UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

HÉLIO CHAVES PEIXOTO NETO

AVALIAÇÃO DO CUSTO-BENEFÍCIO DA APLICAÇÃO DE MÉTODOS DE APRENDIZADO SEMI SUPERVISIONADO PARA DETECÇÃO DE EQUIPAMENTOS DE PROTEÇÃO INDIVIDUAL EM AMBIENTE INDUSTRIAL

Recife
2025

HÉLIO CHAVES PEIXOTO NETO

AVALIAÇÃO DO CUSTO-BENEFÍCIO DA APLICAÇÃO DE MÉTODOS DE APRENDIZADO SEMI SUPERVISIONADO PARA DETECÇÃO DE EQUIPAMENTOS DE PROTEÇÃO INDIVIDUAL EM AMBIENTE INDUSTRIAL

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional

Orientador (a): Teresa Bernarda Ludermir

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Peixoto Neto, Hélio Chaves.
Avaliação do custo-benefício da aplicação de métodos de aprendizado semi supervisionado para detecção de Equipamentos de Proteção Individual em Ambiente Industrial / Hélio Chaves Peixoto Neto. - Recife, 2025.
70f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2025.
Orientação: Teresa Bernarda Ludermir.
Inclui referências.

1. Aprendizado semi supervisionado; 2. Equipamentos de proteção individual; 3. Visão computacional; 4. Detecção de objetos; 5. Segurança ocupacional. I. Ludermir, Teresa Bernarda. II. Título.

UFPE-Biblioteca Central

Hélio Chaves Peixoto Neto

“Avaliação do Custo-Benefício da Aplicação de Métodos de Aprendizado Semi Supervisionado para Detecção de Equipamentos de Proteção Individual em Ambiente Industrial”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 30/07/2025.

BANCA EXAMINADORA

_____

Profa. Dra. Teresa Bernarda Ludermir
Centro de Informática / UFPE
(**orientadora**)

_____

Prof. Dr. Sergio Fernandovitch Chevtchenko
MARCS Institute /University of Western Sydney

_____

Prof. Dr. Tarcísio Daniel Pontes Lucas
Universidade de Pernambuco / Campus Caruaru

AGRADECIMENTOS

À minha esposa, Luisa Lacerda Rique, a José e Carlos Murilo, por todo o apoio, amor e companheirismo durante essa jornada.

Aos meus pais, Geane Luciana de Freitas e Erasmo Chaves Peixoto Sobrinho, por todo o esforço, apoio, conquistas celebradas e amor que me deram antes e durante a caminhada. Sem eles o presente trabalho sequer existiria.

Aos meus irmãos, em especial a Gustavo Chaves Peixoto, pelo amor e parceria nessa jornada.

À minha família, pelo amor, cuidado e ajuda que culminaram neste trabalho.

Aos meus amigos, em especial aos membros do “Só Brincadeiras+”, “PCF Farras 4.0”, “Squad”, “Squad-“, “Amizade + q vddeira" “After nunca mais” e “Squad+" pelas conversas, encontros e companheirismo de sempre.

À Edward, Laís, Vitória, Isabella e Victor, pelo apoio no presente trabalho.

À indústria paulista pela parceria e o uso dos dados.

À professora Teresa, pelos ensinamentos e oportunidades durante toda a jornada. Sem ela o trabalho não aconteceria.

À Universidade Federal de Pernambuco e ao Centro de Informática, pela oportunidade e infraestrutura concedida para a realização do presente trabalho.

RESUMO

A detecção automatizada de Equipamentos de Proteção Individual (EPIs) em ambientes industriais representa um desafio significativo devido aos altos custos e tempo demandado para anotação manual de dados. Este trabalho investigou a eficácia de métodos de aprendizado semi supervisionado como alternativa ao aprendizado totalmente supervisionado para detecção de EPIs, visando responder qual abordagem apresenta melhor relação custo-benefício. A metodologia experimental foi conduzida em ambiente industrial real, utilizando 50.088 imagens capturadas em uma indústria para detecção de quatro classes: capacete, colete, pessoa e abafador tipo concha. Foram implementadas estratégias de anotação parcial com 10%, 20% e 30% dos dados, aplicando técnicas de *pseudo-labeling* através da arquitetura YOLOv8. Os resultados foram avaliados mediante validação cruzada com k=10 repetições e análise estatística usando testes de Friedman e Nemenyi. Os modelos semi supervisionados demonstraram performance comparável ao totalmente supervisionado, com diferenças controladas nas métricas principais: mAP@0.5 de 0.971 (10%), 0.979 (20%) e 0,985 (30%) contra 0.986 (100%) e mAP@0.5:0.95 de 0.767 (10%), 0.771 (20%) e 0,801 (30%) contra 0.805 (100%). A abordagem semi supervisionada resultou em economia substancial de tempo de anotação, reduzindo em 85% o processo manual. Os resultados indicam que métodos semi supervisionados constituem alternativa viável e economicamente vantajosa para desenvolvimento de sistemas de detecção de EPIs, mantendo eficácia técnica com significativa redução de recursos humanos especializados.

Palavras-chave: Aprendizado semi supervisionado; Equipamentos de proteção individual; Visão computacional; Detecção de objetos; Segurança ocupacional.

ABSTRACT

Automated detection of Personal Protective Equipment (PPE) in industrial environments represents a significant challenge due to high costs and time required for manual data annotation. This work investigated the effectiveness of semi-supervised learning methods as an alternative to fully supervised learning for PPE detection, aiming to answer which approach presents the best cost-benefit ratio. The experimental methodology was conducted in a real industrial environment, using 50,088 images captured in a São Paulo industry for detecting four classes: helmet, vest, person, and ear protection. Partial annotation strategies were implemented with 10%, 20% and 30% of data, applying pseudo-labeling techniques through YOLOv8 architecture. Results were evaluated through cross-validation with k=10 repetitions and rigorous statistical analysis using Friedman and Nemenyi tests. Semi-supervised models demonstrated comparable performance to fully supervised approach, with controlled differences in main metrics: mAP@0.5 of 0.971 (10%), 0.979 (20%) and 0,985 (30%) versus 0.986 (100%) and mAP@0.5:0.95 of 0.767 (10%), 0.771 (20%) and 0,801 (30%) versus 0.805 (100%). The semi-supervised approach resulted in substantial annotation time savings, reducing the manual process by 85%. Results indicate that semi-supervised methods constitute a viable and economically advantageous alternative for developing PPE detection systems, maintaining technical efficacy with significant reduction of specialized human resources.

Keywords: Semi-supervised learning; Personal protective equipment; Computer vision; Object detection; Occupational safety.

LISTA DE ILUSTRAÇÕES

Figura 1 - Subconjuntos da Inteligencia Artificial. 16
Figura 2 - Um diagrama simplificado de um neurônio humano. 1 - dendritos, 2 - núcleo do neurônio, 3 - zona de iniciação, 4 - axônio, e 5 - terminais do axônio. 18
Fluxograma 1 - Esquema de uma tarefa auxiliar contrastiva. 22
Fluxograma 2 - Esquema de uma tarefa auxiliar preditiva. 23
Fluxograma 3 - Esquema de uma tarefa auxiliar generativa. 23
Fluxograma 4 - Esquema genérico de um Aprendizado Autossupervisionado. 24
Figura 3 - Lista de métodos que podem ser empregados em problemas de classificação. 27
Figura 4 - Resultados da aplicação do framework STAC em comparação com o Aprendizado Supervisionado. 28
Gráfico 1 - Histograma do dataset. 34
Figura 5 - Mão de trabalhador se torna oclusão e impede os modelos de 10% (azul) e 20% (vermelho) de realizarem a detecção. 55
Figura 6 - Trabalhador se torna oclusão de outro e impede que dois dos modelos (10% e 20%) de realizarem a detecção. 56
Figura 7 - Trabalhador com o celular foi anotado (verde), mas não é detectado em modelo algum. 57
Figura 8 - Trabalhador não anotado (verde) é detectado nos modelos de 20% (rosa), 30% e 100% (vermelho). 58

LISTA DE TABELAS

Tabela 1 - Resultados da precisão média para os 5 modelos em cada classe. 20
Tabela 2 - Lista de aplicações recentes do estado da arte para Aprendizado Autossupervisionado. 25
Tabela 3 - Resultados da aplicação do Unbiased Teacher em comparação com outros para o dataset MS COCO. 29
Tabela 4 - Resultados da aplicação do framework CISO em comparação com outros para o dataset MS COCO. 30
Tabela 5 - Resultados da aplicação do framework CISO em comparação com outros para o dataset VOC. 30
Tabela 6 - Limiar de confiança ótimo por modelo e por classe. 37
Tabela 7 - Estatísticas descritivas do cross-validation. 44
Tabela 8 - Resultados do Teste de Friedman para as métricas. 45
Tabela 9 - Resultados do Teste de Nemenyi para cada comparação par a par. 46
Tabela 10 - Subtrações entre as médias das métricas após o cross-validation, em pontos percentuais. 47
Tabela 11 - Valores de mAP@0,5 por classes. 48
Tabela 12 - Valores de mAP@0,5:0,95 por classes. 49
Tabela 13 - Matriz de confusão normalizada para o modelo com 10% de dados anotados manualmente. 50
Tabela 14 - Matriz de confusão normalizada para o modelo com 20% de dados anotados manualmente. 51
Tabela 15 - Matriz de confusão normalizada para o modelo com 30% de dados anotados manualmente. 51
Tabela 16 - Matriz de confusão normalizada para o modelo com todos os dados anotados manualmente. 51
Tabela 17 - Estatísticas descritivas dos resultados no SH17. 61

SUMÁRIO

1 INTRODUÇÃO **12**
1.1 CONTEXTUALIZAÇÃO DO PROBLEMA E RELEVÂNCIA DO TEMA 12
1.2 OBJETIVOS 14
1.2.1 Objetivo Geral **15**
1.2.2 Objetivos Específicos **15**
1.3 ESTRUTURA DA DISSERTAÇÃO 15
2 REVISÃO BIBLIOGRÁFICA **16**
2.1 APRENDIZAGEM PROFUNDA 16
2.1.1 Origem da Aprendizagem Profunda **17**
2.1.2 Visão Computacional **18**
2.2 MÉTODOS DE APRENDIZAGEM PROFUNDA 19
2.2.1 Aprendizado Supervisionado **19**
2.2.2 Aprendizado Não Supervisionado **21**
2.2.3 Aprendizado Autossupervisionado **21**
2.2.3.1 Aplicações do Aprendizado Autossupervisionado 24
2.2.4 Aprendizado Semi Supervisionado **26**
2.2.4.1 Aplicações do Aprendizado Semi Supervisionado 28
3 METODOLOGIA **32**
3.1. COLETA DOS DADOS 32
3.1.1. Cenário industrial e vídeos capturados **32**
3.1.2. Estruturação dos frames e classes **32**
3.2. ESTRATÉGIAS DE ANOTAÇÃO PARCIAL 35
3.3. MODELO 1 - GERAÇÃO DE RÓTULOS 36
3.4. MODELO 2 - TREINAMENTO FINAL COM ROTULAÇÕES AUTOMÁTICAS 38
3.5. MÉTRICAS DE AVALIAÇÃO 39
3.6 VALIDAÇÃO ESTATÍSTICA 40
3.6.1 Protocolo de Validação Cruzada **40**
3.6.2 Teste de Friedman **41**
3.6.3 Teste de Nemenyi **41**
3.7. FERRAMENTAS, BIBLIOTECAS E AMBIENTE EXPERIMENTAL 42
4 RESULTADOS E DISCUSSÃO **44**

4.1. DESEMPENHO GERAL DOS MODELOS COM DIFERENTES PROPORÇÕES DE ANOTAÇÃO 44
4.1.1 Estatísticas Descritivas 44
4.1.2 Teste de Friedman 45
4.1.3 Teste de Nemenyi 45
4.1.4 Análise da Magnitude dos Efeitos e Discussão 46
4.2. ANÁLISE QUANTITATIVA DAS CLASSES 48
4.2.1 Análise dos mAPs 48
4.2.2 Matrizes de Confusão 50
4.2.3 Comparação com Abordagem Supervisionada 53
4.3. ANÁLISE VISUAL 54
4.4. DISCUSSÃO SOBRE A ECONOMIA DE TEMPO DE ANOTAÇÃO 58
4.5. VALIDAÇÃO EM DATASET PÚBLICO 59
4.5.1 Contexto do SH17 Dataset 60
4.5.2 Resultados 60
4.6. LIMITAÇÕES E CONSIDERAÇÕES METODOLÓGICAS 61
5 CONCLUSÕES E TRABALHOS FUTUROS 63
5.1. PRINCIPAIS CONCLUSÕES DA PESQUISA 63
5.2. IMPLICAÇÕES PRÁTICAS E ACADÊMICAS 63
5.3. SUGESTÕES DE TRABALHOS FUTUROS 64
5.3.1 Aplicação em outros ambientes industriais 64
5.3.2 Uso de técnicas de active learning 65
REFERÊNCIAS BIBLIOGRÁFICAS 66

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA E RELEVÂNCIA DO TEMA

A detecção ou a segmentação de objetos é uma fase importante da anotação de dados em aplicações de Visão Computacional (VC), como diagnósticos médicos, carros autônomos e robótica (AFLALO *et al.*, 2023).

Atualmente, a etapa de anotação de dados em um projeto de Visão Computacional é feita comumente por seres humanos. *Frame* a *frame*, as pessoas detectam quais classes de determinado produto pertence a uma amostra. Esse processo utiliza *bounding boxes* ou segmentação. O tempo investido na atividade é alto, comumente sendo a etapa que mais demora em um projeto de detecção de objetos (JING; TIAN, 2019). Além disso, há falta de padronização pelos diversos agentes que fazem a mesma atividade. Ao mesmo tempo, nem sempre especialistas no contexto do problema realizam essas anotações. Isso gera dúvidas sobre a qual classe pertence o item que será anotado. Por fim, por ser uma atividade repetitiva, a equipe de trabalho pode se desmotivar, o que diminui as chances de sucesso de um projeto comum.

As abordagens tradicionais de aprendizado supervisionado dependem de forma significativa da quantidade de dados anotados disponíveis para treinamento. Embora exista uma grande quantidade de dados acessíveis, a escassez de anotações tem impulsionado os pesquisadores a explorar abordagens alternativas que possam aproveitá-los de maneira eficaz. Nesse contexto, os métodos auto supervisionados e semi supervisionados desempenham um papel fundamental no avanço do aprendizado profundo, permitindo o aprendizado de representações de características sem a necessidade de anotações dispendiosas, aproveitando a supervisão implícita fornecida pelos próprios dados (JAISWAL *et al*., 2020). É por isso que esses métodos ganham cada vez mais atenção para diminuir os problemas citados (ZHANG *et al*., 2024).

A abordagem autossupervisionada é inspirada na forma que os bebês aprendem. Nos primeiros anos de vida, as crianças ganham conhecimento a partir da observação e a interação com o seu redor. O método autossupervisionado busca dividir as imagens não identificadas em tarefas pré-textuais, com o objetivo de

conceder um rótulo àquilo que ainda não é conhecido. Com o rótulo aplicado, as atividades comuns de anotação, como detecção e segmentação, são realizadas pelo próprio modelo (RANI *et al*., 2023)

As aplicações semi supervisionadas utilizam de poucos dados já anotados com seu rótulo de identificação para treinar um modelo que seja capaz de rotular os dados não anotados para aumentar a base de treinamento e refinar o modelo a fim de otimizar a performance final (XU; XIAO; LÓPEZ, 2019).

Apesar dos avanços e estudos desses métodos, poucas são as aplicações em contextos de segurança. Segundo o Ministério do Trabalho e Emprego do Brasil (2023), 499.955 acidentes de trabalho foram reportados com quase 3 mil óbitos associados. Além disso, em 2022, foram registrados 17,9 milhões de dias perdidos por auxílio-doença por acidente de trabalho e 8,4 milhões de dias perdidos por aposentadoria por invalidez por acidente de trabalho no Brasil. Nos 10 anos anteriores, o gasto do INSS com benefícios previdenciais acidentários ultrapassou os R$ 100 bilhões de reais, em dados atualizados do Observatório de Segurança e Saúde do Trabalho (2023), iniciativa coordenada pelo Ministério Público do Trabalho e pelo Escritório da Organização Internacional do Trabalho para o Brasil.

Uma das formas de prevenção, mais precisamente a última linha de defesa na hierarquia de controles de segurança ocupacional, são os Equipamentos de Proteção Individual (EPI). A legislação brasileira, através da Norma Regulamentadora NR-6, define EPI como "o dispositivo ou produto de uso individual utilizado pelo trabalhador, concebido e fabricado para oferecer proteção contra os riscos ocupacionais existentes no ambiente de trabalho" (BRASIL, 2022). Sua importância transcende aspectos meramente regulatórios, impactando diretamente a saúde, a vida dos trabalhadores e a viabilidade econômica das operações industriais.

Mesmo com a reconhecida importância, a fiscalização manual do uso de EPIs apresenta limitações práticas significativas. Supervisores de segurança não conseguem monitorar continuamente todos os trabalhadores em ambientes industriais extensos. Fatores como fadiga, distração e limitações de recursos humanos resultam em lacunas de monitoramento que podem ter consequências graves. Adicionalmente, a fiscalização humana pode ser percebida como invasiva ou punitiva, potencialmente gerando resistência cultural e comprometendo a efetividade

das políticas de segurança. Neste contexto, sistemas automatizados de detecção de EPIs baseados em visão computacional emergem como solução complementar promissora, permitindo monitoramento contínuo, objetivo e não invasivo da conformidade com protocolos de segurança. A implementação de sistemas automatizados de detecção de EPIs pode transformar a gestão de segurança ocupacional de reativa para proativa. Ao invés de identificar não-conformidades apenas após acidentes ou através de inspeções periódicas, estes sistemas permitem identificação em tempo real de situações de risco, possibilitando intervenções preventivas imediatas. Alertas automáticos podem ser enviados a trabalhadores e supervisores quando não-conformidades são detectadas, criando loops de *feedback* que reforçam comportamentos seguros. Dados agregados sobre padrões de conformidade podem informar decisões estratégicas sobre treinamento, design de processos e alocação de recursos de segurança. Além disso, registros automatizados de conformidade fornecem documentação objetiva para auditorias regulatórias e investigações de incidentes, reduzindo riscos legais e reputacionais para as organizações.

Estas lacunas são especialmente críticas considerando que o desenvolvimento de sistemas requer anotação especializada de grandes volumes de dados por profissionais com conhecimento técnico em segurança do trabalho, processo que é tanto custoso quanto demorado. A escassez de *datasets* públicos de EPIs específicos para contextos industriais agrava ainda mais este problema, uma vez que condições operacionais, tipos de equipamentos e características ambientais podem variar significativamente entre diferentes regiões e setores industriais.

1.2 OBJETIVOS

Com todo o contexto apresentado, o presente trabalho busca aplicar métodos de aprendizado semi supervisionado e supervisionado para avaliar a abordagem mais eficiente em rotulações de imagens na criação de um sistema de detecção de Equipamentos de Proteção Individual (EPI), responsáveis por mais de 56% dos acidentes em locais de construção (GALLO *et al*., 2022). Por isso, ao término desta dissertação pretende-se responder "Qual a abordagem mais custo-benéfica para o caso de uso apresentado?"

1.2.1 Objetivo Geral

Avaliar o método de aprendizado mais eficaz em rotulações de imagens de detecção de uso de Equipamentos de Proteção Individual em um contexto real.

1.2.2 Objetivos Específicos

- Discutir o impacto do uso das Aprendizagens Supervisionada, Autossupervisionada e Semi Supervisionada ;
- Comparar os resultados das abordagens Supervisionada e Semi Supervisionada em detecção de uso de Equipamentos de Proteção Individual (EPI) em um contexto real.

1.3 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está organizado em cinco capítulos que abordam de forma sistemática a investigação proposta. O primeiro capítulo apresenta a contextualização do problema, destacando os desafios da anotação manual de dados em projetos de Visão Computacional e a relevância da aplicação de métodos alternativos de aprendizado no contexto de segurança ocupacional, além de estabelecer os objetivos da pesquisa. O segundo capítulo desenvolve uma revisão bibliográfica abrangente sobre os fundamentos teóricos da aprendizagem profunda, métodos de aprendizado supervisionado, autossupervisionado e semi supervisionado, bem como suas aplicações em Visão Computacional e detecção de objetos, temas centrais do trabalho. O terceiro capítulo detalha a metodologia experimental empregada, incluindo os procedimentos de coleta de dados em ambiente industrial real, estratégias de anotação parcial, desenvolvimento dos modelos propostos e métricas de avaliação utilizadas. O quarto capítulo apresenta e discute os resultados obtidos, comparando o desempenho das diferentes abordagens de aprendizado, analisando as métricas quantitativas por classe e realizando análise visual dos padrões de detecção. Por fim, o quinto capítulo sintetiza as principais conclusões da pesquisa, discute as implicações práticas e acadêmicas dos achados e propõe direções para trabalhos futuros na área de detecção automatizada de equipamentos de proteção individual.

2 REVISÃO BIBLIOGRÁFICA

2.1 APRENDIZAGEM PROFUNDA

O recente interesse em tópicos como Inteligência Artificial e Aprendizagem de Máquina é grande. Assistentes como o ChatGPT da OpenAI, Bard do Google e CoPilot da Microsoft, tornam o tema mais popular e impactam mais pessoas no dia a dia. Porém, dentre os pesquisadores outro termo atrai ainda mais curiosidade, a Aprendizagem Profunda. Apesar de ser uma subaérea da Inteligência Artificial e da Aprendizagem de Máquina, como mostrado na Figura 1, o interesse nos métodos de aprendizado profundo se deve ao fato de que eles demonstraram superar as técnicas anteriores de última geração em várias tarefas, além da abundância de dados complexos de diferentes fontes como, por exemplo, visuais, auditivos, médicos, sociais e sensoriais (VOULODIMOS *et al*., 2018).

Figura 1 - Subconjuntos da Inteligencia Artificial.

Fonte: O autor (2025).

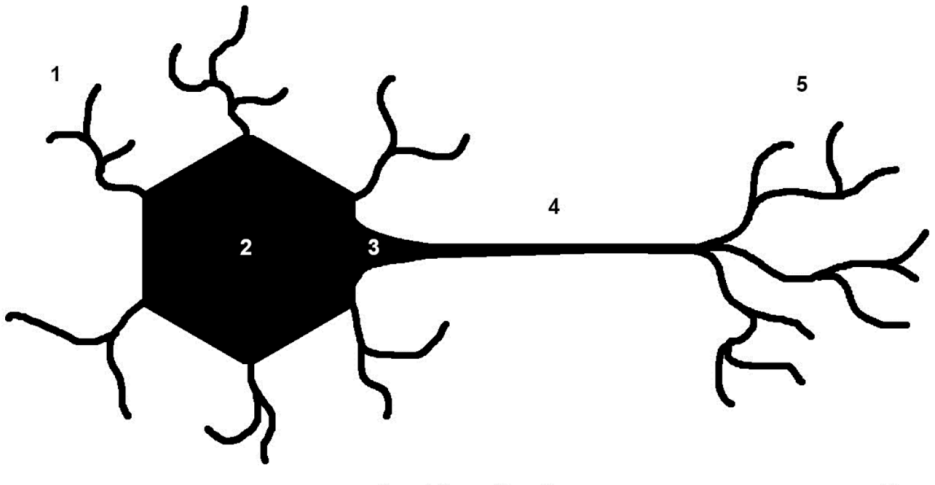
Ao mesmo tempo, o aumento do poder de processamento nas GPUs e CPUs, a diminuição do custo dos equipamentos e o avanço nos algoritmos trouxeram novos olhares ao tema.

Os primeiros sistemas precursores das técnicas de Aprendizagem Profunda (do inglês *Deep Learning*) surgiram entre as décadas de 1960 e 1970, porém a expressão "*Deep learning*" só foi criada em 2006 (SCHMIDHUBER, 2015). O nome provém da característica de entender informações padronizadas e cada vez mais específicas dos dados disponíveis com várias camadas especializadas, o que torna o aprendizado profundo.

2.1.1 Origem da Aprendizagem Profunda

Este tipo de aprendizado foi inspirado na forma como os humanos realizam tarefas complexas no dia a dia. O cérebro humano é um órgão complexo, que desempenha funções essenciais de processamento de informações em um intervalo temporal extremamente reduzido. As unidades fundamentais responsáveis por essas funções são os neurônios, mostrados na Figura 2, que facilitam a transmissão e o processamento de dados através de suas interações. Por exemplo, no corpo humano vários sensores transmitem informações sobre cheiros, imagens, temperatura, equilíbrio, dentre outros. Os neurônios, a partir dessas transmissões faz com que podemos responder a esses estímulos externos (DOUGHERTY, 2013). Além da resposta a estímulos, a habilidade do cérebro de identificar, como o reconhecimento um rosto familiar em meio a uma multidão, em poucos segundos, aumentou o desejo de criar artificialmente essas funções, dando origem às Redes Neurais Artificiais (RNA).

Figura 2 - Um diagrama simplificado de um neurônio humano. 1 - dendritos, 2 - núcleo do neurônio, 3 - zona de iniciação, 4 - axônio, e 5 - terminais do axônio.



Fonte: Kufel *et al*, 2023.

Nas Redes Neurais Artificiais, as informações de entrada são imagens, textos, vídeos, dentre outros. Estes dados atravessarão um conjunto de neurônios artificiais que aprenderão suas características específicas. Quando esses neurônios estão dispostos em múltiplas camadas, temos o que é chamado de Aprendizagem Profunda e as Redes Neurais (KUFEL *et al*., 2023).

2.1.2 Visão Computacional

Neste contexto de reprodução das funções humanas através das Redes Neurais Artificiais, a resolução de problemas multidimensionais com imagens e vídeos, seus processamentos e identificações é uma das áreas de interesse, chamada de Visão Computacional. A VC é uma tecnologia voltada para a automação e integração de diversos processos, por meio da extração de informações presentes em imagens ou vídeos, baseando-se nos princípios biológicos da visão. Sua fundamentação teórica remonta ao final da década de 1950, quando surgiu juntamente com o avanço da Inteligência Artificial (GROSSI, 2020). Uma pessoa observa qualquer objeto com seus olhos, mas o principal processo ocorre no cérebro. O cérebro é responsável por reconhecer, interpretar, compreender e

classificar o objeto por meio dos sinais recebidos dos olhos. Em seguida, ele gera e transmite informações sobre esse objeto. Ao mesmo tempo, o cérebro é capaz de reconhecer e classificar novos objetos ao compará-los com objetos já conhecidos e suas características, como ocorre quando uma pessoa vê um objeto pela primeira vez (KHANG *et al*., 2024). Então, a Visão Computacional utiliza algoritmos de reconhecimento de padrões para treinar máquinas com grandes quantidades de dados visuais. A máquina então processa as imagens de entrada, pode rotular os objetos nessas imagens e encontra padrões nesses objetos (CHATTERJEE, 2022).

2.2 MÉTODOS DE APRENDIZAGEM PROFUNDA

Os tipos de aprendizagem em um problema de Aprendizagem Profunda mais comuns são os supervisionados, não supervisionados e semi supervisionados. Será dado um foco maior nos dois últimos que farão parte do escopo metodológico deste trabalho.

2.2.1 Aprendizado Supervisionado

O aprendizado supervisionado pode ser definido como o processo de aprendizado de uma função que mapeia uma entrada para uma saída (LÓPEZ; LÓPEZ; CROSSA, 2022). Os dados de treinamento dessa função consiste em pares de objetos: o dado de entrada e a outra, o resultado desejado. A saída da função pode ser um valor numérico ou um rótulo de classe, caso este relacionado aos problemas de Visão Computacional. O objetivo final é aprender uma função que, dada uma amostra de dados e os resultados desejados, melhor aproxime a relação entre entrada e saída. Essa função deve ser capaz de prever o valor correspondente a qualquer entrada válida após ter visto uma série de exemplos dos dados de treinamento. Sob condições ideais, o algoritmo determina corretamente os rótulos de classe para instâncias desconhecidas, o que implica em um algoritmo de aprendizado capaz de generalizar para dados não vistos.

Em 2013, um problema compartilhado no site Kaggle envolveu uma competição de quem construía o algoritmo mais otimizado para identificar numa

imagem se o animal em questão era um gato ou cachorro. Apesar da resolução vencedora utilizar uma abordagem supervisionada, ou seja, rotular na base de dados que continham 25.000 imagens quais eram de gato ou cachorro, é razoável conjecturar que o tempo e o esforço aplicados podem ter sido grandes para resolver o problema.

Em 2022, Gallo *et al.* aplicaram cinco modelos de Aprendizado Supervisionado para detecção do uso de Equipamentos de Proteção Individual. Neste caso, foram escolhidos capacete, colete e luva para o experimento. No Aprendizado Supervisionado, todas as classes são rotuladas, inclusive as negativas. Só pode ser detectada a falta de capacete quando o modelo aprende o que é uma cabeça; quando não há luva, a mão é detectada; a mesma lógica é verdade para o colete e o busto. Portanto, são seis classes de interesse, das quais mais de 65 mil exemplos foram anotados manualmente. Os resultados são mostrados na Tabela 1 abaixo.

Tabela 1 - Resultados da precisão média para os 5 modelos em cada classe.

	Head	Helmet	Chest	Vest	Hand	Glove
YOLOv4	**96.4**	**98.2**	**94.9**	**86.7**	**80.4**	**63.6**
YOLOv4-Tiny	*92.6*	94.6	*91.8*	*79.8*	57.0	35.9
SSD MobileNet V2	77.3	86.2	81.3	69.9	43.1	21.3
CenterNet Resnet50 V2	*92.6*	*95.4*	91.1	75.2	*64.5*	*39.9*
EfficientDet D0	83.8	86.6	90.1	79.3	39.5	34.1

Fonte: Gallo *et al*, 2022.

O modelo YOLO (*You Only Look Once*) foi o melhor em todas as classes e representa atualmente o estado da arte em detecção de objetos. Após seu lançamento em 2015, o YOLO rapidamente se destacou como uma técnica inovadora, uma vez que, por meio de uma nova abordagem, foi capaz de alcançar uma precisão equivalente ou superior à dos métodos de detecção de objetos disponíveis na época, mas com uma velocidade de detecção significativamente superior (REDMON *et al*., 2016). Outro fator crucial para o sucesso do YOLO é o fato de ser totalmente de código aberto e livre de licenças de uso. Em outras palavras, tanto o código-fonte quanto a arquitetura da rede neural e os pesos pré-

treinados estão disponíveis para qualquer pessoa e podem ser utilizados de diversas formas, sem restrições.

Apesar do resultado citado, é importante notar que em casos com *datasets* maiores, a impossibilidade de rotulação ou o desconhecimento das classes alvo leva à segunda abordagem, o Aprendizado Não Supervisionado.

2.2.2 Aprendizado Não Supervisionado

O Aprendizado Não Supervisionado difere do aprendizado supervisionado pela utilização de dados não anotados, ou seja, dados que não foram previamente rotulados por seres humanos ou algoritmos (KUFEL *et al*., 2023). Nesse tipo de aprendizado, o modelo aprende a partir dos dados de entrada sem nenhum conhecimento prévio sobre as saídas rotuladas ou variáveis de resposta correspondentes. Em vez de rotular ou prever saídas, o algoritmo foca em agrupar ou associar os dados com base em suas características, na busca da identificação de padrões. Porém, há um caso específico que é possível classificar essas imagens mesmo sem dados rotulados: o Aprendizado Autossupervisionado.

2.2.3 Aprendizado Autossupervisionado

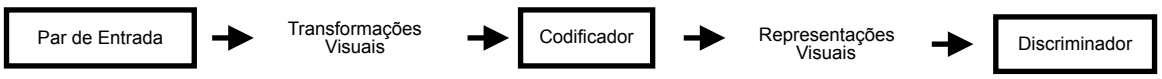
O Aprendizado Autossupervisionado foi definido pela primeira vez por Raina *et al*. em 2007. Os autores explicam essa abordagem afirmando que um algoritmo começa utilizando dados não rotulados para aprender uma representação concisa das entradas. Por exemplo, se os dados forem vetores de valores de intensidade de pixels que representam imagens, o algoritmo usará os dados de entrada para aprender os elementos básicos que compõem uma imagem. Esse processo pode envolver, por exemplo, a descoberta de fortes correlações entre as linhas de pixels, simplesmente ao examinar as estatísticas das imagens não rotuladas. Assim, o algoritmo pode perceber que a maioria das imagens contém várias bordas. Ao aprender essas correlações, o algoritmo passa a representar as imagens não mais com base nos valores brutos de intensidade dos pixels, mas em termos das bordas que aparecem nelas. Essa representação da imagem, agora centrada nas bordas,

em vez dos valores de pixels brutos, constitui uma forma mais abstrata ou de nível superior de representar a entrada. Essa abordagem permite que o algoritmo aprenda de maneira mais eficiente, simplificando a análise dos dados rotulados ao abstrair detalhes complexos, como a intensidade dos pixels, em características mais gerais e significativas, como as bordas (RAINA *et al*., 2007).

Em comparação com os métodos de Aprendizado Supervisionado, que exigem um par de dados Xi e Yi, no qual Yi é anotado por humanos, o Aprendizado Autossupervisionado também é treinado com os dados Xi, juntamente com seu pseudo rótulo Pi, sendo que o Pi é gerado automaticamente para uma tarefa auxiliar predefinida, sem envolver qualquer anotação humana (JING; TIAN, 2019). As tarefas auxiliares geralmente são categorizadas em três grupos: contrastivas, preditivas e generativas.

No primeiro caso, o objetivo é otimizar a discriminação entre imagens contrastantes, ou seja, minimizar a distância entre pares positivos e maximizar a distância entre pares negativos. Vamos imaginar que em um exemplo seja necessário detectar um capacete de proteção individual em uma pessoa. A situação positiva seria o indivíduo no ato do uso do capacete; a negativa, seria o constraste disso, ou seja, o não do uso do acessório. Então, dado um par de dados de entrada, os codificadores contrastivos selecionados aprenderão a partir desses para calcular as representações das vistas e, em seguida, usarão um módulo discriminador para comparar a similaridade das instâncias e calcular a perda contrastiva, como mostrado no Fluxograma a seguir (TIAN *et al*., 2020).

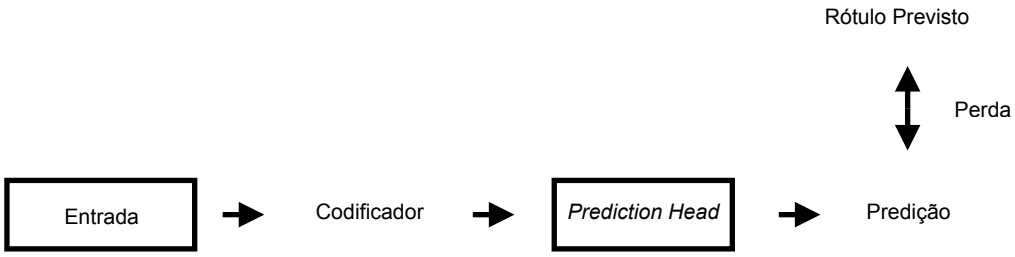
Fluxograma 1 - Esquema de uma tarefa auxiliar contrastiva.



Fonte: O autor, 2025.

Na segunda abordagem, após identificar propriedades ou partes específicas dos dados de entrada, os modelos preditivos irão prever um novo rótulo. Em geral, as tarefas preditivas consistem em um codificador e uma ou mais *prediction head*, como mostrado no Fluxograma a seguir. O sistema compara os rótulos reais e previstos para fornecer uma medida de perda (ZHAO *et al*., 2024).

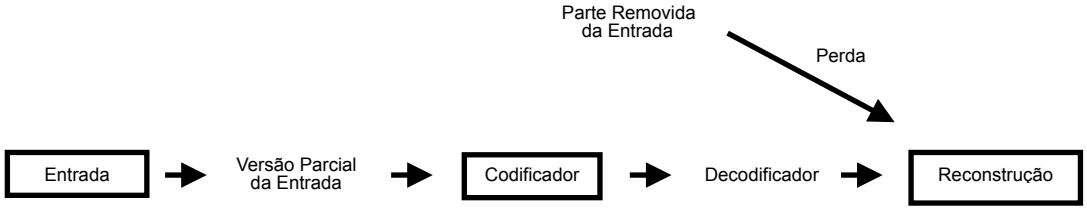
Fluxograma 2 - Esquema de uma tarefa auxiliar preditiva.



Fonte: O autor, 2025.

No último caso, o modelo é treinado para reconstruir uma parte do dado de entrada original a partir de uma versão parcialmente completa para o aprendizado de características. A ideia é que o modelo pode recuperar as informações ausentes se as características contextuais forem bem aprendidas (GUO; ZHU; LI, 2021). A perda é calculada pela comparação entre a parte faltante no dado de entrada com a parte gerada pelo algoritmo. As etapas de construção estão demonstradas no Fluxograma abaixo.

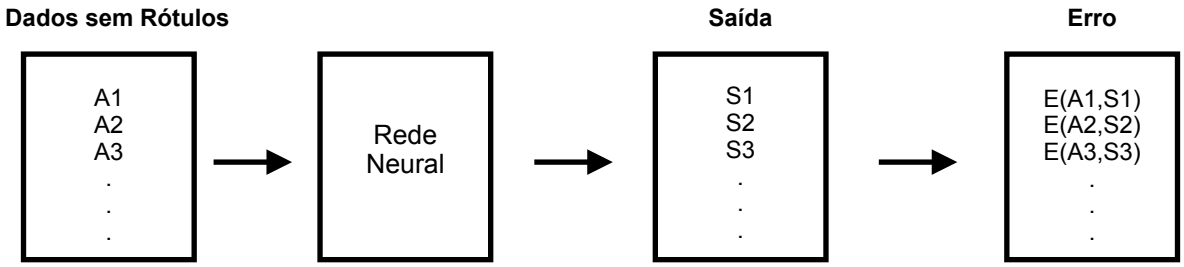
Fluxograma 3 - Esquema de uma tarefa auxiliar generativa.



Fonte: O autor, 2025.

Independentemente da abordagem utilizada nas tarefas auxiliares, o objetivo é minimizar o erro entre os pseudo rótulos e as predições da Rede Neural treinada, como mostrado no Fluxograma abaixo.

Fluxograma 4 - Esquema genérico de um Aprendizado Autossupervisionado.



Fonte: O autor, 2025.

2.2.3.1 Aplicações do Aprendizado Autossupervisionado

Zhao *et al*. (2024) construíram uma tabela com 80 aplicações recentes, entre 2020 e 2023, que obtiveram performance no nível do estado da arte utilizando aprendizado autossupervisionado. Divididas em 11 áreas de aplicação, que variam de indústria à física, os artigos estão listados na Tabela 2 a seguir.

Tabela 2 - Lista de aplicações recentes do estado da arte para Aprendizado Autossupervisionado.

No.	Methods	Articles	Application tasks	Categories	Architectures	Performance (%)
1	TL	Xue et al. (2020)	Cervical histopathology	Medical/Image classification	Ensembled TL	98.61/Acc
2	TL	Lopes et al. (2021)	Heart disease detection	Medical/Image classification	Sex identify model	83/Acc
3	TL	De Bois, El Yacoubi, and Ammi (2021)	Blood glucose prediction	Medical/Image classification	Adversarial FCN	18.94/RMSE*
4	SSL	Fedorov et al. (2021)	Alzheimer prediction	Medical/Image classification	DCGAN	95/AUC
5	TL	Bargshady et al. (2022)	COVID-19 detection	Medical/Image classification	Inception-CycleGAN	94.2/Acc
6	TL	Gardner, Bull, Dervilis, and Worden (2022)	Structural monitoring	Medical/Image classification	KBTL	98.5/Acc
7	TL	Kathamuthu et al. (2023)	COVID-19 detection	Medical/Image classification	VGG-16	98/Acc
8	SSL	Jiao, Droste, Drukker, Papageorghiou, and Noble (2020)	Ultrasound video	Medical/Video classification	Siamese network	75.7/F1
9	SSL	Xue and Salim (2021)	COVID-19 classification	Medical/Audio classification	Transformer-CP	84.43/Acc
10	TL	McCreery, Katariya, Kannan, Chablani, and Amatriain (2020)	COVID-19 FAQs	Medical/Text generation	BERT-QA	84.5/Acc
11	SSL	Yoon, Zhang, Jordon, and van der Schaar (2020)	Patient treatment prediction	Medical/Tabular prediction	VIME	86.02/AUC
12	TL	Ankenbrand et al. (2021)	MRI segmentation	Medical/Image segmentation	Modified TL	90/DSC
13	TL	Zhao et al. (2021)	Fundus segmentation	Medical/Image segmentation	TL based U-Net	99.75/Acc
14	SSL	Azizi et al. (2023)	Medical vision	Medical/Multi-task	REMEDIS	95.8/AUC
15	SSL	Moor et al. (2023)	Generalist medical AI	Medical/Multi-task	GMAI	–
16	TL	Zhang et al. (2020)	Ball screw diagnosis	Industrial/Image classification	IEDT	94.95/Acc
17	TL	Xin, Cheng, Diender, and Veljkovic (2020)	Bridge engineering	Industrial/Image classification	TL-based CNN	99.05/Acc
18	TL	Zhu, Chen, Anduv, Jin, and Du (2021)	Chillers diagnosis	Industrial/Image classification	FDD based TL	81.27/Acc
19	TL	Michau and Fink (2021)	Anomaly detector	Industrial/Image classification	Adversarial TL	99.8/Acc
20	TL	Asanuma, Doi, and Igarashi (2020)	Electric motor	Industrial/Regression	VGG16	–
21	SSL	Ren, Wang, Laili, and Zhang (2021)	Soft sensor	Industrial/Regression	LSTM-DeepFM	0.7479/RMSE*
22	SSL	Akrim, Gogu, Vingerhoeds, and Salaün (2023)	Fatigue damage prediction	Industrial/Regression	SSL-based	13.24/MAPE
23	TL	Zhang, Wang, and Li (2023)	Fault diagnosis	Industrial/Federated learning	Blockchain framework	99.8/Acc
24	SSL	Yuan and Lin (2020)	City classification	Satellite/Image classification	SITS-BERT	98.76/Acc
25	SSL	Li et al. (2023)	Scene classification	Satellite/Image classification	MES^{2}L	95.03/Acc
26	SSL	Muhtar, Zhang, and Xiao (2022)	Object segmentation	Satellite/Image segmentation	IndexNet	71.07/mIoU
27	TL	Chen, Sun, Li, and Hou (2022)	Plane detection	Satellite/Object detection	DA R-CNN	90.17/AP
28	TL	Gomroki, Hasanlou, and Reinartz (2023)	City change detection	Satellite/Object detection	EfficientNetV2 T-Unet	97.66/Acc
29	SSL	Xiao et al. (2023)	Image super-resolution	Satellite/Image degradation	SSL network	–
30	SSL	Liu, Jiang, Xiong, Yang, and Ye (2020)	Chat bot	NLP/Text generation	MRTM	87.7/F1
31	SSL	Cao, Jin, Wan, and Yu (2020)	Auto essay scoring	NLP/Text generation	HA-LSTM+SST+DAT	79.7/QWK
32	TL	Van Nguyen, Nguyen, Min, and Nguyen (2021)	Crosslingual translation	NLP/Text translation	CCCAR	72.1/F1
33	TL	Omran, Sharef, Grosan, and Li (2023)	Crosslingual translation	NLP/Text translation	LSTM model	96.65/F1
34	TL	Lu et al. (2021)	Event extraction	NLP/Information retrieval	TEXT2EVENT	72.7/F1
35	TL	Elmadany, Abdul-Mageed, et al. (2022)	Crosslingual task	NLP/Multi-task	AraT5	24.37/Bleu
36	TL	Xu, Dinkel, Wu, Xie, and Yu (2021)	Audio caption generation	NLP/Multi-modal	CNN10	65.5/Bleu
37	TL	Sung, Cho, and Bansal (2022)	Vision-text generation	NLP/Multi-modal	CLIP-BART	79.2/Acc
38	SSL	Li et al. (2020)	3D pose estimation	Motion/Regression	Modified SSL	41.4/AUC
39	SSL	Bhatnagar, Sminchisescu, Theobalt, and Pons-Moll (2020)	3D pose estimation	Motion/Regression	LoopReg	–
40	SSL	Dai et al. (2020)	3D depth estimation	Motion/Regression	Depth-net	72.31/IoU
41	SSL	Spurr, Dahiya, Wang, Zhang, and Hilliges (2021)	Hand pose estimation	Motion/Regression	PeCLR	74/AUC
42	SSL	Cheng, Chen, Zhang, and Lin (2021)	Action recognition	Motion/Regression	Motion-Transformer	91.9/Acc
43	SSL	Ma, Li and Li (2023)	3D pose estimation	Motion/regression	Factorisation network	40.4/AUC
44	TL	Huang, Fu, Liu and Ostadabbas (2021)	Infant pose estimation	Motion/Image estimation	FiDIP	93.6/mAP
45	SSL	Baur et al. (2021)	3D scene segmentation	Motion/Image segmentation	SLIM	0.0668/AEE*
46	SSL	Huang et al. (2021)	Action recognition	Motion/Video classification	MoSI	71.8/Acc
47	SSL	Bi, Hu, Zhao, Li, and Sun (2023)	Action recognition	Motion/Video classification	Contrastive-SSL	75.7/Acc
48	TL	Cao and Xiang (2020)	Garbage classification	Environment/Image classification	Modified Inception-V3	99.3/Acc
49	TL	Choe, Choi, and Kim (2020)	Parrot classification	Environment/Image classification	Keras-based TL	96.75/F1
50	TL	Rehman et al. (2021)	Leaf disease recognition	Environment/Image classification	ResNet+Mask RCNN	96.6/Acc
51	TL	Huang, Chuang, and Liao (2022)	Tomato pest identification	Environment/Image classification	ResNet50+DA	97.12/Acc
52	TL	Attallah (2023)	Tomato leaf classification	Environment/Image classification	Resnet-18	99.34/Acc
53	TL	Palanisamy, Singhania, and Yao (2020)	Environment sound classification	Environment/Audio classification	Ensemble DenseNet	92.89/Acc
54	TL	Chen et al. (2021)	Water quality prediction	Environment/Text prediction	TrAdaBoost-LSTM	0.47/RMSE*
55	SSL	Shen et al. (2020)	Taxonomy expansion	Web/Information retrieval	TaxoExpan-FWFS	72.3/F1
56	SSL	Yao et al. (2021)	Item recommendation	Web/Information retrieval	SSL-DNN	53.55/Recall
57	SSL	Zhang et al. (2021)	Group recommendation	Web/Information retrieval	$\mathcal{S}^2$-HHGR	88.3/Recall
58	SSL	Zhou, Dou, Zhu, and Wen (2021)	Personalised search	Web/Information retrieval	PSSL	83.01/mAP
59	SSL	Feng, Wan, Wang, Li, and Luo (2021)	Twitter bot account	Web/Text classification	SATAR	95.09/Acc
60	TL	Ameer et al. (2023)	Emotion classification	Web/Text classification	RoBERTa-MA	62.4/Acc
61	TL	Arbane, Benlamri, Brik, and Alahmar (2023)	Emotion classification	Web/Text classification	Bi-LSTM	93/Recall
62	SSL	Huh, Heo, Kang, Watanabe, and Chung (2020)	Speaker recognition	Speaker/Audio classification	Light ResNet-34	8.65/EER
63	SSL	Chi et al. (2021)	Speaker classification	Speaker/Audio classification	Audio ALBERT	99.3/Acc
64	SSL	Xia, Zhang, Weng, Yu, and Yu (2021)	Speaker verification	Speaker/Audio classification	MoCo+WavAug	8.63/EER
65	SSL	Sang, Li, Liu, Arnold, and Wan (2022)	Speaker verification	Speaker/Audio classification	Thin-ResNet34	6.99/EER
66	SSL	Chen et al. (2023)	Speaker verification	Speaker/Audio classification	WavLM-based SS	16.5/WER
67	SSL	Cai, Wang, and Li (2021)	Speaker recognition	Speaker/Audio segmentation	Iterative SSL	3.45/EER
68	SSL	Huang, Raj, García, and Khudanpur (2023)	Speaker recognition	Speaker/Object localisation	JSM	7.58/WER
69	SSL	Hu et al. (2020)	Speaker localisation	Speaker/Multi-modal	Modified SSL	51.9/IoU
70	SSL	Song, Wang, Fan, Tan, and Zhang (2022)	Speaker localisation	Speaker/Multi-modal	SSPL	61/AUC
71	SSL	Li, Huang, and Zhang (2021)	Vehicle re-identify	Vehicle/Image classification	SSL network	98.7/Acc
72	SSL	Kothandaraman, Chandra, and Manocha (2021)	Road segmentation	Vehicle/Image segmentation	SS-SFDA	95.63/F1
73	SSL	Kumar et al. (2021)	Road segmentation	Vehicle/Image segmentation	SynDistNet	1.668/RSME*
74	TL	Hu, Zhao and Zhang (2020)	Infrared pedestrian detection	Vehicle/Object detection	Faster R-CNN	84.78/AP
75	TL	Farid et al. (2023)	Vehicle detection	Vehicle/Object detection	YOLO-v5	99.94/Acc
76	SSL	Wei et al. (2023)	Depth estimation	Vehicle/Image prediction	Transformer	6.835/RMSE*
77	SSL	Yan, Zhu, Jin, and Bohg (2020)	Robot manipulation	Robot/Image prediction	Dynamic model	–
78	SSL	Nubert, Khattak, and Hutter (2021)	LiDAR localisation	Robot/Image estimation	Modified SSL	–
79	SSL	Ze, Hansen, Chen, Jain, and Wang (2023)	Motor control	Robot/Image reconstruction	3D auto-encoder	–
80	TL	Inubushi and Goto (2020)	Fluid turbulence prediction	Physics/Nonlinear prediction	Modified TL	0.1/MSE

Fonte: Zhao *et al*, 2024.

Mesmo com a variedade listada, não há nenhum trabalho citado no contexto de segurança. Apesar de haver quatro artigos com a finalidade de detecção de objetos, foco da presente dissertação, todos utilizam *transfer learning* a partir de anotações manuais de toda a base de dados, ou seja, Aprendizado Supervisionado. Mesmo que esse seja um pequeno recorte, é de fato escasso o número de trabalhos que utilizam Aprendizado Autossupervisionado para detecção de EPIs. Por isso, esta dissertação se aterá a citar artigos que aplicam Aprendizado Autossupervisionado em detecção de objetos.

Em 2020, Li *et al.* utilizaram Aprendizado Autossupervisionado para aumentar uma base de dados de cadeiras e mochilas. Isso foi feito pois naquela data, a acurácia para detecção desses objetos, com a base devidamente rotulada, era abaixo dos 40%. No artigo de comparação, escrito por Lin *et al* em 2014, foram utilizadas 70 mil horas de trabalho manuais e atribuídos 2.5 milhões de rótulos em 328 mil imagens. Com o uso de 5 mil imagens de mochilas e quase 13 mil de cadeiras do artigo de referência, Li *et al* adicionaram mais de 250 mil imagens da internet, sendo cerca de um terço para o primeiro objeto e o restante para o segundo, sem a necessidade de anotação manual, para ao fim obter uma acurácia acima dos 50%.

2.2.4 Aprendizado Semi Supervisionado

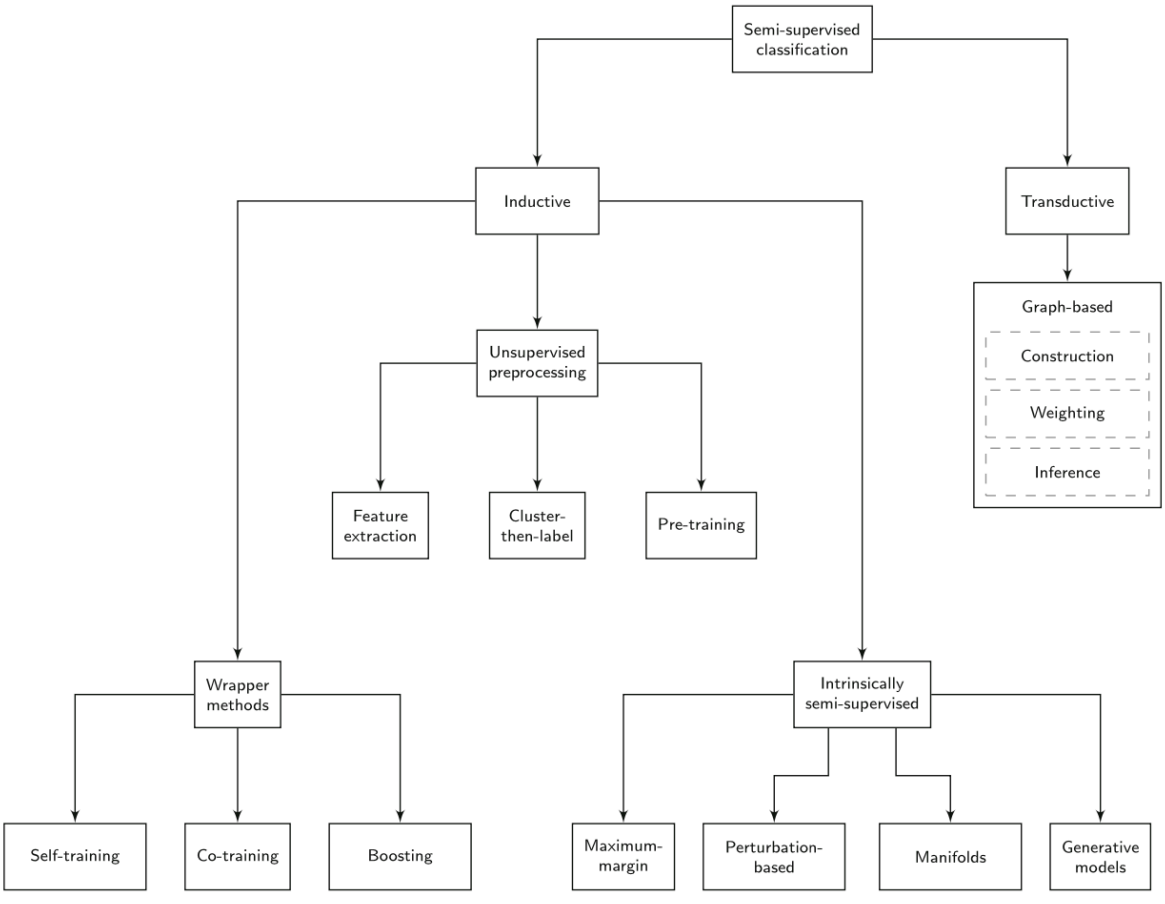
O aprendizado semi supervisionado é uma abordagem de aprendizado que constrói algoritmos que utilizam dados rotulados e não rotulados (YANG *et al*., 2021). Alguns resultados recentes mostraram que, em certos casos, o Aprendizado Semi Supervisionado se aproxima do desempenho do Aprendizado Supervisionado, mesmo quando uma parte dos rótulos em um determinado conjunto de dados foi descartada. Esses resultados são demonstrados ao pegar uma base de dados de classificação existente e utilizar apenas uma pequena parte dele como dados rotulados, com o restante tratado como não rotulado (OLIVER *et al*., 2019).

O aprendizado semi supervisionado pode ser usado em três grandes tipos de problema de Inteligência Artificial: Classificação, Agrupamento e Regressão, sendo o primeiro essencial para o presente trabalho. Por ser uma combinação do

Aprendizado Supervisionado e Não Supervisionado, essa abordagem mistura suas características para surgir como uma nova possibilidade de resolução.

Uma condição importante para o aprendizado semi supervisionado é que a distribuição dos dados de entrada contenha informações sobre a distribuição dos rótulos de saída. Em outras palavras, os dados não rotulados devem ter alguma relação com os rótulos que queremos prever. Se isso for verdade, podemos usar os dados não rotulados para aprender mais sobre os dados de entrada e, assim, também sobre os rótulos de saída. No entanto, se essa condição não for atendida, ou seja, se os dados de entrada não fornecerem informações úteis sobre os rótulos, não será possível melhorar a precisão das previsões usando os dados não rotulados adicionais ou até piorá-la (ENGELER; HOOS, 2019). Com essa premissa, diversos métodos podem ser empregados para resolver problemas de classificação, como mostrado na Figura 3 a seguir.

Figura 3 - Lista de métodos que podem ser empregados em problemas de classificação.



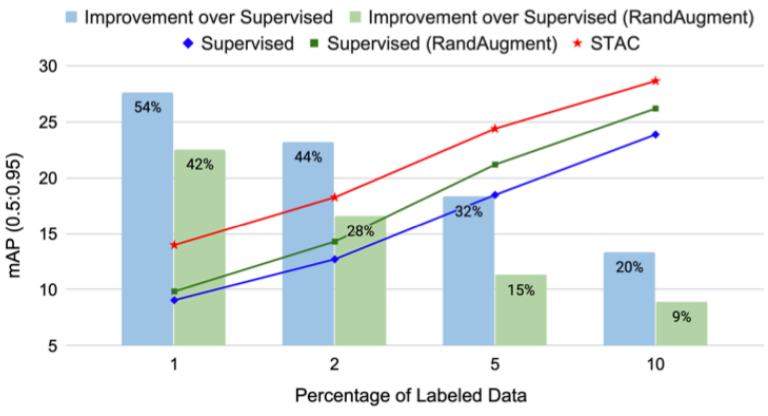
Fonte: Engeler & Hoos, 2019.

2.2.4.1 Aplicações do Aprendizado Semi Supervisionado

Em 2020, Sohn *et al*. introduziram o *FixMatch*, que representou uma simplificação significativa dos métodos de aprendizado semi supervisionado mantendo uma performance competitiva. Os resultados do *FixMatch* estabeleceram novos marcos à época no tema em comparação com outros resultados. No CIFAR-10 com apenas 250 exemplos rotulados, o método alcançou 94.93% de acurácia comparado aos 93.73% do estado da arte anterior. Além disso, o *FixMatch* demonstrou eficácia fora do comum em cenários de extrema escassez de rótulos, obtendo 88.61% de acurácia no CIFAR-10 com apenas 4 rótulos por classe.

Em 2021, Sohn *et al.* propuseram um novo *framework* para aprendizado semi supervisionado, chamado STAC, que introduz apenas dois novos hiperparâmetros: o limiar de confiança e o peso da perda não supervisionada. O teste foi feito em dois banco de dados conhecidos (MS COCO e VOC07) para comparação com os trabalhos do estado da arte de algoritmos supervisionados e com modificações na porcentagem de dados rotulados utilizados, sendo 1%, 2%, 5% e 10% do total, com a finalidade de detectar objetos. Quanto menor o percentual, melhor foi o resultado, como mostrado na Figura 4 abaixo.

Figura 4 - Resultados da aplicação do *framework* STAC em comparação com o Aprendizado Supervisionado.



Fonte: Sohn *et al*, 2021.

É importante notar que o resultado chegou a ser 54% melhor que o estado da arte de algoritmos supervisionados, com a melhora em todas as quatro distribuições.

Em 2021, Liu *et al*. propuseram o *Unbiased Teacher*, um *framework* que aborda especificamente o problema de viés em *pseudo-labels* causado pelo desequilíbrio de classes inerente à detecção de objetos. O método introduz uma abordagem *Teacher-Student* que treina mutuamente dois modelos: o *Teacher* gera *pseudo-labels* para treinar o *Student*, enquanto o *Student* atualiza gradualmente o *Teacher* via *Exponential Moving Average* (EMA). O *framework* também incorpora *Focal loss* para mitigar o problema de desbalanceamento entre classes, questão crítica em detecção de objetos que não é adequadamente endereçada pelos métodos de classificação de imagens tradicionais. Os resultados experimentais do *Unbiased Teacher* demonstraram melhorias substanciais em relação aos métodos existentes. No *dataset* MS-COCO superou significativamente o STAC, representando uma performance melhor em todos os casos sobre o estado da arte no momento, como pode ser visto na Tabela 3. Além disso, obteve 10 pontos percentuais a mais de mAP quando utilizado com menos de 5% de dados rotulados comparado ao *baseline* supervisionado.

Tabela 3 - Resultados da aplicação do *Unbiased Teacher* em comparação com outros para o *dataset* MS COCO.

	0.5%	1%	2%	5%	10%
Supervised	6.83 ± 0.15	9.05 ± 0.16	12.70 ± 0.15	18.47 ± 0.22	23.86 ± 0.81
CSD*	7.41 ± 0.21 (+0.58)	10.51 ± 0.06 (+1.46)	13.93 ± 0.12 (+1.23)	18.63 ± 0.07 (+0.16)	22.46 ± 0.08 (-1.40)
STAC	9.78 ± 0.53 (+2.95)	13.97 ± 0.35 (+4.92)	18.25 ± 0.25 (+5.55)	24.38 ± 0.12 (+5.86)	28.64 ± 0.21 (+4.78)
Unbiased Teacher	**16.94 ± 0.23** (+10.11)	**20.75 ± 0.12** (+11.72)	**24.30 ± 0.07** (+11.60)	**28.27 ± 0.11** (+9.80)	**31.50 ± 0.10** (+7.64)

Fonte: Liu *et al*, 2021.

Em 2024, Qi, Nguyen e Yan propuseram um outro *framework* para Aprendizado Semi supervisionado, chamado CISO. Para maximizar a utilização dos dados de *pseudo-labels* e lidar com a escassez de dados de pseudo-rótulos devido a configurações de limiar alto, foi proposta uma abordagem de iteração média, onde todos os dados não rotulados são aplicados a cada iteração de treinamento. O teste foi aplicado no conjunto de dados do MS COCO e VOC, com 1%, 5% e 10% de dados rotulados. Neste caso, quanto maior o percentual de dados rotulados, melhor foi o resultado. Em todos as situações, o CISO teve uma performance superior que o STAC, como pode ser visto nas Tabelas 4 e 5 abaixo. Para o *dataset* VOC, o resultado é melhor que todas as outras abordagens.

Tabela 4 - Resultados da aplicação do *framework* CISO em comparação com outros para o *dataset* MS COCO.

Method		1%	5%	10%
Anchor based	Supervised	9.05±0.16	18.47±0.22	23.86±0.81
	CSD [15]	10.20±0.15	18.90±0.10	24.50±0.15
	STAC [40]	13.97±0.35	24.38±0.12	28.64±0.21
	DETReg [5]	14.58±0.30	24.80±0.20	29.12±0.20
	Instant Teaching [60]	18.05±0.15	26.75±0.05	30.40±0.05
	ISMT [51]	18.88±0.38	26.37±0.24	30.53±0.52
	Unbiased Teacher [25]	20.75±0.12	28.27±0.11	31.50±0.10
	Soft Teacher [50]	20.46±0.39	30.74±0.08	34.04±0.14
	LabelMatch [7]	25.81±0.28	32.70±0.18	35.49±0.17
Anchor free	HT [43]	16.96±0.36	27.70±0.15	31.61±0.28
	Ours (CISO*)	21.04±0.18	29.50±0.21	34.20±0.12
	Ours (CISO)	22.00±0.17	30.90±0.15	36.20±0.26

Fonte: Qi; Nguyen; Yan, 2024.

Tabela 5 - Resultados da aplicação do *framework* CISO em comparação com outros para o *dataset* VOC.

Methods	AP_{50}	$AP_{50:95}$
Supervised	72.75	42.04
CSD [15]	74.70	-
STAC [40]	77.45	44.64
Instant Teaching [60]	79.20	50.00
Ours (CISO)*	80.39	51.77
Ours (CISO)	81.44	52.98
CSD [15]	75.10	-
STAC [40]	79.08	46.01
Instant Teaching [60]	79.90	50.80
Ours (CISO*)	83.03	53.83
Ours (CISO)	84.48	55.30

Fonte: Qi; Nguyen; Yan, 2024.

Em 2024, Liu e Wang abordaram o problema da detecção de EPIs em contextos com escassez de dados através de técnicas de rotulação semi-automática. Os autores criaram um *dataset* considerando a detecção de capacetes e roupas refletivas. O modelo proposto, denominado AL-YOLOv5, incorpora mecanismos de atenção para melhorar a extração de características e uma função de perda aprimorada para enfrentar desafios relacionados à detecção de roupas refletivas e à sobreposição de *bounding boxes*.

Os resultados experimentais demonstraram avanços notáveis de 0,9 AP na categoria com dados limitados e 0,4 mAP no geral comparado ao YOLOv5 *baseline*, com progresso particularmente substancial na detecção de roupas refletivas, reduzindo significativamente as falsas detecções e melhorando *frames* de sobreposição. Este trabalho é especialmente relevante no contexto da presente dissertação por três razões: primeiro, demonstra que a combinação de rotulação semi-automática com arquiteturas aprimoradas pode reduzir significativamente o esforço de anotação manual; segundo, evidencia que classes como roupas refletivas, podem se beneficiar de abordagens semi supervisionadas; terceiro, estabelece um precedente metodológico para o uso de técnicas de rotulação semi-automática em ambientes industriais reais, alinhando-se diretamente com os objetivos desta pesquisa.

3 METODOLOGIA

3.1. COLETA DOS DADOS

3.1.1. Cenário industrial e vídeos capturados

A coleta de dados foi feita em uma indústria localizada no interior do estado de São Paulo, onde foi instalada uma câmera VIP1230B da marca Intelbras para captura contínua de uma seção na entrada da fábrica. O monitoramento operou com a gravação de vídeos durante diferentes períodos do dia, incluindo os turnos da manhã e da tarde. Esta diferença de horário é importante para promover maior variabilidade nas condições de iluminação, proporcionando maior robustez ao *dataset* resultante.

O processo de aquisição de dados inicial resultou em um conjunto de 26 vídeos capturados em um local específico, no qual os colaboradores deviam checar o uso dos equipamentos de proteção individual. A decisão de incluir todos os vídeos gravados no período de coleta, sem aplicação de critérios de seleção específicos, foi motivada pela necessidade de manter representatividade parecida com as condições operacionais reais. Isso inclui variações na densidade de trabalhadores, diferentes formas de checagem e variabilidade no uso efetivo dos EPIs. A pequena quantidade coletada também influenciou essa decisão.

A extração de *frames* resultou em um total de 50.088 imagens estáticas que compõem o *dataset* final. Este processo de decomposição dos vídeos permite a aplicação de técnicas de detecção de objetos *frame* a *frame*, essencial para a metodologia proposta do caso em questão.

3.1.2. Estruturação dos *frames* e classes

A estruturação do *dataset* seguiu o padrão de anotação na qual cada *frame* é acompanhado por um par correspondente contendo as coordenadas normalizadas dos objetos detectados. Todas as anotações foram feitas em um total de 77 dias,

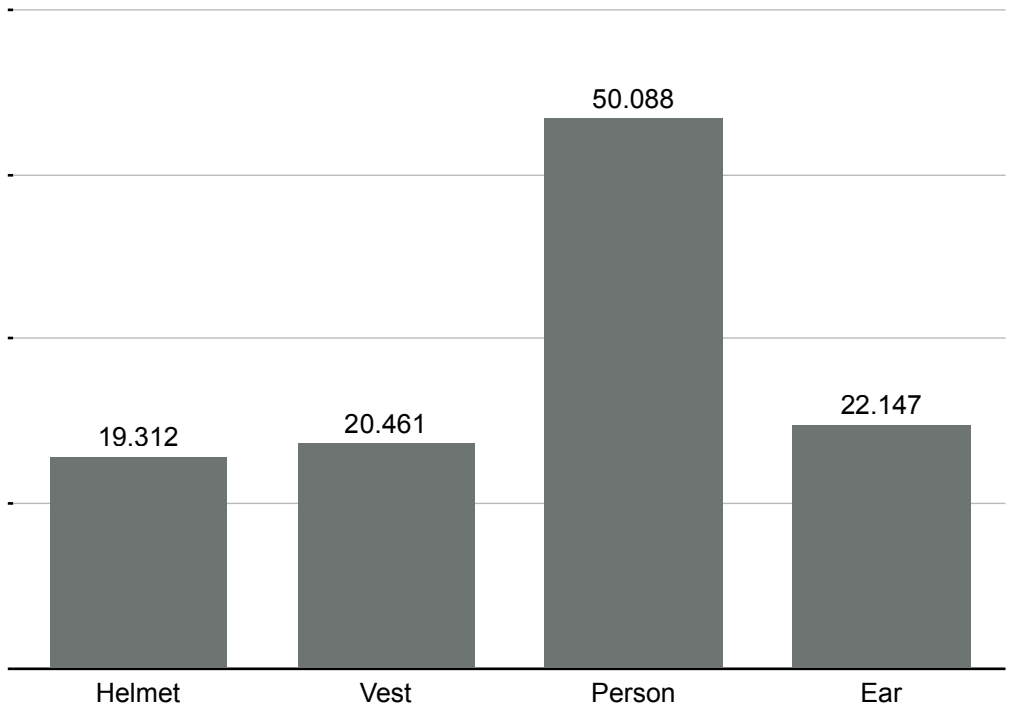
com o uso da plataforma CVAT (*Computer Vision Annotation Tool*), uma ferramenta *open source* que permite a segmentação e detecção de objetos, este último o caso do presente trabalho. A carga diária variou entre 1h30 e 3h de anotações, feitas majoritariamente no turno da manhã.

O sistema foi desenvolvido para a detecção automática de quatro classes principais de objetos relacionados à segurança ocupacional da indústria em questão, definidas com base nas normas regulamentadoras brasileiras (NR-6) e práticas internacionais de segurança industrial:

- *Helmet* (Capacete de Segurança): destinado à proteção do crânio contra impactos, perfurações e choques elétricos;
- *Vest* (Colete de Segurança): destinado a aumentar a visibilidade do trabalhador, reduzindo o risco de acidentes, especialmente em ambientes com pouca luz ou em áreas de risco com presença de veículos ou máquinas;
- *Person* (Pessoa): detecção e localização de indivíduos presentes no ambiente industrial, fundamental para análise contextual do uso de EPIs;
- *Ear* (Protetor Auricular tipo Concha): destinado a bloquear ou reduzir a intensidade do som, protegendo os ouvidos de danos causados pela exposição a níveis de ruído perigosos.

A análise estatística preliminar do *dataset* revelou uma distribuição desbalanceada entre as classes, característica inerente a cenários reais, observada no Gráfico abaixo.

Gráfico 1 - Histograma do *dataset*.



Fonte: O autor, 2025.

A classe *Person* apresenta frequência significativamente superior às demais, uma vez que trabalhadores estão constantemente presentes no ambiente, independentemente do uso de EPIs específicos, mas o contrário não é válido, ou seja, EPIs que apareceram no *frame,* mas não foram vestidos por indivíduos, não foram anotados, pois são casos que não devem ser detectados pelo sistema final. Esta distribuição não foi artificialmente corrigida através de técnicas de balanceamento sintético, pois reflete fielmente as condições operacionais reais nas quais nem todos os trabalhadores utilizam simultaneamente todos os equipamentos de proteção requeridos. Dados rotulados e não rotulados podem não aderir à suposição de dados independentes e distribuídos de forma idêntica. Isso ocorre porque dados não rotulados podem originar-se de cenários diferentes daqueles dos dados rotulados em casos reais (QI *et al.*, 2023) Por fim, o balanceamento destas classes quase impossibilitaria a aplicação, a não ser que todos os *frames* tivesse a presença de uma pessoa com todos os EPIs. Porém, neste caso, o algoritmo não

iria ter em sua base nenhum *frame* de um indivíduo sem EPI e provavelmente falharia nessas situações.

3.2. ESTRATÉGIAS DE ANOTAÇÃO PARCIAL

A implementação de estratégias de anotação parcial constitui o núcleo metodológico deste trabalho, visando avaliar a viabilidade de uma abordagem semi supervisionada em uma indústria real. Foi desenvolvido um algoritmo baseado na distribuição de classes para seleção dos subconjuntos de treinamento, garantindo representatividade estatística. Inicialmente, é calculada a frequência total de cada classe no *dataset* completo, estabelecendo uma distribuição de referência que representa as características estatísticas globais dos dados. Em seguida, o algoritmo procede à seleção sequencial de vídeos e avalia iterativamente qual combinação melhor preserva a proporção original das classes no subconjunto selecionado. O processo de seleção utiliza uma função de distância que calcula a divergência entre a distribuição do subconjunto candidato e a distribuição global de referência. Esta função de distância considera as frequências relativas de todas as classes simultaneamente, penalizando seleções que resultem em desvios significativos da distribuição original. A cada iteração, o algoritmo seleciona o vídeo que, quando adicionado ao subconjunto atual, resulta na menor distância possível em relação à distribuição de referência.

Para o cenário experimental de 10% de dados anotados manualmente, a seleção resultou em uma divisão estratégica na qual 2 vídeos foram destinados ao conjunto de treinamento, 3 vídeos ao conjunto de validação e mais 18 vídeos foram reservados para posterior anotação automática. Similarmente, para o cenário de 20% de dados anotados manualmente, a estratégia de seleção identificou 4 vídeos para treinamento e 3 vídeos para validação, deixando 16 vídeos para anotação automática posterior. Para o cenário de 30%, foram destinados 6 vídeos para treinamento, 3 para validação e os 14 restantes foram anotados automaticamente em seguida. Adicionalmente, para os vídeos restantes (3 em cada caso), foi estabelecida uma reserva estratégica dos dados totais exclusivamente para avaliação final, garantindo que a performance dos modelos seja testada em dados completamente não vistos durante qualquer etapa do processo de desenvolvimento.

Esta abordagem é fundamental para evitar vazamento de informações entre conjuntos e garantir avaliação imparcial dos resultados.

3.3. MODELO 1 - GERAÇÃO DE RÓTULOS

A implementação do Modelo 1, em ambos os casos de 10%, 20% e 30% de rotulação manual, representa a fase inicial da abordagem semi supervisionada proposta, utilizando a técnica de *pseudo-labeling* para expandir automaticamente a base de dados anotados. Esta estratégia fundamenta-se no princípio de que um modelo treinado com dados parcialmente anotados pode generalizar suficientemente bem. Dessa forma, pode produzir anotações automáticas de qualidade aceitável em dados não rotulados, incluindo um *dataset* desbalanceado (LEE, 2013). O Modelo 1 foi implementado utilizando a arquitetura YOLOv8 *medium*, escolhida por sua comprovada eficiência em tarefas de detecção de objetos em tempo real e sua capacidade de generalização em diferentes cenários. O YOLOv8 representa o estado da arte em detecção de objetos, combinando velocidade de processamento com precisão de detecção através de sua arquitetura baseada em redes neurais convolucionais profundas.

O processo de treinamento do modelo base iniciou-se com a utilização de pesos pré-treinados fornecidos pela *Ultralytics*, aproveitando representações de características aprendidas em *datasets* de larga escala. Esta estratégia de t*ransfer learning* é fundamental para compensar a limitação de dados anotados disponíveis, permitindo que o modelo inicie o treinamento com conhecimento prévio sobre detecção de objetos genéricos (ALI *et al*., 2023). O treinamento foi conduzido por 150 épocas, período determinado através de análise da convergência do modelo nos conjuntos de validação. Foram realizados experimentos preliminares com durações variadas (50, 100, 150 e 200 épocas) nos quais se observou que a métrica mAP@0.5 no conjunto de validação estabilizava consistentemente após aproximadamente 100-120 épocas, apresentando variações inferiores a 0.5% nas épocas subsequentes. O limite de 150 épocas foi estabelecido para garantir convergência completa em todos os cenários experimentais, considerando que o uso de diferentes proporções de dados anotados poderia resultar em taxas de convergência ligeiramente distintas. Durante este processo, foram utilizados os

subconjuntos de dados parcialmente anotados (10%, 20% ou 30%), com monitoramento contínuo das métricas de performance para possível detecção de *overfitting* e ajuste automático de hiperparâmetros.

Após a conclusão do treinamento inicial, o modelo foi aplicado aos dados não anotados para geração automática de *pseudo-labels*. Este processo de inferência utilizou um limiar de confiança adaptativo para todas as classes, com o limiar ótimo mostrado na Tabela 6 a seguir.

Tabela 6 - Limiar de confiança ótimo por modelo e por classe.

Classe	10%	20%	30%	100%
Helmet	0.3	0.3	0.3	0.3
Vest	0.4	0.5	0.7	0.7
Person	0.3	0.4	0.3	0.3
Ear	0.6	0.3	0.5	0.6

Fonte: O autor, 2025.

O processo de geração de *pseudo-labels* incluiu a aplicação de supressão de não-máximos (Non-Maximum Suppression - NMS) para eliminação de detecções redundantes e sobrepostas, evitando duplicações que poderiam confundir o treinamento subsequente do Modelo 2. As coordenadas das detecções foram normalizadas seguindo o padrão YOLO, na qual cada *bounding box* é representada pelas coordenadas do centro (x, y) e dimensões (largura, altura), todas normalizadas em relação às dimensões da imagem. Este formato padronizado facilita o processamento posterior e garante compatibilidade com os dados originalmente anotados.

O controle de qualidade dos *pseudo-labels* gerados foi realizado através de inspeção visual posterior, processo manual no qual todos *frames* das anotações automáticas foram verificados para identificação de erros grosseiros ou padrões sistemáticos de falhas. Isso só foi possível devido ao pequeno *dataset* em questão, sendo difícil de ser realizado em casos mais complexos. Embora esta abordagem não permita correção automatizada dos erros, fornece insights importantes sobre a qualidade geral do processo de *pseudo-labeling* e possíveis direções para melhorias futuras.

3.4. MODELO 2 - TREINAMENTO FINAL COM ROTULAÇÕES AUTOMÁTICAS

O desenvolvimento do Modelo 2 representa a última etapa experimental, na qual a base de dados expandida através dos *pseudo-labels* gerados pelo Modelo 1 é utilizada para treinamento de um modelo final com teórica capacidade de generalização superior. Esta estratégia visa combinar o conhecimento extraído dos dados originalmente anotados com as informações adicionais fornecidas pelos *pseudo-labels*. O resultado é um sistema mais robusto e preciso. O Modelo 2 foi implementado utilizando a mesma arquitetura YOLOv8 do Modelo 1. Esta escolha arquitetural permitiu a análise direta do impacto da expansão da base de dados através de *pseudo-labeling*, isolando este fator de possíveis variações introduzidas por diferentes arquiteturas de rede. A configuração de treinamento do Modelo 2 também foi feita com 150 épocas, período mostrado suficiente para convergência completa considerando a base de dados expandida. O otimizador AdamW foi utilizado por sua capacidade de adaptação automática da taxa de aprendizado e eficiente regularização através de *weight decay* integrado (LOSHCHILOV; HUTTER, 2019). O sistema de taxa de aprendizado adaptativa com *scheduler* cosseno foi empregado para garantir convergência estável e eficiente (LOSHCHILOV; HUTTER, 2017). Esta estratégia inicia o treinamento com taxa de aprendizado relativamente alta para exploração rápida do espaço de parâmetros, reduzindo gradualmente a taxa conforme o treinamento progride para refinamento fino dos pesos da rede (NAKAMURA *et al*., 2021). O padrão cosseno de decaimento proporciona transições suaves que evitam oscilações indesejadas na função de perda. O *batch size* foi determinado automaticamente baseado na capacidade de hardware disponível, maximizando a utilização de recursos computacionais enquanto mantém estabilidade numérica durante o treinamento. Esta abordagem adaptativa é particularmente importante quando se trabalha com diferentes configurações de hardware, garantindo reprodutibilidade dos resultados independentemente da plataforma computacional utilizada. As técnicas de regularização, *dropout* e o já citado *weight decay*, foram mantidas em suas configurações padrão para tentar prevenir *overfitting* (SRIVASTAVA *et al.*, 2014). Estas técnicas são especialmente importantes quando se trabalha com *pseudo-labels*, na qual a qualidade dos rótulos automáticos pode introduzir ruído no processo de aprendizado (ARAZO *et al.*, 2020).

O conjunto de dados de treinamento do Modelo 2 foi composto apenas pelos *pseudo-labels* gerados pelo Modelo 1. Esta estratégia permitiu testar a qualidade da anotação automática diretamente, uma vez que o *dataset* pequeno poderia ser significativamente influenciado pela rotulação manual.

Durante o treinamento, foi implementado monitoramento contínuo das métricas de performance no conjunto de validação, permitindo ajuste de hiperparâmetros, se necessário. A implementação de *early stopping* baseado na métrica mAP@0.5 garantiu que o treinamento fosse interrompido no ponto ótimo de generalização, evitando degradação da performance por treinamento excessivo.

3.5. MÉTRICAS DE AVALIAÇÃO

A avaliação quantitativa dos modelos desenvolvidos fundamentou-se em métricas que oferecem perspectivas complementares sobre diferentes aspectos da performance dos modelos. A métrica principal *Mean Average Precision* (mAP) com limiar de IoU de 0.5 (mAP@0.5) representa o padrão para avaliação de sistemas de detecção de objetos. Esta métrica calcula a precisão média através de diferentes níveis de *recall* para cada classe, posteriormente calculando a média geral entre todas as classes (EVERINGHAM *et al*., 2010). O limiar de IoU (*Intersection over Union*) de 0.5 determina que uma detecção é considerada correta quando a sobreposição entre a *bounding box* predita e a anotação verdadeira atinge pelo menos 50% (EVERINGHAM *et al*., 2010). A métrica mAP@0.5:0.95 proporciona uma avaliação mais rigorosa ao calcular a média das precisões médias em múltiplos limiares de IoU, variando de 0.5 a 0.95 com incrementos de 0.05. Esta abordagem multi-limiar fornece análise mais detalhada da qualidade das localizações preditas, penalizando detecções com imprecisão espacial mesmo quando a classificação está correta (PADILLA; NETTO; DA SILVA, 2020). Para aplicações críticas de segurança, nas quais a localização precisa dos EPIs pode ser fundamental para análises subsequentes, esta métrica oferece dados valiosos sobre a qualidade das fronteiras da anotação. Quando somamos a avaliação da detecção com a localização, teremos uma avaliação global do resultado.

Além dessas métricas, utilizaremos a matriz de confusão das classes, que fornece uma análise detalhada das classificações realizadas pelo modelo,

mostrando não apenas acertos e erros, mas também os padrões específicos de confusão entre classes (KOHAVI; PROVOST, 1998). Esta ferramenta é fundamental para identificar classes frequentemente confundidas pelo modelo, permitindo análise qualitativa dos tipos de erros mais comuns e direcionando esforços de melhoria. Nela, podemos calcular a precisão, definida como a proporção de detecções corretas em relação ao total de detecções realizadas para cada categoria específica (POWERS, 2011). Matematicamente, a equação que permite seu cálculo é TP/(TP+FP), na qual TP representa verdadeiros positivos e FP representa falsos positivos, o que torna possível a identificação de classes que apresentam tendência a gerar falsos alarmes. De mesmo modo, com a matriz de confusão, também conseguimos calcular o *recall,* responsável por medir a capacidade do modelo de detectar corretamente todos os objetos existentes de uma determinada classe (POWERS, 2011). Sua equação é representada como TP/(TP+FN), na qual FN representa falsos negativos. Em contextos de segurança industrial, baixo *recall* das classes pode resultar em falhas na identificação de situações de risco.

Para garantir contextualização adequada dos resultados, foi estabelecida comparação sistemática com uma abordagem totalmente supervisionada, treinada utilizando 100% dos dados manualmente anotados e os mesmos 3 vídeos de teste utilizados nas estratégias de semi supervisão. Esta proposta permite quantificação direta da eficiência da abordagem semi supervisionada, determinando quanto da performance do modelo totalmente supervisionado pode ser atingida com o uso de frações dos dados anotados.

3.6 VALIDAÇÃO ESTATÍSTICA

Para garantir robustez e validade estatística dos resultados experimentais, foi implementado protocolo rigoroso de validação baseado em múltiplas repetições independentes e testes estatísticos apropriados para o design experimental empregado.

3.6.1 Protocolo de Validação Cruzada

O protocolo de validação seguiu abordagem de validação cruzada com k=10 repetições independentes para cada cenário experimental (10%, 20%, 30% e 100% de dados anotados manualmente). O valor de k = 10 foi decidido a partir do número de combinações possíveis entre 2 vídeos de treino acrescidos de 3 de validação, resultando em uma combinação de 5 elementos tomados 2 a 2, caso experimentado na abordagem de 10%. Como os testes feitos precisam do seu par experimental, os modelos de 20%, 30% e 100% também tiveram 10 experimentos. Esta estratégia visa estimar a variabilidade natural do processo de treinamento e fornecer base estatística robusta para comparações entre as abordagens (KOHAVI, 1995). Cada repetição consistiu no treinamento completo de um modelo YOLOv8 com inicialização aleatória diferente, garantindo independência estatística entre as observações. Para cada repetição, foram registradas quatro métricas de performance: *precision*, *recall*, mAP@0.5 e mAP@0.5:0.95.

3.6.2 Teste de Friedman

Para avaliar a significância das diferenças entre tratamentos, foi empregado o teste não-paramétrico de Friedman (FRIEDMAN, 1937), apropriado para experimentos de blocos que não atendem pressupostos de normalidade (HOLLANDER *et al*., 2013). O teste de Friedman é robusto e amplamente recomendado para comparações múltiplas em aprendizado de máquina (DEMŠAR, 2006).

3.6.3 Teste de Nemenyi

Quando o teste de Friedman indicou diferenças significativas, procedeu-se à análise *post-hoc* através do teste de Nemenyi (NEMENYI, 1963), que controla o erro em comparações múltiplas par-a-par (HOLLANDER *et al*., 2013), o que resulta na compreensão de quais pares resultaram em um $p < 0,05$. O teste de Nemenyi é especificamente desenvolvido para uso após o teste de Friedman e é considerado conservador, reduzindo a probabilidade de descobertas falsas (DEMŠAR, 2006).

3.7. FERRAMENTAS, BIBLIOTECAS E AMBIENTE EXPERIMENTAL

A implementação foi desenvolvida com Python como linguagem principal, aproveitando seu ecossistema de bibliotecas especializadas em aprendizado de máquina e processamento de imagens. Como citado, o YOLOv8 constituiu a base arquitetural principal do sistema, sendo esta escolha motivada pela validação da comunidade científica e capacidade de integração com *pipelines* de treinamento automatizado. Além da arquitetura de rede neural, o YOLO oferece ferramentas completas para treinamento, validação e inferência, reduzindo significativamente a complexidade de implementação.

A biblioteca OpenCV foi utilizada para processamento de imagens e manipulação de vídeos, o que inclui operações de carregamento, redimensionamento e salvamento de *frames*. O PyTorch funciona como a base tecnológica essencial para experimentos com redes neurais profundas, oferecendo ferramentas otimizadas para cálculos com tensores e com gradientes necessários para o treinamento dos modelos. A integração nativa com CUDA permite aproveitamento eficiente de recursos de GPU quando disponíveis, acelerando significativamente os processos de treinamento e inferência. A biblioteca NumPy foi empregada para operações matemáticas e manipulação de *arrays* multidimensionais, a qual fornece base computacional eficiente para processamento de dados numéricos. O sistema de configuração baseado em PyYAML permite parametrização flexível de todos os aspectos experimentais, desde caminhos de diretórios até hiperparâmetros de treinamento. Para visualização de resultados e análise de métricas, foram utilizadas as bibliotecas Matplotlib e Seaborn. Estas ferramentas são fundamentais para interpretação dos resultados e comunicação efetiva dos achados experimentais. O ambiente computacional foi configurado com detecção automática de recursos de *hardware*, com o uso de GPU CUDA quando disponível e *fallback* para CPU se necessário. Esta abordagem adaptativa garante funcionalidade em diferentes configurações de *hardware*, desde estações de trabalho mais avançadas até laptops convencionais.

A implementação seguiu arquitetura modular com separação clara de responsabilidades, facilitando manutenção e extensão futura do código. Funções específicas foram desenvolvidas para cada etapa do fluxo experimental, desde

preparação de dados até avaliação de resultados, garantindo clareza metodológica e facilidade de *debugging*. Para garantir reprodutibilidade, foram implementados mecanismos de controle de *seeds* aleatórias em todas as bibliotecas utilizadas, garantindo que experimentos repetidos produzam resultados idênticos. Esta prática é fundamental para validação de resultados e comparação objetiva entre diferentes configurações experimentais.

O sistema de monitoramento experimental inclui registro contínuo de métricas durante o treinamento, salvamento automático de *checkpoints* do modelo, e geração de relatórios detalhados de performance. Estas funcionalidades facilitam análise posterior dos experimentos e identificação de padrões na evolução das métricas de performance.

4 RESULTADOS E DISCUSSÃO

4.1. DESEMPENHO GERAL DOS MODELOS COM DIFERENTES PROPORÇÕES DE ANOTAÇÃO

4.1.1 Estatísticas Descritivas

Os experimentos conduzidos sugerem que a abordagem semi supervisionada proposta é capaz de manter performance elevada mesmo com redução significativa na quantidade de dados manualmente anotados, o que corrobora o reportado por Sohn *et al*. (2021), que demonstraram melhorias substanciais com uso limitado de dados rotulados. A validação cruzada com k=10 repetições independentes forneceu base estatística para comparação entre as quatro abordagens experimentais. O Tabela 7 apresenta as estatísticas descritivas das métricas de performance para cada cenário de anotação, revelando padrões consistentes de hierarquia entre as abordagens.

Tabela 7 - Estatísticas descritivas do *cross-validation*.

Métrica	10%	20%	30%	100%
Precision	0.949±0.015	0.962±0.011	0.964±0.008	0.967±0.007
Recall	0.948±0.012	0.960±0.015	0.973±0.010	0.975±0.009
mAP@0,5	0.971±0.010	0.979±0.007	0.985±0.005	0.986±0.003
mAP@0,5:0,95	0.767±0.016	0.771±0.015	0.801±0.012	0.805±0.014

Fonte: O autor, 2025.

A análise descritiva inicial sugere uma melhora de performance, com a abordagem supervisionada (100%) apresentando valores médios superiores em todas as métricas. Além disso, o desvio padrão diminui conforme a redução da quantidade de dados anotados.

4.1.2 Teste de Friedman

Para avaliar estatisticamente se as diferenças observadas são significativas, foi aplicado o teste não-paramétrico de Friedman a cada métrica. O Tabela 8 sumariza os resultados da análise global.

Tabela 8 - Resultados do Teste de Friedman para as métricas.

Métrica	**Rank Médio**				**Estatística**
	10%	**20%**	**30%**	**100%**	**χ^2(df=2)**
Precision	3.80	2.80	1.90	1.50	19.140
Recall	3.90	3.10	1.60	1.40	26.460
mAP@0,5	3.80	3.00	1.70	1.50	22.680
mAP@0,5:0,95	3.60	2.90	1.80	1.70	18.420

Fonte: O autor, 2025.

Os resultados do teste de Friedman rejeitam a hipótese nula de igualdade entre as abordagens para todas as métricas avaliadas ($p < 0.05$). As estatísticas χ^2 observadas (18.420-26.460) superam os valores críticos tanto para $\alpha = 0.05$ quanto para $\alpha = 0.01$, indicando evidência robusta de diferenças entre as abordagens.

A análise dos *ranks* médios revela hierarquia consistente: as abordagens com 100% e 30% obtém sistematicamente os melhores *ranks* (1.50-1.90), seguida pela semi supervisionada com 20% (2.80-3.10) e, por último, a semi supervisionada com 10% (3.60-3.90).

4.1.3 Teste de Nemenyi

Dado que o teste de Friedman indicou diferenças significativas em todas as métricas, procedeu-se à análise post-hoc através do teste de Nemenyi para identificar quais pares de abordagens diferem significativamente, com os resultados mostrados no Tabela 9. A diferença crítica (CD) para $\alpha = 0.05$ com $k = 4$ tratamentos e $N = 10$ blocos é maior ou igual a 1.213.

Tabela 9 - Resultados do Teste de Nemenyi para cada comparação par a par.

Métrica	100% x 30%	100% x 20%	100% x 10%	30% x 20%	30% x 10%	20% x 10%
Precision	0.4	1.3	2.3	0.9	1.9	1.0
Recall	0.2	1.7	2.5	1.5	2.3	0.8
mAP@0,5	0.2	1.5	2.3	1.3	2.1	0.8
mAP@0,5:0,95	0.1	1.2	1.9	1.1	1.8	0.7

Fonte: O autor, 2025.

A análise *post-hoc* revela padrões específicos de significância:

1. **Comparação do modelo supervisionado**: A abordagem com 100% dos dados é estatisticamente superior à abordagem com 10% em todas as métricas (4/4 comparações significativas), à abordagem com 20% em 3/4 métricas (exceto mAP@0.5:0.95). Em contrapartida, não possui diferença estatisticamente significativa em relação à abordagem com 30%.
2. **Diferenças entre abordagens semi supervisionadas**: O modelo com 30% é estaticamente superior ao modelo com 10% em todas as métricas, e em 2/4 métricas (*Recall e mAP@0,5*) em relação ao de 20%. Não foram detectadas diferenças significativas entre os modelos com 20% e 10% de dados anotados em nenhuma das métricas avaliadas.
3. **Diferenças das métricas *Recall e mAP@0,5***: Essas métricas demonstraram maiores diferenças entre abordagens, sendo a única na qual em todas as comparações com a abordagem supervisionada, exceto 100% x 30%, resultaram em resultados estatisticamente significativos.

4.1.4 Análise da Magnitude dos Efeitos e Discussão

Embora as diferenças sejam estatisticamente significativas, é importante avaliar sua relevância prática. O Tabela 10 quantifica as magnitudes das diferenças em termos das métricas originais em pontos percentuais.

Tabela 10 - Subtrações entre as médias das métricas após o *cross-validation*, em pontos percentuais.

Métrica	100% x 30%	100% x 20%	100% x 10%	30% x 20%	30% x 10%
Precision	0.3%	0.5%	1.9%	0.7%	2.1%
Recall	0.2%	1.5%	2.7%	1.3%	2.5%
mAP@0,5	0.1%	0.7%	1.5%	0.6%	1.4%
mAP@0,5:0,95	0.4%	3.4%	3.8%	3.0%	3.4%

Fonte: O autor, 2025.

A análise das magnitudes revela que, embora geralmente estatisticamente significativas, as diferenças absolutas são relativamente modestas para a maioria das métricas. A maior perda ocorre na métrica mAP@0.5:0.95, onde a redução de 100% para 10% de dados anotados resulta em deterioração de 3.8%. Este resultado alinha-se com a literatura que demonstra maior sensibilidade de métricas rigorosas de localização a reduções na qualidade dos dados de treinamento (PADILLA; NETTO; DA SILVA, 2020). Ao mesmo tempo, esta diferença pode fazer pouco efeito na prática e não ser custo-benéfica em relação ao esforço de anotação empregado para alcançar essa performance. Isto é notado na comparação entre 100% e 30%, na qual a métrica com maior diferença de resultado tem resultado apenas 0,4% menor.

Os resultados obtidos oferecem validação de padrões reportados na literatura. Xu *et al*. (2024) demonstraram que *frameworks* semi supervisionados mantêm performance competitiva com modelos pré treinados em condições de escassez de dados anotados, achado que encontra correspondência direta nos resultados nos quais a abordagem de 30% mantém 99.5% da performance na métrica mais rigorosa (mAP@0.5:0.95: 0.801 vs 0.805).

Além disso, as métricas alcançadas corroboram com a discussão de Jin *et al*. (2021) sobre a existência de pontos ótimos na quantidade de dados anotados para maximizar qualidade de *pseudo-labels*. Enquanto a diferença 100% vs 30% não é significativa, a diferença 100% vs 10% é significativa em todas as métricas, sugerindo que existe um limiar crítico entre 10% e 30% na qual a degradação se acelera.

Apesar dessas diferenças, as abordagens semi supervisionadas mantém diferenças controladas: *precision* (-1.9%), *recall* (-2.8%), mAP@0.5 (-1.5%) e

mAP@0.5:0.95 (-3.8%) para o cenário de 10%, o pior caso no quesito métricas. Estas magnitudes são piores que as reportadas em trabalhos anteriores, porém o *dataset* utilizado é menor em números absolutos.

4.2. ANÁLISE QUANTITATIVA DAS CLASSES

4.2.1 Análise dos mAPs

As análises dos mAPs por classe revelam comportamentos que refletem a complexidade inerente à detecção de diferentes tipos de EPIs, alinhando-se com os desafios reportados por Gallo *et al*. (2022). A heterogeneidade na performance entre classes é um fenômeno bem documentado na literatura de detecção de objetos (FANG *et al*., 2017), particularmente em domínios especializados com um *background* complexo envolvido, como o caso das indústrias.

A classe *Vest* apresentou a melhor performance em todos os cenários experimentais, com mAP@0.5 evoluindo de 0.985 (10%) para 0.988 (20%), 0,993 (30%) até 0.994 (100%), demonstrando excelente detectabilidade. Todos os resultados são vistos na Tabela 11 abaixo.

Tabela 11 - Valores de mAP@0,5 por classes.

Classes	mAP@0,5 - 10%	mAP@0,5 - 20%	mAP@0,5 - 30%	mAP@0,5 - 100%
Helmet	0,975	0,982	0,990	0,991
Vest	0,985	0,988	0,993	0,994
Person	0,962	0,976	0,979	0,981
Ear	0,962	0,970	0,977	0,978

Fonte: O autor, 2025.

Este resultado pode ser atribuído às características visuais distintivas dos coletes de segurança, que apresentam cores contrastantes e padrões reflexivos que facilitam a detecção automática. Esta observação é consistente com os princípios de design de EPIs, que intencionalmente maximizam a visibilidade através de cores de alta saturação e materiais refletivos para serem processados mais eficientemente

por sistemas de visão (TREISMAN; GELADE, 1980). A aplicação desta teoria ao contexto de redes neurais convolucionais sugere que características de baixo nível como cores e bordas associadas aos coletes são mais facilmente aprendidas e generalizadas.

A classe *Helmet* demonstrou ótima performance, com variação mínima entre os diferentes cenários. Os modelos variaram de 0,975 (10%) até 0,991 (100%) no mAP@0.5. A consistência na detecção de capacetes pode ser atribuída à sua forma geométrica distintiva e posicionamento padronizado na região superior da cabeça.

A classe *Person* apresentou mAP@0.5 de 0.962 (10%), 0.976 (20%), 0.979 (30%) e 0.981 (100%), mostrando progressão consistente, porém baixa, com o aumento de dados anotados. Esta classe representa um desafio fundamental em ambientes industriais devido à variabilidade de poses, oclusões parciais, condições de iluminação variáveis, a presença de equipamentos industriais que podem causar confusão visual e sua importância no contexto de detecção de segurança. A localização das *bounding boxes* dessa classe demonstrou bons resultados nos cenários testados, com valores de mAP@0.5:0.95, visualizados no Tabela 12, entre 0.765-0.815. Este padrão é consistente com a arquitetura YOLOv8 utilizada, que foi otimizada para detecção de pessoas em diversos contextos.

Tabela 12 - Valores de mAP@0,5:0,95 por classes.

Classes	mAP@0,5:0,95 - 10%	mAP@0,5:0,95 - 20%	mAP@0,5:0,95 - 30%	mAP@0,5:0,95 - 100%
Helmet	0,785	0,789	0,820	0,825
Vest	0,815	0,821	0,842	0,845
Person	0,765	0,771	0,809	0,815
Ear	0,703	0,703	0,735	0,735

Fonte: O autor, 2025.

A classe *Ear* apresentou a maior variabilidade entre os cenários. Com mAP@0.5 variando entre 0.962-0.978 e mAP@0.5:0.95 entre 0.703-0.735, esta classe representa o maior desafio técnico do sistema proposto. A menor performance na métrica mais rigorosa indica dificuldades específicas na localização precisa destes objetos, provavelmente devido ao seu tamanho reduzido e alta susceptibilidade a oclusões parciais. Esta dificuldade é consistente com os princípios

de detecção de objetos pequenos em redes neurais convolucionais, nas quais a perda de informação espacial pode impactar significativamente a detecção de objetos que ocupam pequenas porções da imagem (SINGH; DAVIS, 2018). Além disso, protetores auriculares tipo concha apresentam características visuais menos distintivas comparados a outros EPIs, frequentemente confundindo-se com cabelo escuro ou sombras. Entretanto, Pathiraja, Gunawardhana e Khan (2023) propõem uma nova abordagem de que visa calibrar conjuntamente a confiança multiclasse preditiva e a localização da *bounding box*, o que pode indicar um caminho para otimizar o caso dessa classe.

4.2.2 Matrizes de Confusão

A análise das matrizes de confusão normalizadas, apresentadas nas Tabelas 13, 14, 15 e 16 abaixo, revela padrões de performance que complementam e questionam o entendimento dos resultados de mAP apresentados anteriormente. Enquanto as métricas gerais de mAP indicaram performance comparável entre as abordagens semi supervisionadas e a supervisionada, as matrizes de confusão expõem nuances no comportamento classificatório dos modelos em questão.

Tabela 13 - Matriz de confusão normalizada para o modelo com 10% de dados anotados manualmente.

Helmet	0,97	-	-	-	0,22
Vest	-	0,99	-	-	0,06
Person	-	-	0,92	-	0,54
Ear	-	-	-	0.96	0,18
Background	0,03	0,01	0,08	0,04	-
	Helmet	*Vest*	*Person*	*Ear*	*Background*

Fonte: O autor, 2025.

Tabela 14 - Matriz de confusão normalizada para o modelo com 20% de dados anotados manualmente.

Helmet	0,98	-	-	-	0,33
Vest	-	0,99	-	-	0,07
Person	-	-	0,94	-	0,48
Ear	-	-	-	0.97	0,12
Background	0,02	0,01	0,06	0,03	-
	Helmet	*Vest*	*Person*	*Ear*	*Background*

Fonte: O autor, 2025.

Tabela 15 - Matriz de confusão normalizada para o modelo com 30% de dados anotados manualmente.

Helmet	0,99	-	-	-	0,19
Vest	-	1,00	-	-	0,12
Person	-	-	0,96	-	0,41
Ear	-	-	-	0.98	0,21
Background	0,01	-	0,04	0,02	-
	Helmet	*Vest*	*Person*	*Ear*	*Background*

Fonte: O autor, 2025.

Tabela 16 - Matriz de confusão normalizada para o modelo com todos os dados anotados manualmente.

Helmet	1,00	-	-	-	0,17
Vest	-	1,00	-	-	0,13
Person	-	-	0,96	-	0,39
Ear	-	-	-	0.98	0,31
Background	-	-	0,04	0,02	-
	Helmet	*Vest*	*Person*	*Ear*	*Background*

Fonte: O autor, 2025.

A classe *Helmet* demonstra consistência em todos os cenários experimentais, com ótima performance (acima de 0,97) tanto nos modelos semi supervisionados quanto no supervisionado. Este resultado corrobora os achados de mAP@0.5

previamente reportados (0.975 para 10%, 0.982 para 20%, 0,990 para 30% e 0.991 para 100%) e alinha-se com a literatura que argumentam que redes neurais convolucionais desenvolvem hierarquias de características, nas quais características simples (bordas, contornos) são progressivamente combinadas em representações mais complexas. Para capacetes, este processamento hierárquico pode ser particularmente eficiente porque a forma distintiva pode criar padrões de bordas e gradientes que são facilmente detectáveis nas camadas iniciais e consistentemente agregados nas camadas superiores (LECUN; BENGIO; HINTON; 2015). Entretanto, a análise da confusão com o *background* revela um padrão diferente de evolução com o aumento de dados anotados. O modelo com 10% apresenta taxa de falsos positivos de 0,22 para capacetes classificados incorretamente como *background*, aumentando para 0,33 no modelo de 20%, e reduzindo para 0,19 em 30% e 0,17 no 100% supervisionado. Bartlett e Mendelson (2002) sugerem que a relação entre quantidade de dados de treinamento e performance de generalização não é necessariamente diretamente proporcional, especialmente em regimes de dados limitados nos quais diferentes mecanismos de generalização podem dominar. Além disso, o pequeno *dataset* pode levar a taxas ruins de *signal to noise* e afetar a capacidade de generalização de um modelo (ADVANI; SAXE, 2017), principalmente sendo ele multi-classe.

A classe *Vest* mantém ótima performance (acima de 0,99) em todos os cenários, confirmando as observações qualitativas sobre as características visuais altamente contrastantes dos coletes de segurança e a citada propriedade de características distintivas desses objetos. A taxa de falsos positivos para *background* mantém-se consistentemente baixa (0,06-0,13), indicando que, neste caso, os coletes não são confundidos com elementos do ambiente industrial. A robustez da detecção de coletes em diferentes proporções de dados anotados sugere a hipótese de que esta classe poderia potencialmente servir como referência confiável em sistemas multi-classe. Esta possibilidade, embora não diretamente validada na literatura existente, merece investigação futura para determinar se detecções como esta podem ser sistematicamente utilizadas para calibração de outros componentes do sistema.

A classe *Person* apresenta o padrão mais complexo, com sua métrica variando entre 0,92-0,96 nos diferentes cenários. O aprendizado supervisionado

alcança performance superior (0,96) comparado aos modelos semi supervisionados de 10% e 20% (0,92 para 10% e 0,94 para 20%), mas com o mesmo valor para 30% (0,96). Porém, a taxa de confusão com background apresenta padrão elevado (0,54 para 10%, 0,48 para 20%, 0,41 para 30% e 0,39 para 100%) em todos os casos, o que contrasta com a estabilidade mostrada pela classe no mAP. O primeiro fator explicativo reside no desequilíbrio inerente entre as classes no *dataset* estudado. Conforme reportado no histograma do *dataset* original, a classe *Person* apresenta 50.088 instâncias, significativamente superior às demais classes (*Helmet*: 22.147, *Vest*: 20.461, *Ear*: 19.312). Este desequilíbrio cria um viés estatístico fundamental que pode afetar diferentemente as métricas de precisão e as taxas de confusão. Shahnawaz e Kumar (2025) reportam que em *datasets* desequilibrados, modelos tendem a desenvolver viés em direção à classe majoritária, resultando em alta precisão para essa classe devido ao grande número de verdadeiros positivos, mesmo na presença de falsos positivos substanciais. Zeiler e Fergus (2013) demonstraram como características aprendidas pelos modelos podem ser altamente sensíveis por estruturas com padrões visuais similares. Equipamentos, estruturas e sombras podem ativar detectores de características humanas, resultando em falsos positivos que contribuem para a alta taxa de confusão com *background*.

A classe *Ear* varia pouco entre cenários (0,96 para 10%, 0,97 para 20%, 0,98 para 30% e 0,98 para 100%) e sua taxa de confusão com o background segue uma tendência diferente (0,18 para 10%, 0,12 para 20%, 0,21 para 30% e 0,31 para 100%). Porém, o aumento substancial de falsos positivos com o aumento de dados anotados pode sugerir um possível *overfitting* aos padrões específicos do *dataset* de treinamento, fenômeno documentado em cenários nos quais a complexidade do modelo excede a diversidade dos dados disponíveis (YING, 2019).

4.2.3 Comparação com Abordagem Supervisionada

A comparação revela que a abordagem semi supervisionada proposta consegue se aproximar das métricas de performance do modelo treinado com 100% dos dados anotados. Este resultado é um indicativo importante, considerando a redução na quantidade de anotações manuais requeridas. Os resultados obtidos são comparáveis aos reportados por Sohn *et al*. (2021) em seu *framework* STAC, no qual

demonstraram que modelos semi supervisionados podem até superar estratégias supervisionadas quando treinados com 5% de dados anotados manualmente, por exemplo. Similarmente, o *framework* CISO proposto por Qi *et al*. (2023) demonstrou performance superior às abordagens supervisionadas em *datasets* como MS COCO e VOC com proporções similares de dados anotados.

Na métrica mAP@0.5, os modelos semi supervisionados praticamente equiparam-se ao supervisionado, com diferenças inferiores a 2%. Este resultado é particularmente significativo quando contextualizado dentro da literatura de detecção de EPIs, na qual Gallo *et al*. (2022) reportaram o uso de 70.000 *frames* anotados manualmente para atingir performance comparável em uma abordagem completamente supervisionada. A métrica mAP@0.5:0.95, mais exigente em termos de precisão de localização, mostrou bons resultados e comportamento similar.

Estes resultados contraintuitivos, nos quais modelos treinados com menos dados anotados tem performance próxima a abordagem supervisionada, têm precedentes na literatura e podem ser explicados por certos mecanismos teóricos.

Primeiro, o processo de *pseudo-labeling* pode ter funcionado como uma forma de ruído positivo, similar aos efeitos observados por Xie *et al*. (2020) em seu trabalho sobre *self-training*. A diversidade adicional introduzida pelos *pseudo-labels* pode ter otimizado a capacidade de generalização dos modelos, especialmente considerando o tamanho relativamente pequeno do *dataset* original.

Depois, a seleção estratégica dos vídeos para anotação manual, baseada na preservação de distribuições de classes, pode ter capturado exemplos particularmente informativos do universo amostral. Esta hipótese é suportada pela teoria do *active learning*, que sugere que seleção inteligente de exemplos de treinamento pode resultar em performance superior comparada à amostragem aleatória (SETTLES, 2009).

4.3. ANÁLISE VISUAL

A análise visual dos *frames* de detecção revelou padrões específicos de erro que fornecem informações fundamentais sobre as limitações e potencialidades da abordagem proposta. Esta análise identificou que as classes *Ear* e *Person* apresentam mais erros de detecção, o que corrobora os resultados quantitativos e

os padrões teóricos esperados para estas categorias de objetos. Além disso, foi observado que o modelo de 30% obteve exatamente a mesma localização de detecções em relação ao de 100% nas Figuras abaixo, o que reforça a semelhança dos resultados exibidos anteriormente.

Para a classe *Ear*, foram identificados múltiplos casos nos quais protetores auriculares anotados não foram detectados pelos modelos, particularmente quando há interferência das mãos dos trabalhadores, o que converge com a hipótese de oclusões levantada no tópico anterior. Isto pode ser visto na Figura 5.

Figura 5 - Mão de trabalhador se torna oclusão e impede os modelos de 10% (azul) e 20% (vermelho) de realizarem a detecção.

Fonte: o autor, 2025.

Recorrentemente, observa-se a oclusão por membros superiores, o que constitui um fator limitante crítico, fenômeno bem documentado na literatura de detecção de objetos pequenos (SINGH; DAVIS, 2018). Esta limitação específica pode ser explicada através do fato de que objetos parcialmente obstruídos requerem inferência contextual que pode exceder a capacidade de generalização dos modelos treinados, especialmente com dados limitados. A presença desses falsos negativos também pode sugerir que os *pseudo-labels* gerados podem ter introduzido ruído sistemático que foi posteriormente amplificado durante o treinamento do Modelo 2.

A classe *Person* apresentou padrões de erro relacionados principalmente a oclusões e sobreposições entre trabalhadores, problema fundamental em detecção de pessoas em ambientes povoados (DOLLÁR *et al*., 2011), observado na Figura 6.

Figura 6 - Trabalhador se torna oclusão de outro e impede que dois dos modelos (10% e 20%) de realizarem a detecção.

Fonte: o autor, 2025.

A análise qualitativa também identificou casos nos quais pessoas foram anotadas mas não detectadas quando posicionadas atrás de outros indivíduos ou quando localizadas em segundo plano, evidenciado na Figura 7.

Figura 7 - Trabalhador com o celular foi anotado (verde), mas não é detectado em modelo algum.

Fonte: O autor, 2025.

Estes padrões são particularmente relevantes para aplicações como esta, nas quais a detecção de todos os trabalhadores presentes na cena é fundamental para análise contextual do uso de EPIs, e um erro pode resultar em falhas críticas na avaliação de conformidade com normas de segurança. A dificuldade em detectar pessoas parcialmente oclusas ou em segundo plano reflete limitações conhecidas das arquiteturas YOLO, que podem ter maior facilidade com objetos mais proeminentes no *frame* (HU *et al*., 2023). De mesmo modo foram evidenciados casos de detecções corretas de pessoas não originalmente anotadas, como na Figura 8, sugerindo que os modelos podem desenvolver capacidade de generalização superior à cobertura das anotações manuais originais.

Figura 8 - Trabalhador não anotado (verde) é detectado nos modelos de 20% (rosa), 30% e 100% (vermelho).

Fonte: o autor, 2025.

A capacidade dos modelos de identificar instâncias válidas não anotadas pode indicar que o processo de *pseudo-labeling* pode ter sido capaz de enriquecer o *dataset* além das limitações das anotações manuais iniciais.

4.4. DISCUSSÃO SOBRE A ECONOMIA DE TEMPO DE ANOTAÇÃO

A implementação da abordagem semi supervisionada proposta resulta em uma economia substancial de recursos humanos dedicados à anotação de dados, de acordo com as motivações fundamentais que impulsionam o desenvolvimento de métodos semi supervisionados nos trabalhos sobre visão computacional. O cenário com 10% de anotação manual requereu 66 dias a menos (77 x 11), enquanto o cenário de 20% reduziu em 54 dias o processo (77 x 23) e o cenário de 30% reduziu em 45 dias as anotações (77 x 32). Esta economia de tempo tem implicações econômicas significativas, especialmente considerando que anotação de dados especializados requer expertise técnica e um custo associado que pode ser elevado, a depender da complexidade (SNOW *et al.*, 2008). A comparação com custos reportados na literatura reforça a relevância econômica dos resultados obtidos. Lin

et al. (2014) reportaram que a criação do *dataset* MS COCO requereu cerca de 60.000 horas humana. A redução no esforço de anotação, mantendo performance comparável, representa avanço significativo na viabilidade prática de sistemas de detecção automatizada para aplicações industriais. A análise econômica deve também considerar os custos de oportunidade associados ao tempo de especialistas. Em contextos industriais, profissionais qualificados em segurança do trabalho frequentemente possuem outras responsabilidades críticas, tornando o tempo dedicado à anotação de dados uma limitação prática significativa. A redução de 77 para 11, 23 e 32 dias de anotação representa liberação substancial de recursos humanos especializados para outras atividades de valor agregado, além da diminuição da fadiga e possível aumento da motivação no trabalho.

A escalabilidade da abordagem proposta constitui uma vantagem adicional com implicações práticas importantes. Uma vez estabelecida o fluxo do experimento semi supervisionado, novos dados podem ser incorporados com mínimo esforço de anotação manual adicional, permitindo adaptação contínua e melhoria iterativa dos modelos de detecção. Esta característica é especialmente relevante para aplicações industriais, nas quais condições operacionais, equipamentos, *layouts* e procedimentos podem evoluir ao longo do tempo, requerendo adaptação dos sistemas de monitoramento.

4.5. VALIDAÇÃO EM *DATASET* PÚBLICO

Para avaliar a capacidade de generalização da abordagem proposta, além do contexto industrial específico no qual os modelos foram desenvolvidos, foi conduzida validação adicional utilizando um *dataset* público para detecção de EPIs. Importante ressaltar que para esta validação, a metodologia completa foi reaplicada, com os seguintes passos: primeiro, seleção de proporções de dados para anotação manual (10%, 20%, 30%); segundo, treinamento do Modelo 1 com dados parcialmente anotados; terceiro, geração de *pseudo-labels* para dados não anotados; e por fim, treinamento do Modelo 2 com *pseudo-labels* gerados. Portanto, os modelos avaliados no *dataset* público foram treinados especificamente nesse *dataset* utilizando a mesma estratégia metodológica, não representando transferência direta dos modelos desenvolvidos no *dataset* industrial coletado pelo autor deste trabalho.

4.5.1 Contexto do SH17 Dataset

O SH17 Dataset[1] representa um dos *datasets* mais abrangentes disponíveis publicamente para detecção de EPIs em contextos industriais. Composto por 8.099 imagens anotadas contendo 75.994 instâncias distribuídas em 17 classes de equipamentos de proteção individual, o *dataset* foi coletado em diversos ambientes industriais reais, incluindo construção civil e ambientes de produção. Uma característica distintiva deste *dataset* é a prevalência de objetos pequenos: 52% das anotações ocupam menos de 1% da área total da imagem, e 78% ocupam menos de 5%.

As 17 classes incluem não apenas EPIs propriamente ditos (*helmet, vest, mask, safety glasses, gloves, safety shoes, earmuff, earplug*), mas também suas correspondentes classes negativas (*no-helmet, no-vest, no-mask, no-safety glasses, no-gloves, no-safety shoes*), além das classes *person*, *head* e *body*.

4.5.2 Resultados

Para o presente estudo, a avaliação focou nas classes que apresentam correspondência conceitual com o *dataset* industrial original deste trabalho: *helmet*, *vest*, *person* e *earmuffs*. É importante ressaltar que os resultados dos autores do *dataset* contemplam todas as 17 classes (AHMAD; RAHIMI, 2024), estabelecendo referência de performance para os resultados deste trabalho.

A Tabela 17 mostra os resultados alcançados em cada métrica para cada percentual anotado, e a última coluna traz os resultados dos autores originais para a mesma arquitetura utilizada.

[1] Disponível em: https://www.kaggle.com/datasets/mugheesahmad/sh17-dataset-for-ppe-detection. Acesso em: 19 set. 2025

Tabela 17 - Estatísticas descritivas dos resultados no SH17.

Métrica	10%	20%	30%	100%	Original
Precision	0.673±0.031	0.712±0.026	0.748±0.021	0.781±0.018	0,815
Recall	0.414±0.035	0.458±0.029	0.503±0.024	0.538±0.020	0,557
mAP@0,5	0.492±0.029	0.548±0.024	0.571±0.019	0.609±0.016	0,637
mAP@0,5:0,95	0.323±0.036	0.379±0.031	0.398±0.026	0.405±0.022	0,417

Fonte: O autor, 2025.

A comparação revela uma performance equivalente aos resultados reportados por Ahmad e Rahimi (2024). A métrica mAP@0.5:0.95 para o modelo com 100% de anotação tem o melhor resultado, atingindo 0.405, ficando apenas 1.2% abaixo do reportado pelos autores originais. A diminuição generalizada pode ser explicada por múltiplos fatores que caracterizam a falta de metodologia pensada especificamente para o contexto, uma vez que as imagens já anotadas foram incorporadas ao processo.

Primeiro, o contexto operacional fundamentalmente distinto entre os ambientes afeta profundamente a detecção. Enquanto o *dataset* original foi coletado em ambiente industrial controlado de manufatura com ângulos de câmera fixos, distâncias padronizadas e iluminação homogênea, o SH17 apresenta diversidade extrema: imagens de múltiplos setores industriais, variações de ângulo de câmera, distâncias diferentes e condições de iluminação heterogêneas. Por último, o desafio de objetos pequenos no SH17 (52% das anotações < 1% da área) torna difícil a generalização para aprendizados semi-supervisionados.

4.6. LIMITAÇÕES E CONSIDERAÇÕES METODOLÓGICAS

Apesar dos resultados promissores, diversas limitações metodológicas devem ser cuidadosamente consideradas na interpretação dos achados apresentados. A identificação dessas é fundamental para contextualizar adequadamente as contribuições e orientar desenvolvimentos futuros.

A primeira limitação significativa refere-se ao tamanho e especificidade da base de dados utilizada. Com 50.088 imagens derivadas de um único ambiente

industrial, o *dataset* é relativamente pequeno comparado aos padrões estabelecidos na literatura de detecção de objetos, na qual *datasets* como MS COCO contêm cerca de 328.000 imagens (LIN *et al*., 2014). Isto impactou a generalização dos resultados para outros contextos industriais com características visuais, condições de iluminação e/ou tipos de equipamentos diferentes, como visto na seção 4.6. A especificidade do ambiente de coleta também representa uma consideração importante. A base foi coletada em uma única instalação industrial no interior de São Paulo, com características, equipamentos e procedimentos específicos. A generalização para outras indústrias, regiões geográficas ou tipos de EPIs sempre irá requerer validação adicional, como feita na seção 4.5.

Outro tópico a ser discutido é a dependência crítica da qualidade dos *pseudo-labels* gerados pelo Modelo 1. Erros sistemáticos introduzidos durante esta etapa podem propagar-se e amplificar-se durante o treinamento do Modelo 2 (OLIVER *et al*., 2019) e levar a vieses de confirmação (ARAZO *et al*., 2020). A inspeção visual manual implementada como controle de qualidade, embora forneça algum nível de verificação, possui limitações inerentes. Esta abordagem não é escalável para *datasets* maiores e pode não capturar os padrões de erro relevantes. Métodos automatizados de avaliação de qualidade de *pseudo-labels* podem fornecer controle de qualidade mais rigoroso e escalável.

Finalmente, a análise qualitativa revelou que certas condições operacionais como oclusões e posicionamento de trabalhadores continuam representando desafios significativos para todos os modelos testados. Estas limitações são inerentes à complexidade do domínio de aplicações reais e refletem desafios fundamentais em visão computacional que podem requerer estratégias complementares para serem adequadamente endereçadas. As oclusões, em particular, representam um desafio técnico que pode requerer abordagens arquiteturais mais sofisticadas, como redes neurais com mecanismos de atenção (GUO; XU; LIU, 2022).

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1. PRINCIPAIS CONCLUSÕES DA PESQUISA

Este trabalho teve como objetivo central avaliar o método de aprendizado mais eficaz para rotulação de imagens em detecção de Equipamentos de Proteção Individual, respondendo à questão: "Qual a abordagem mais custo-benéfica para o caso de uso apresentado?". Com base na investigação experimental conduzida, conclui-se que há indícios que a abordagem semi supervisionada constitui uma alternativa viável e economicamente vantajosa ao aprendizado totalmente supervisionado para o domínio específico de detecção de EPIs no ambiente industrial. Esta constatação pode representar uma mudança na forma como sistemas de detecção automatizada podem ser desenvolvidos e implementados nestes contextos, nos quais tradicionalmente o gargalo reside na disponibilidade de dados especializados, na anotação por profissionais qualificados, no tempo dedicado, na fadiga, e na falta de motivação para a execução de uma tarefa repetitiva. A redução substancial do tempo necessário para anotação manual libera profissionais de segurança do trabalho para atividades de maior valor agregado, como desenvolvimento de políticas de segurança, treinamento de funcionários e análise estratégica de riscos ocupacionais.

O presente estudo sugere que diferentes classes de EPIs apresentam comportamentos distintos no processo de aprendizado semi supervisionado, acrescentando que estratégias de treinamento personalizadas por categoria podem otimizar ainda mais os resultados. Isto indica que a abordagem utilizada pode não ser ideal para todos os tipos de equipamentos de proteção, indicando um caminho para metodologias que trate cada objeto com a especialidade para o contexto de cada caso de uso.

5.2. IMPLICAÇÕES PRÁTICAS E ACADÊMICAS

O trabalho promove à discussão um precedente metodológico para implementação de sistemas de visão computacional em ambientes com recursos

limitados para anotação de dados, como foi aplicado na indústria paulista. A demonstração feita pode acelerar a adoção de tecnologias de monitoramento automatizado em empresas, que podem enfrentar barreiras de entrada devido aos custos de desenvolvimento de *datasets* especializados, facilitando a democratização do acesso a tecnologias avançadas de segurança ocupacional.

Do ponto de vista acadêmico, este trabalho contribui para o crescente corpo de literatura sobre aplicações práticas de aprendizado semi supervisionado em visão computacional. A pesquisa fornece indícios de que princípios teóricos bem estabelecidos podem ser efetivamente traduzidos para aplicações, preenchendo lacuna importante entre teoria e prática no campo de aprendizado de máquina aplicado, com a constituição de uma base de dados real.

Esta investigação também destaca a importância de considerar fatores econômicos e operacionais no projeto de sistemas de aprendizado de máquina. Frequentemente, pesquisas acadêmicas focam exclusivamente em métricas de performance técnica, negligenciando considerações práticas que determinam a viabilidade de implementação. É importante notar que a otimização conjunta de performance técnica e eficiência econômica pode levar a soluções mais sustentáveis e amplamente aplicáveis.

5.3. SUGESTÕES DE TRABALHOS FUTUROS

5.3.1 Aplicação em outros ambientes industriais

A generalização da metodologia proposta para diferentes setores industriais representa uma direção natural de expansão desta pesquisa. Cada ambiente apresenta desafios únicos relacionados a condições de iluminação, densidade populacional e protocolos específicos de segurança. Investigações futuras deveriam explorar a possibilidade do uso de *transfer learning* dos modelos desenvolvidos entre diferentes contextos industriais, avaliando tanto a robustez da abordagem quanto a necessidade de adaptações específicas por setor.

A aplicação em ambientes de alta complexidade visual, como plataformas petrolíferas, minas subterrâneas ou plantas químicas, nos quais condições extremas

de operação podem desafiar sistemas convencionais de visão computacional pode se tornar uma pesquisa enriquecedora. Tais ambientes frequentemente apresentam limitações adicionais de conectividade e processamento que poderiam beneficiar-se de abordagens que maximizam o custo-benefício.

Além disso, a criação de uma unidade entre as indústrias para desenvolvimento colaborativo de *datasets* padronizados poderia acelerar significativamente o progresso na área, como acontece no campo da ciber segurança. Compartilhamento de dados anonimizados entre organizações do mesmo setor ou setores relacionados permitiria desenvolvimento de modelos mais robustos e generalizáveis, enquanto distribui custos de desenvolvimento entre múltiplos *stakeholders*.

5.3.2 Uso de técnicas de *active learning*

A integração de estratégias de *active learning* representa evolução natural da abordagem semi supervisionada proposta. Sistemas que identificam automaticamente quais exemplos não rotulados deveriam ser, pelo impacto no contexto, poderiam otimizar ainda mais o processo de desenvolvimento de *datasets*. Esta direção é particularmente relevante considerando que nem todos os exemplos não rotulados contribuem igualmente e podem até prejudicar o aprendizado do modelo, enquanto a rotulação de poucos *frames* adicionais pode trazer um ganho significativo no processo (YOO; KWEON, 2019). Além disso, desenvolvimento de interfaces que facilitem a interação entre algoritmos de *active learning* e equipes especialistas em segurança poderia revolucionar o processo de anotação. Sistemas que apresentam casos de incerteza máxima ou exemplos representativos de regiões inexploradas do espaço de características permitiriam que essas equipes contribuam de forma mais estratégica e eficiente.

REFERÊNCIAS BIBLIOGRÁFICAS

ADVANI, M. S.; SAXE, A. M. High-dimensional dynamics of generalization error in neural networks. Neural Networks, v. 132, p. 428-446, 2017.

AFLALO, Amit *et al*. DeepCut: unsupervised segmentation using graph neural networks clustering. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS (ICCVW), 2023. Proceedings [...]. [S.l.: s.n.], 2023. p. 32-41.

AHMAD, H. M.; RAHIMI, A. SH17: A dataset for human safety and personal protective equipment detection in manufacturing industry. Journal of Safety Science and Resilience, 2024.

ALI, A. *et al*. Transfer learning: a new promising techniques. Mesopotamian Journal of Big Data, v. 2023, p. 31-32, fev. 2023.

ARAZO, E. *et al*. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2020. Proceedings... [S.l.: s.n.], 2020. p. 1-8.

BARTLETT, P. L.; MENDELSON, S. Rademacher and gaussian complexities: risk bounds and structural results. Journal of Machine Learning Research, v. 3, p. 463-482, 2002.

BRASIL. Ministério do Trabalho e Emprego. Norma Regulamentadora nº 6 (NR-6): Equipamento de Proteção Individual - EPI. Brasília: MTE, 2022. Disponível em: https://www.gov.br/trabalho-e-emprego/pt-br/acesso-a-informacao/participacao-social/conselhos-e-orgaos-colegiados/comissao-tripartite-partitaria-permanente/arquivos/normas-regulamentadoras/nr-06-atualizada-2022-1.pdf. Acesso em: 28 ago. 2025.

CHATTERJEE, S. Computer vision: principles, algorithms, applications, learning. Journal of Computer Vision, v. 15, n. 2, p. 78-95, 2022.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, v. 7, p. 1-30, 2006.

DOLLÁR, P. *et al*. Pedestrian detection: an evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 34, n. 4, p. 743-761, 2011.

DOUGHERTY, G. Pattern Recognition and Classification: An Introduction. New York: Springer-Verlag, 196 p., 2013.

ENGELER, C.; HOOS, H. H. A survey on semi-supervised learning. Machine Learning, v. 109, n. 2, p. 373-440, 2019.

EVERINGHAM, M. *et al*. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, v. 88, n. 2, p. 303-338, 2010.

FANG, Y. *et al*. Object detection meets knowledge graphs. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 26., 2017, Melbourne. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne: [s.n.], 2017. p. 1661-1667.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, v. 32, n. 200, p. 675-701, 1937.

GALLO, G. *et al*. A smart system for personal protective equipment detection in industrial environments based on deep learning at the edge. IEEE Access, v. 10, p. 110862-110878, 2022.

GROSSI, Caroline Dias. Desenvolvimento de um software de visão computacional para estudo do escoamento de cascalhos em peneira vibratória. 2020. 168 f. Dissertação (Mestrado em Engenharia Química) - Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2020.

GUO, M.; XU, T.; LIU, J. Attention mechanisms in computer vision: A survey. Computational Visual Media, v. 8, p. 331-368, 2022.

GUO, X.; ZHU, L.; LI, Y. Self-supervised learning for computer vision: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 44, n. 8, p. 4182-4200, 2021.

HOLLANDER, M. *et al*. Nonparametric statistical methods. 3. ed. New York: John Wiley & Sons, 2013.

HU, M. *et al*. Efficient-lightweight YOLO: improving small object detection in YOLO for aerial images. Sensors, v. 23, p. 6423, 2023.

JAISWAL, A. *et al*. A survey on contrastive self-supervised learning. Technologies, v. 9, n. 2, p. 1-22, 2020.

JIN, H. *et al*. Evolutionary optimization based pseudo labeling for semi-supervised soft sensor development of industrial processes. Chemical Engineering Science, v. 237, p. 116560, 2021.

JING, L.; TIAN, Y. Self-supervised visual feature learning with deep neural networks: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 43, n. 11, p. 4037-4058, 2019.

KHANG, A. *et al*. (org.). Computer vision and AI-integrated IoT technologies in the medical ecosystem. Boca Raton: CRC Press, 2024.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., 1995. Proceedings... [S.l.: s.n.], 1995. p. 1137-1143.

KOHAVI, R.; PROVOST, F. Glossary of terms. Machine Learning, v. 30, n. 2-3, p. 271-274, 1998.

KUFEL, J. *et al*. What is machine learning, artificial neural networks and deep learning? Examples of practical applications in medicine. Diagnostics, v. 13, n. 15, p. 2582, 2023.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature, v. 521, n. 7553, p. 436-444, 2015.

LEE, D. H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: WORKSHOP ON CHALLENGES IN REPRESENTATION LEARNING, ICML, 2013. Proceedings... [S.l.: s.n.], 2013. p. 1-6.

LI, Y. *et al*. Improving object detection with selective self-supervised self-training. In: VEDALDI, Andrea *et al*. (ed.). Computer Vision – ECCV 2020: 16th European Conference on Computer Vision, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV. Cham: Springer, 2020. p. 589-607.

LIN, T. Y. *et al*. Microsoft COCO: common objects in context. In: FLEET, D. et al. (org.). Computer vision – ECCV 2014: 13th European conference on computer vision, Zurich, Switzerland, September 6-12, 2014, proceedings, part V. Cham: Springer International Publishing, 2014. p. 740-755.

LIU, Y. F. *et al*. Unbiased teacher for semi-supervised object detection. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2021. Proceedings... [S.l.: s.n.], 2021. p. 1-15.

LIU, Y.; WANG, J. Personal Protective Equipment Detection for Construction Workers: A Novel Dataset and Enhanced YOLOv5 Approach. IEEE Access, v. 12, p. 47338-47358, 2024.

LÓPEZ, O.; LÓPEZ, A.; CROSSA, J. Multivariate statistical machine learning methods for genomic prediction. Cham: Springer Nature Switzerland, 2022.

LOSHCHILOV, I.; HUTTER, F. SGDR: stochastic gradient descent with warm restarts. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2017. Proceedings... [S.l.: s.n.], 2017. p. 1-16.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2019. Proceedings... [S.l.: s.n.], 2019. p. 1-19.

NAKAMURA, K. *et al*. Learning-rate annealing methods for deep neural networks. Electronics, v. 10, n. 16, p. 2029, 2021.

NEMENYI, P. Distribution-free multiple comparisons. 1963. Thesis (Ph.D.) - Princeton University, Princeton, 1963.

OBSERVATÓRIO DE SEGURANÇA E SAÚDE DO TRABALHO. Dados consolidados de acidentes de trabalho no Brasil. Brasília: MPT/OIT, 2023. Disponível em: https://smartlabbr.org/sst. Acesso em: 14 jan. 2025.

OLIVER, A. *et al*. Realistic evaluation of deep semi-supervised learning algorithms. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 2018. Proceedings... [S.l.: s.n.], 2019. p. 3235-3246.

PADILLA, R.; NETTO, S. L.; DA SILVA, E. A. B. A survey on performance metrics for object-detection algorithms. In: INTERNATIONAL CONFERENCE ON SYSTEMS, SIGNALS AND IMAGE PROCESSING, 2020. Proceedings... [S.l.: s.n.], 2020. p. 237-242.

PATHIRAJA, B.; GUNAWARDHANA, M.; KHAN, M. Multiclass confidence and localization calibration for object detection. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2023. Proceedings... [S.l.: s.n.], 2023.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, v. 2, n. 1, p. 37-63, 2011.

QI, J.; NGUYEN, M.; YAN, W. Q. CISO: co-iteration semi-supervised learning for visual object detection. Multimedia Tools and Applications, v. 83, p. 33941-33957, 2024.

RAINA, R. *et al*. Self-taught learning: transfer learning from unlabeled data. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 2007. Proceedings... [S.l.: s.n.], 2007. p. 759-766.

RANI, V. *et al*. Self-supervised learning: a succinct review. Archives of Computational Methods in Engineering, v. 30, p. 2761-2775, 2023.

REDMON, J. *et al*. You only look once: unified, real-time object detection. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016. Proceedings... [S.l.: s.n.], 2016. p. 779-788.

SCHMIDHUBER, J. Deep learning in neural networks: an overview. Neural Networks, v. 61, p. 85-117, 2015.

SETTLES, B. Active learning literature survey. Madison: University of Wisconsin-Madison, 2009. (Computer Sciences Technical Report 1648).

SHAHNAWAZ, M.; KUMAR, M. A comprehensive survey on big data analytics: characteristics, tools and techniques. ACM Computing Surveys, v. 57, n. 8, p. 196, mar. 2025.

SINGH, B.; DAVIS, L. S. An analysis of scale invariance in object detection SNIP. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2018. Proceedings... [S.l.: s.n.], 2018. p. 3578-3587.

SNOW, R. *et al*. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2008. Proceedings...[S.l.: s.n.], 2008. p. 254-263.

SOHN, K. *et al*. FixMatch: simplifying semi-supervised learning with consistency and confidence. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 2020. Proceedings... [S.l.: s.n.], 2020. p. 596-608.

SOHN, K. *et al*. A simple semi-supervised learning framework for object detection. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2021. Proceedings... [S.l.: s.n.], 2021. p. 4654-4663.

SRIVASTAVA, N. *et al*. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, v. 15, n. 1, p. 1929-1958, 2014.

TIAN, Y. *et al*. Contrastive multiview coding. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2020. Proceedings... [S.l.: s.n.], 2020. p. 776-794.

TREISMAN, A. M.; GELADE, G. A feature-integration theory of attention. Cognitive Psychology, v. 12, n. 1, p. 97-136, 1980.

VOULODIMOS, A. *et al*. Deep learning for computer vision: a brief review. Computational Intelligence and Neuroscience, v. 2018, p. 1-13, 2018.

XIE, Q. *et al*. Self-training with noisy student improves ImageNet classification. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2020. Proceedings... [S.l.: s.n.], 2020. p. 10687-10698.

XU, J.; XIAO, L.; LÓPEZ, A. Self-Supervised Domain Adaptation for Computer Vision Tasks. IEEE Access, v. 7, p. 156694-156706, 2019.

XU, Y. *et al*. Revisiting pretraining for semi-supervised learning in the low-label regime. Neurocomputing, v. 565, 2024.

YANG, X. *et al*. A survey on deep semi-supervised learning. Knowledge-Based Systems, v. 226, p. 107-135, 2021.

YING, X. An overview of overfitting and its solutions. Journal of Physics: Conference Series, v. 1168, n. 2, p. 022-028, 2019.

YOO, D.; KWEON, I. S. Learning loss for active learning. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2019. Proceedings... [S.l.: s.n.], 2019. p. 93-102.

ZEILER, D.; FERGUS, R. Visualizing and Understanding Convolutional Networks. In: FLEET, D.; PAJDLA, T.; SCHIELE, B.; TUYTELAARS, T. (ed.). Computer Vision – ECCV 2014: 13th European Conference on Computer Vision. Cham: Springer, 2014. p. 818-833. (Lecture Notes in Computer Science, v. 8689)

ZHANG, K. *et al*. PatchNet: Maximize the Exploration of Congeneric Semantics for Weakly Supervised Semantic Segmentation. IEEE Transactions on Neural Networks and Learning Systems, v. 35, n. 8, p. 10984-10995, ago. 2024.

ZHAO, Z. *et al*. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Systems With Applications, v. 242, p. 122807, 2024.