uso de Edge Servers com fallback para CPU, aliado ao filtro humano de falsos positivos, respondendo diretamente ao problema central desta pesquisa. A pesquisa atingiu os objetivos propostos. Quanto ao desenvolvimento e treinamento do modelo YOLO26l (Objetivo I), a arquitetura provou ser altamente adaptável ao domínio policial, atingindo mAP50 de 97,19% e mitigando o viés histórico entre classes de armas, garantindo detecção confiável mesmo de artefatos de ocorrência estatística menor. Na avaliação de desempenho e confiabilidade (Objetivo II), os 28,9 FPS aliados a uma Precisão de 95,21% atestam que a ferramenta opera em tempo real, sendo a supervisão humana o mecanismo indispensável para filtrar alarmes espúrios e validar as ocorrências frente às limitações de detecção de objetos distantes. Por fim, sobre a viabilidade da arquitetura de Edge Server (Objetivo III), o sistema comprovou resiliência mandatória para missões de segurança: a migração automática para CPU em caso de falha de GPU assegura que o policiamento não fica desassistido, e o armazenamento local preserva a soberania das evidências.

Ao centralizar o processamento em um servidor local, a ferramenta atua como um verdadeiro multiplicador de força: uma única guarnição de plantão consegue monitorar dezenas de câmeras com a eficácia de uma equipe dedicada. O sistema não substitui o policial; atua como um sensor avançado sob a égide do *Human-in-the-Loop* (CAO et al., 2019; AMERSHI et al., 2019), respeitando a doutrina militar que exige a decisão humana para o emprego de força e garantindo a evolução contínua do algoritmo.

Para estudos futuros, recomenda-se a expansão do *dataset* para abarcar outras classes de armas brancas recorrentes em crimes no interior do estado, o desenvolvimento de módulos de alta resolução para mitigar a perda de desempenho em objetos muito distantes, e a integração direta do sistema com o NIOp, viabilizando o envio automático das evidências em vídeo no exato instante em que a ameaça é confirmada pelo operador.

REFERÊNCIAS

AMERSHI, S. et al. Guidelines for Human-AI Interaction. In: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2019, Glasgow. **Proceedings** [...]. New York: ACM, 2019. p. 1-13. DOI: 10.1145/3290605.3300233. Disponível em: https://dl.acm.org/doi/10.1145/3290605.3300233. Acesso em: 22 abr. 2026.

BRASIL. Decreto-Lei nº 3.688, de 3 de outubro de 1941. Lei das Contravenções Penais. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 6 out. 1941. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del3688.htm. Acesso em: 22 abr. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 20 abr. 2026.

CAO, S. et al. Toward Human-In-The-Loop Prohibited Item Detection in X-Ray Baggage Images. In: CHINESE AUTOMATION CONGRESS (CAC), 2019, [S.l.]. **Proceedings** [...]. [S.l.]: IEEE, 2019. p. 4360-4364. Disponível em: https://ieeexplore.ieee.org/document/8996933. Acesso em: 24 abr. 2026.

CAWLEY, G. C.; TALBOT, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. **Journal of Machine Learning Research**, [S.l.], v. 11, p. 2079-2107, 2010.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

HOLZINGER, A. Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop? **Brain Informatics**, [S.l.], v. 3, n. 2, p. 119-131, 2016. Disponível em: https://link.springer.com/article/10.1007/s40708-016-0042-6. Acesso em: 20 abr. 2026.

NVIDIA. What's the Difference: Edge Computing vs. Cloud Computing. **NVIDIA Blog**, [S.l.], 5 jan. 2022. Disponível em: https://blogs.nvidia.com/blog/difference-between-cloud-and-edge-computing/. Acesso em: 20 abr. 2026.

PARÁ. Polícia Militar do Pará. Departamento-Geral de Operações. **Relatório de produtividade: ano 2024 (01 jan. a 31 de dez.)**. Belém: PMPA, 2024. Disponível em: https://www.pm.pa.gov.br/. Acesso em: 16 abr. 2026.

PARÁ. Secretaria de Estado de Segurança Pública e Defesa Social. **Plano Estadual de Segurança Pública e Defesa Social 2022-2031**. Belém: SEGUP, 2022. 131 f.: il. color. Disponível em:

http://sistemas.segup.pa.gov.br/transparencia/wpcontent/uploads/2023/03/Plano-Estadual_compressed.pdf. Acesso em: 1 maio 2026.

REDMON, J.; FARHADI, A. YOLOv3: An Incremental Improvement. **arXiv preprint**, [S.l.], 2018. arXiv:1804.02767. Disponível em: https://arxiv.org/abs/1804.02767. Acesso em: 23 abr. 2026.

SHI, W. et al. Edge Computing: Vision and Challenges. **IEEE Internet of Things Journal**, [S.l.], v. 3, n. 5, p. 637-646, out. 2016. DOI: 10.1109/JIOT.2016.2579198. Disponível em: https://ieeexplore.ieee.org/document/7488250. Acesso em: 22 abr. 2026.

ULTRALYTICS. **YOLO26**. [S.l.]: Ultralytics, 2026. Disponível em: https://docs.ultralytics.com/pt/models/yolo26/. Acesso em: 15 abr. 2026.

WARM, J. S.; PARASURAMAN, R.; MATTHEWS, G. Vigilance requires hard mental work and is stressful. **Human Factors**, [S.l.], v. 50, n. 3, p. 433-441, 2008. DOI: 10.1518/001872008X312152. Disponível em: https://journals.sagepub.com/doi/10.1518/001872008X312152. Acesso em: 18 abr. 2026.

YIN, R. K. **Case study research and applications: design and methods**. 6. ed. Los Angeles: SAGE Publications, 2018.

ZAUNER, C.; RIPPERGER, T.; INNERHOFER-OBERPERFLER, F. A survey of perceptual hashing for multimedia. In: INTERNATIONAL JOINT CONFERENCE ON E-BUSINESS AND TELECOMMUNICATION NETWORKS (ICETE), 8., 2011, Seixal. **Proceedings** [...]. Berlin: Springer, 2011. p. 322-337.