XAI-Driven Loss Optimization via Activation Map Alignment in Multi-Task CNNs: A Systematic Literature Review

Gabriel Fragoso [1], **Teresa Ludermir**[1], **Júlia Fragoso**[1], **Diego Calixto**[1], **Gabriel Vital**

[1]Centro de Informática – Universidade Federal de Pernambuco (Cin – UFPE)
Recife – PE – Brasil

{gamf,tbl,jamf,dhvc, gvmov}@cin.ufpe.br

Abstract. *This paper presents a systematic literature review on the integration of explainable artificial intelligence techniques into loss functions for multi-task CNN. Following the guidelines of Kitchenham and Charters, studies published between 2020 and 2026 were collected from ACM Digital Library, IEEE Xplore, and Scopus. The selected papers were analyzed according to their alignment mechanisms, supervision signals, optimization strategies, and performance outcomes. The results indicate a growing shift from passive interpretability toward explanation-guided optimization, especially in medical imaging and remote sensing. The review also highlights challenges related to computational cost, hyperparameter sensitivity, and balancing attention constraints with contextual information.*

Resumo. *Este artigo apresenta uma revisão sistemática da literatura sobre a integração de técnicas de XAI em funções de perda de CNNs. Seguindo as diretrizes de Kitchenham e Charters, foram analisados estudos publicados entre 2020 e 2026 nas bases ACM Digital Library, IEEE Xplore e Scopus. Os trabalhos selecionados foram avaliados quanto aos seus mecanismos de alinhamento, sinais de supervisão, estratégias de otimização e desempenho. Os resultados indicam uma transição da interpretabilidade passiva para abordagens de otimização guiadas por explicações. A revisão também destaca desafios relacionados ao custo computacional, à sensibilidade, à hiperparametrização e ao equilíbrio entre a restrição da atenção do modelo e a preservação de informações contextuais.*

1. Introdução

Nos últimos anos, o Aprendizado Multitarefa consolidou-se como uma abordagem promissora na área de Visão Computacional, permitindo que redes neurais convolucionais (CNNs) aprendam múltiplas tarefas relacionadas por meio de representações compartilhadas, favorecendo a generalização e a eficiência computacional [Caruana 1997]. Estes modelos permitem que algoritmos executem tarefas complementares, como classificação e segmentação supervisionada, de forma simultânea a partir de uma representação compartilhada [Misra et al. 2016].

Contudo, o aumento exponencial na complexidade dessas arquiteturas deixa nítida a opacidade dos modelos, gerando sistemas operacionais do tipo ”caixa-preta”. A evolução acelerada desses algoritmos preditivos traz consigo desafios críticos quanto

ao viés algorítmico e dilemas éticos decorrentes da falta de transparência, tornando necessária a transição e o desenvolvimento de metodologias fundamentadas em Inteligência Artificial Explicável (XAI) [Ludermir 2021].

Tradicionalmente, ferramentas de XAI, como o Mapeamento de Ativação de Classe (CAM) [Zhou et al. 2016] e suas variantes baseadas em gradiente (Grad-CAM) [Selvaraju et al. 2017], têm sido aplicadas quase exclusivamente a um papel de diagnóstico post-hoc, utilizadas apenas para visualizar as regiões de interesse após a conclusão do treinamento para justificar decisões já tomadas pelo classificador [Shi 2025, Zhu et al. 2022]. Embora ferramentas clássicas de XAI venham sendo propostas para fornecer auditorias visuais e diagnósticos de forma passiva após o treinamento [de Araujo Morais et al. 2024], a integração direta desses mapas explicativos no cerne das funções de perda (loss functions) durante a etapa de treinamento para a resolução de conflitos de aprendizado permanece uma lacuna pouco explorada na literatura vigente.

A justificativa para romper com a passividade dos diagnósticos post-hoc e adotar a XAI como um mecanismo ativo de regularização reside no princípio pioneiro de que um modelo computacional deve ser ”certo pelos motivos certos”(right for the right reasons), conforme preconizado por Ross, Hughes e Doshi-Velez [Ross et al. 2017]. Ao estender esse conceito para o domínio das CNNs, por meio da penalização de desvios em mapas de ativação internos, induz-se um alinhamento espacial onde a rede é coagida a focar nas regiões semanticamente corretas do dado de entrada [Zhu et al. 2022, Rieger et al. 2020]. Essa abordagem não apenas atenua a volatilidade e o tratamento de incertezas inerentes a cenários complexos de modelagem, aspectos cruciais ressaltados por Morais e Ludermir [de Araujo Morais and Ludermir 2026], mas também atua como um regularizador estrutural capaz de suavizar a competição entre os gradientes das diferentes tarefas.

Diante desse cenário e das limitações observadas nas abordagens tradicionais, emerge a seguinte questão de pesquisa: de que forma técnicas de explicabilidade (XAI) são integradas às funções de perda para otimizar o alinhamento e mitigar conflitos de aprendizado em arquiteturas CNN multitask? Para responder a este questionamento, o objetivo desta pesquisa é identificar e categorizar as principais estratégias de integração de técnicas de Explainable AI (XAI) em funções de perda para a otimização de modelos CNN multitask.

Para atingir o objetivo, este artigo apresenta quatro contribuições principais. Primeiramente, propõe uma taxonomia das estratégias de Inteligência Artificial Explicável incorporadas à função de perda em redes neurais convolucionais multitarefa. Em seguida, realiza uma comparação entre diferentes mecanismos de alinhamento explicativo, destacando suas características e aplicações. Além disso, analisa os efeitos dessas abordagens sobre o desempenho dos modelos, a interpretabilidade, o custo computacional e a dinâmica de otimização durante o treinamento. Por fim, define uma agenda de pesquisa que mapeia desafios abertos considerados críticos para o avanço da área.

O restante deste artigo está organizado da seguinte forma: a Seção 2 detalha a metodologia de categorização e os cenários estabelecidos para o estudo das funções de perda. Na Seção 3, são expostos e discutidos os resultados comparativos obtidos. Por fim, a Seção 4 apresenta as considerações finais do trabalho, destacando as limitações encontradas e propondo direcionamentos para investigações futuras.

2. Método de Pesquisa

Este estudo adota a abordagem de Revisão Sistemática de Literatura (RSL) fundamentada nas diretrizes metodológicas propostas por Kitchenham e Charters [Kitchenham et al. 2007]. O método é estruturado com o intuito de mitigar vieses e garantir a transparência e a reprodutibilidade no mapeamento do estado da arte. Esta seção detalha as fases de planejamento da pesquisa, protocolo de intenções, aplicação das strings de busca nas bases selecionadas e relatórios do estudo.

2.1. Planejamento da Pesquisa

A fundamentação deste planejamento parte da necessidade de investigar como técnicas de Inteligência Artificial Explicável (XAI) podem atuar como mecanismos de regularização ativa durante o treinamento de redes profundas multitarefa. O objetivo principal desta pesquisa é identificar e categorizar as principais estratégias de integração de técnicas de XAI em funções de perda para a otimização de modelos CNN *multitask*.

Para nortear o mapeamento sistemático e a coleta de evidências, definiu-se a seguinte Questão de Pesquisa Central (**QPC**):

QPC: De que forma técnicas de explicabilidade (XAI) são integradas às funções de perda para otimizar o alinhamento e mitigar conflitos de aprendizado em arquiteturas CNN multitask?

Para desdobrar a questão central e viabilizar uma análise taxonômica precisa do domínio, foram estabelecidas Questões de Pesquisa Secundárias (**QPs**):

- **QP1:** Qual é a fonte de supervisão utilizada pela loss de alinhamento?
- **QP2:** Como a função de perda operacionaliza o alinhamento explicativo?
- **QP3:** Quais características permitem organizar as funções de perda de alinhamento explicativo em uma taxonomia coerente?
- **QP4:** Quais são os impactos do alinhamento explicativo no desempenho e na eficiência computacional?

2.2. Estratégia de Busca e Seleção de Fontes

A seleção das fontes de informação priorizou bases de dados digitais consolidadas e representativas no domínio da ciência da computação e engenharia de software. Desse modo, foram selecionadas três bases: **ACM Digital Library**, **IEEE Xplore** e **Scopus**.

A estratégia de busca fundamenta-se na varredura automatizada por meio de uma *search string* genérica. Esta consulta foi calibrada iterativamente para capturar de forma abrangente os eixos conceituais do estudo: Aprendizado Multitarefa, Arquiteturas Convolucionais, Explicabilidade e Otimização de Funções de Perda. A estrutura da *string* genérica é apresentada a seguir:

(”Multi-task”OR ”Multitask”OR ”MTL”OR ”Joint Learning”) AND (”CNN”OR ”Convolutional”OR ”Deep Learning”) AND (”XAI”OR ”Explainable AI”OR ”Saliency Map*”OR ”Activation Map*”OR ”Attribution”) AND (”Loss Function*”OR ”Alignment”OR ”Regularization”OR ”Penalty”)

Com o intuito de mapear as contribuições mais recentes e os avanços de fronteira na interseção entre XAI e otimização multitarefa, a janela temporal da pesquisa compreendeu o período entre os anos de 2020 e 2026. A execução do protocolo de busca e a consequente extração dos resultados foram realizadas no dia 04 de maio de 2026.

Para a execução prática do protocolo, a *string* genérica sofreu adequações sintáticas específicas para atender aos requisitos gramaticais, aos operadores lógicos e aos mecanismos de indexação de cada plataforma, resultando nas consultas estruturadas abaixo:

- **Scopus:**
  TITLE-ABS-KEY ((”Multi-task”OR ”Multitask”OR ”MTL”OR ”Joint Learning”) AND (”CNN”OR ”Convolutional”OR ”Deep Learning”) AND (”XAI”OR ”Explainable AI”OR ”Saliency Map*”OR ”Activation Map*”OR ”Attribution”) AND (”Loss Function*”OR ”Alignment”OR ”Regularization”OR ”Penalty”)) AND PUBYEAR $>$ 2020 AND PUBYEAR $<$ 2026
- **IEEE Xplore:**
  (((”All Metadata”:”Multi-task learning”OR ”Multitask”OR ”MTL”OR ”Joint Learning”) AND (”All Metadata”:”CNN”OR ”Convolutional”OR ”Deep Learning”) AND (”All Metadata”:”XAI”OR ”Explainable AI”OR ”Saliency Map*”OR ”Activation Map*”OR ”Attribution”) AND (”All Metadata”:”Loss Function*”OR ”Alignment”OR ”Regularization”OR ”Penalty”)))
- **ACM Digital Library:**
  [Full Text: ”multi-task”] OR [Full Text: ”multitask”]] AND [[Full Text: ”cnn”] OR [Full Text: ”convolutional”]] AND [[Full Text: ”xai”] OR [Full Text: ”explainable ai”]] AND [[Full Text: ”loss function”] OR [Full Text: ”alignment”] OR [Full Text: ”regularization”]] AND [Full Text: ”classification”] AND [Full Text: ”segmentation”] AND [E-Publication Date: (01/01/2022 TO 12/31/2026)]

2.3. Condução da Revisão

A execução das *search strings* customizadas nas bases de dados selecionadas resultou em um montante inicial de 115 estudos primários. A distribuição quantitativa inicial por indexador eletrônico consolidou-se da seguinte forma: 68 artigos recuperados na **ACM Digital Library**, 26 artigos no **IEEE Xplore** e 21 artigos na **Scopus**. Esses resultados foram exportados e organizados de forma centralizada em uma planilha eletrônica para dar início ao processo de filtragem.

Para garantir o alinhamento ao escopo, aplicaram-se sistematicamente os seguintes critérios: **CI1:** CNNs multitarefa com *backbone* compartilhado; **CI2:** refinamento mútuo dinâmico entre mapas de explicação e máscaras de segmentação via retropropagação; **CE1:** arquiteturas baseadas puramente em *Transformers*; **CE2:** XAI estritamente *post-hoc* (não integrada à perda); **CE3:** idiomas diferentes de inglês, português ou espanhol; **CE4:** estudos secundários ou terciários; e **CE5:** desalinhamento temático.

2.4. Processo de Triagem e Filtragem

A busca automatizada, realizada por Gabriel Fragoso em 04/05/2026, retornou um total de 115 artigos. O processo de seleção e o fluxo de afunilamento dos estudos seguiram as diretrizes do protocolo PRISMA, conforme ilustrado no fluxograma da Figura 1. Esta dinâmica de filtragem e dividiu-se em quatro etapas sequenciais:

1. Filtragem e Duplicatas: Os 115 resultados foram consolidados em uma planilha do Google Sheets, onde 4 duplicatas foram removidas manualmente, restando 111 estudos para triagem.

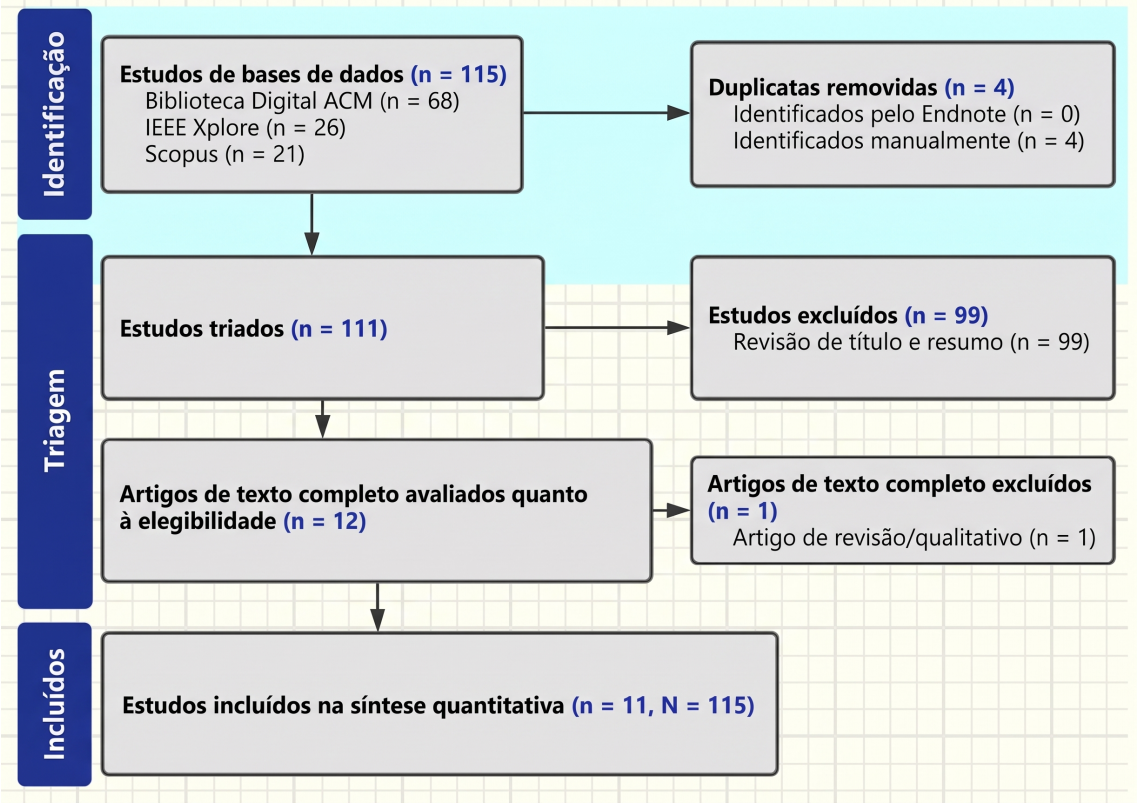


Figura 1. Fluxograma de seleção de estudos baseado no protocolo PRISMA.

2. Triagem por Título e Resumo: Realizou-se a avaliação pareada dos 111 estudos. Foram excluídos 99 artigos que violavam os critérios de inclusão. Os principais fatores de rejeição foram a ausência de arquiteturas CNN e o emprego de XAI apenas como análise *post-hoc* (sem integração à *loss function*). Esta fase resultou em 12 artigos candidatos para leitura integral (2 ACM, 6 IEEE Xplore, 4 Scopus).

3. Avaliação de Qualidade: Os 12 textos completos foram submetidos a um protocolo de extração e avaliação baseado em 5 critérios essenciais (clareza metodológica, delimitação prática, definição da proposta, discussão empírica e ameaças à validade). Cada critério foi pontuado em 0,0 (não conformidade), 0,5 (parcial) ou 1,0 (integral), exigindo-se um desempenho mínimo de 70% (3,5 pontos) para inclusão.

4. Seleção Final: Após a leitura crítica, 1 artigo (ACM) foi eliminado por insuficiência metodológica. A composição final da revisão consolidou-se em 11 estudos selecionados, cujos dados e mapeamento estão sintetizados na Tabela 1.

Tabela 1. Fluxo de seleção dos estudos por base de dados.

Fonte de Pesquisa	Registros encontrados	Após triagem	Após avaliação de qualidade	Estudos finais
ACM Digital Library	68	2	1	1
IEEE Xplore	26	6	6	6
Scopus	21	4	4	4
Total	**115**	**12**	**11**	**11**

3. Resultados

A distribuição temporal dos estudos selecionados evidencia um aumento no número de publicações ao longo dos anos. Foram identificados 2 artigos publicados em 2022, 1 artigo em 2023, 2 artigos em 2024, 5 artigos em 2025 e 1 artigo em 2026. Observa-se que 2025 concentrou o maior número de publicações da amostra, correspondendo a aproximadamente 45,5% dos estudos selecionados, sugerindo crescimento recente do interesse por mecanismos de interpretabilidade integrados ao treinamento.

No que se refere à distribuição geográfica, a pesquisa revela um ecossistema científico globalizado, com forte concentração na Ásia, região que lidera a produção com 5 dos estudos selecionados, provenientes da China. A América do Norte responde por 3 das publicações. A Europa está representada pela França, responsável por 1 artigo. Por fim, os 2 artigos restantes caracterizam-se pelo alcance intercontinental da área, dividindo-se de forma equitativa entre um estudo desenvolvido em Ugand e um consórcio multilateral que integra a cooperação conjunta entre pesquisadores de Omã, Índia e Malásia.

No que tange aos aspectos metodológicos, verifica-se a ausência de estudos de caso estritamente qualitativos, as pesquisas integradas a esta revisão adotam abordagens experimentais quantitativas.

3.1. Taxonomia das Estratégias de Alinhamento e Explicabilidade

Uma das contribuições centrais deste artigo consiste na formalização de uma taxonomia analítica para categorizar as funções de perda de alinhamento. A taxonomia foi construída por meio de análise indutiva dos estudos selecionados. As categorias foram definidas segundo o mecanismo utilizado para incorporar conhecimento explicativo durante o treinamento, considerando a origem do sinal supervisório e a forma de alinhamento entre a decisão do modelo e evidências relevantes.

Durante a extração e síntese dos dados, observou-se que as abordagens diferem principalmente quanto à origem do sinal supervisório utilizado para orientar o alinhamento explicativo e à estratégia empregada para incorporar esse conhecimento ao processo de treinamento. A partir dessas características, as funções de perda foram agrupadas em sete categorias fundamentais: (i) alinhamento por penalização direta de saliência, (ii) alinhamento por máscara supervisionada, (iii) alinhamento por pseudo-rótulos XAI, (iv) alinhamento multimodal, (v) alinhamento por representação interpretável, (vi) alinhamento multitarefa baseado em biomarcadores e (vii) alinhamento por restrições espaciais.

A Tabela 2 sintetiza a taxonomia proposta para as funções de perda de alinhamento explicativo. Para cada categoria explicitam-se o sinal de supervisão dominante e as características arquiteturais distintivas. Essa organização permite compreender as diferenças conceituais entre as abordagens e oferece uma visão sistemática das estratégias empregadas para incorporar interpretabilidade ao treinamento de modelos de Deep Learning.

Em suma, a taxonomia proposta organiza as funções de perda pelo seu modo de funcionamento. Esta estrutura pode servir como um guia para pesquisadores escolherem a melhor estratégia para desenvolver modelos que já nascem explicáveis.

3.2. Sistematização e Análise dos Estudos Selecionados

Com o objetivo de fornecer um panorama sobre o estado da arte, realizou-se um mapeamento sistemático dos principais frameworks que integram restrições explicativas ao

Tabela 2. Taxonomia unificada das estratégias de alinhamento explicativo e métodos de supervisão.

Categoria	Sinal / Fonte de Supervisão	Característica Principal
Penalização Direta de Saliência	Mapas de ativação (Saliência)	Penaliza a atenção fora da região de interesse.
Máscara Supervisionada	Máscaras clínicas	Alinha explicações a regiões anotadas por especialistas.
Pseudo-rótulos XAI	Grad-CAM / CAM	Utiliza explicações geradas automaticamente como supervisão.
Multimodal	Imagem + Texto	Integra evidências visuais e semânticas.
Representação Interpretável	Modelo explicável	Alinha o espaço latente a representações interpretáveis.
Multitarefa com Biomarcadores	Biomarcadores (Tarefas auxiliares)	Aprende simultaneamente o diagnóstico e fatores causais.
Restrições Espaciais	Restrições Coordenadas	Penaliza a distância ou sobreposição para guiar o foco da explicação.

ciclo de otimização de redes profundas. A Tabela 3 sumariza os onze estudos selecionados, classificando-os de acordo com a taxonomia proposta nesta investigação, além de detalhar seus respectivos domínios de aplicação, mecanismos de incorporação na função de perda (*loss function*), ganhos empíricos reportados e limitações.

A análise comparativa revela uma forte predominância do emprego dessas técnicas no domínio de imagens médicas, abrangendo investigações em oftalmologia [Shi 2025], neurologia [Mukherjee et al. 2025], oncologia [Dai et al. 2025, Amyar et al. 2022], ortopedia [Aravinda et al. 2026] e pneumologia [Zhu et al. 2022]. Essa concentração pode ser justificada pela criticidade da área, em que a explicabilidade atua como requisito fundamental para a validação e confiabilidade clínica dos diagnósticos automatizados. Fora do escopo médico, observam-se contribuições relevantes voltadas ao sensoriamento remoto e análise de imagens aéreas [Li et al. 2023, Liu et al. 2024], engenharia de software [Xu et al. 2025] e visão computacional aplicada à avaliação estética [Yu and Liu 2024] ou cenários complexos superlotados [Hu 2025].

No que tange aos mecanismos de alinhamento, os artigos estudados mostram que a co-otimização multitarefa e o uso de máscaras supervisionadas são mais eficazes para garantir a fidelidade explicativa sem penalizar severamente a acurácia do classificador.

Os métodos de explicabilidade baseados em mapas de ativação endógenos (como CAM e Grad-CAM) foram mapeados em mais de metade da amostra analítica devido à sua flexibilidade arquitetural e independência de modificações estruturais complexas nas redes convolucionais. Paralelamente, os pesquisadores estão deixando de usar apenas mapas visuais e passando a criar sistemas focados em conceitos práticos. Essas técnicas traduzem os mapas de calor complexos em palavras e termos que os profissionais da área conseguem compreender facilmente.

Além disso, os resultados encontrados indicam que o paradigma do aprendizado multitarefa estabelece-se como uma técnica de alinhamento implícito. Ao unificar tarefas acopladas em uma mesma representação compartilhada, o processo de retropropagação

Tabela 3. Mapeamento comparativo dos estudos baseados em funções de perda de alinhamento explicativo.

Estudo/Ano	Taxonomia XAI	Domínio / Tarefa	Como a XAI é Incorporada à Função de Perda	Ganhos Reportados	Limitações Chave
Aravinda et al. (2026)	Pseudo-rótulos XAI	Imagens Médicas (Raio-X) / Classificação	Mapas Grad-CAM atuam como pseudo-rótulos geoespaciais. Uma perda de consistência BCE penaliza desvios entre a cabeça auxiliar e o mapa retropropagado.	No dataset MURA: acurácia de 81,49% e F1 de 0,815 supera ResNet50 (76,5%) e ViT (72,1%).	Latência de explicadores pós-hoc inviabiliza tempo real (SHAP 1,20s; LIME 2,30s)
Dai et al. (2025)	Multimodal / Híbrido	Imagens Médicas (Ultrassom) / Classificação	Alinhamento hierárquico em 4 níveis (Global, Máscara Local, Auto-atenção e Conceitos). Ancoragem entre a geometria da máscara e a taxonomia do laudo.	Co-otimização de acurácia e fidelidade explicativa; maior robustez e transferibilidade transdomínio.	Dependente da completude de anotações especialistas; vulnerável a ruídos intrínsecos do ultrassom.
Hu (2025)	Restrições Espaciais	Visão Computacional / Segm. fina e bordas	Pré-processamento via *early fusion* (soma ponderada imagem-saliência) agindo como limitador de escopo geométrico.	Incrementos em mIoU e Precisão Média (AP); robustez visual contra ruídos.	Altos custos computacionais e de memória (3 dias de treino em GPU 1080).
Li et al. (2023)	Pseudo-rótulos XAI	Sensoriamento Remoto / Classificação	Máscara de atenção seletiva baseada em saliência XAI para isolar o alvo e suprimir o background.	Redução drástica do viés de desordem (*clutter bias*) do plano de fundo.	Forte dependência de hiperparametrização e *data augmentation*.
Liu et al. (2024)	Restrições Espaciais	Sensoriamento Remoto / Segm. Hierárquica	Decomposição taxonômica macro/micro; máscara macro concatenada como restrição suave na entrada micro.	Ganhos em mIoU e OA ao interromper propagação de erros; coerência geométrica.	Alto uso de GPU; miopia em classes minoritárias com texturas similares.
Shi (2025)	Multitarefa com Biomarcadores	Imagens Médicas / Classif. e Segm.	Otimização conjunta via Entropia Cruzada Binária (BCE) e Perda de Dice morfológica com pesos compartilhados.	AUC evolui de 0,905 para 0,932; R^2 ajustado de 56% comprova nexo causal.	Desbalanceamento de classes; imprecisão residual no mapeamento do copo óptico.
Mukherjee et al. (2025)	Máscara Supervisionada	Imagens Médicas (MRI) / Classificação	*Parts Regularizer* calcula a perda de Dice entre a ativação latente e a máscara especialista (cérebro/CSF).	Acurácia de 90,19% e AUC de 91,79% (3 classes); isolamento de microestruturas.	Falhas em *edge cases* por variabilidade geográfica e cobertura restrita.
PET-CAM (2025)	Penalização Direta de Saliência	Imagens Médicas / Detecção Fraca	Minimização da Perda de Distância geométrica entre o centroide do CAM e um ponto central p da lesão.	Coeficientes Dice de 0,77 (pulmão) e 0,73 (esôfago); mitiga falsas ativações.	Ausência de validação multicêntrica; forte dependência de *priors* espaciais.
Xu et al. (2025)	Representação Interpretável	Engenharia de Software / Classif. Binária	Loop de feedback DP-CNN e SDT; penalidade via Perdas de Fidelidade de Saída e de Recursos na loss conjunta.	Omissão Falsa -59,30%; F-measure +14,67%; AUC +4,85%.	Baixo desempenho em datasets muito pequenos; alta sensibilidade ao *fine-tuning*.
Yu & Liu (2024)	Multitarefa com Biomarcadores	Avaliação Estética / Regressão de atributos	Otimização multitarefa acoplando Perda de Classificação aos Pares e Perda Euclidiana para atributos formais.	Alinhamento com julgamento humano no EVA (composição ρ=0,91, semântica ρ=0,88).	Restrições por simplicidade estrutural; volume restrito do dataset disponível.
Zhu et al. (2022)	Penalização Direta de Saliência	Imagens Médicas (CXR) / Classificação	Camada CAM nativa; otimização via Cross-Entropy e Perda de Erro Quadrático Médio Seletivo contra *human gaze*.	ResNet+GG-CAM eleva interpretabilidade de cardiomegalia para 94,5%	Dilatação do tempo de convergência; rigidez para generalização arquitetural.

força a coerência topológica das representações latentes. O modelo é induzido a aprender, simultaneamente, as propriedades discriminatórias do objeto e sua respectiva localização espacial.

3.3. Complexidade Operacional e Trade-offs Computacionais

A integração de mecanismos de explicabilidade (*XAI*) e alinhamento estruturado ao ciclo de otimização impõe desafios de custo computacional e latência [Aravinda et al. 2026, Hu 2025, Liu et al. 2024, Zhu et al. 2022]. No treinamento, a modificação do fluxo convolucional por funções de perda complexas prolonga a convergência e o ajuste de hiperparâmetros [Li et al. 2023, Xu et al. 2025].

Em arquiteturas multitarefa com alinhamento por restrições espaciais [Hu 2025, Liu et al. 2024], a fusão precoce de mapas de saliência e restrições hierárquicas estende o processamento para até 3 dias em GPUs dedicadas (NVIDIA GeForce GTX 1080) e eleva drasticamente o consumo de VRAM.

No cenário de inferência, os estudos mostram que a escolha do método de *XAI* dita a viabilidade em tempo real. Enquanto abordagens integradas ativamente (Grad-CAM) mantêm baixa latência, métodos *pós-hoc* como SHAP e LIME geram gargalos operacionais para sistemas contínuos [Aravinda et al. 2026]. Além disso, frameworks de alinhamento multimodal [Dai et al. 2025] ou representação interpretável [Li et al. 2023, Xu et al. 2025] dependem da completude de anotações especialistas e sofrem com *overfitting* ou ruídos físicos em cenários de poucos dados, exigindo rotinas complexas de ajuste fino. Esses modelos exibem alta sensibilidade ao volume e qualidade dos dados.

3.4. Impacto da Explicabilidade na Dinâmica de Otimização e Desempenho

A introdução de sinais explicativos nas funções de perda altera a dinâmica de otimização de uma simples minimização de erro de rótulo (classificação) para uma restrição e ajuste espacial de características.

Em modelos multitarefa de alinhamento explicativo, as diferentes frentes de predição são processadas por uma base compartilhada. O processo de retropropagação conjunta das múltiplas funções de perda obriga a rede a compartilhar representações neurais entre as tarefas. Isso força o modelo a tomar suas decisões com base em características semanticamente ricas e alinhadas entre si, limitando drasticamente as chances da rede utilizar atalhos prejudiciais.

A introdução de uma loss de supervisão concorre diretamente com a loss de classificação. Os estudos mostram que o ajuste dessa dinâmica pode ser otimizado, através do uso de pesos de incerteza homoscedástica ou parâmetros dinâmicos de regularização, para dar ênfase à supervisão da atenção visual apenas nas épocas iniciais e, depois, focar no erro de classificação para evitar a degradação da performance preditiva [Zhu et al. 2022, Liu et al. 2024].

Os achados dos diferentes estudos apontam para avanços quantitativos em tarefas de classificação e segmentação, com destaque para o ambiente de diagnóstico médico. Os modelos avaliados registraram saltos de desempenho, como o aumento da AUC na classificação de glaucoma (de 0,905 para 0,932) [Shi 2025] e o alcance de 90,19% de

acurácia em cenários classificação de três classes, sendo capazes de isolar com precisão microestruturas de necrose e inflamação [Mukherjee et al. 2025]. Em tarefas de segmentação anatômica [Amyar et al. 2022], os coeficientes Dice atingiram 0,77 e 0,73, superando amplamente a U-Net tradicional, enquanto o framework MoRE-XAI superou arquiteturas consolidadas como a ResNet50 e o Vision Transformer no dataset MURA [Aravinda et al. 2026].

Paralelamente à eficiência métrica, os estudos demonstram uma evolução na fidelidade e na interpretabilidade da IA Explicável (XAI). A arquitetura ResNet+GG-CAM [Zhu et al. 2022], por exemplo, elevou a congruência semântica na identificação de cardiomegalia para 94,5%, mitigando ativações fora do alvo anatômico, enquanto a técnica PET-CAM delimitou estritamente a massa tumoral por meio de uma perda por distância [Amyar et al. 2022].

Dessa forma, o alinhamento explicativo apresenta-se não apenas como um mecanismo secundário de auditoria post-hoc, mas como um princípio de design para a próxima geração de sistemas multitarefa, onde a precisão estatística e a fundamentação semântica operam de maneira integrada.

4. Considerações Finais

Este trabalho apresentou uma Revisão Sistemática da Literatura sobre o uso de sinais explicativos incorporados diretamente às funções de perda de redes neurais convolucionais multitarefa. A revisão foi conduzida seguindo as diretrizes de Kitchenham e Charters e o protocolo PRISMA, abrangendo publicações entre 2020 e 2026 indexadas nas bases ACM Digital Library, IEEE Xplore e Scopus. A partir de um conjunto inicial de 115 estudos, foram selecionados 11 trabalhos primários para análise detalhada.

Os resultados indicam um movimento gradual da interpretabilidade baseada exclusivamente em métodos pós-hoc para abordagens que integram informações explicativas ao próprio processo de treinamento dos modelos. Nesse contexto, foi proposta uma taxonomia que organiza as estratégias de alinhamento explicativo de acordo com a origem do sinal de supervisão e a forma de integração à arquitetura. Observou-se também que essas abordagens são aplicadas principalmente em domínios nos quais a interpretabilidade desempenha papel relevante, como diagnóstico por imagens médicas e sensoriamento remoto. Além disso, os estudos analisados sugerem que a combinação entre aprendizado multitarefa e alinhamento explicativo pode contribuir para a obtenção de representações latentes mais consistentes e para a redução da dependência de correlações espúrias presentes nos dados.

Apesar dos resultados reportados na literatura, foram identificados desafios importantes. Entre eles destacam-se o aumento do custo computacional associado ao treinamento, a necessidade de ajuste cuidadoso dos pesos entre as funções de perda e a dificuldade de definir o grau adequado de restrição das regiões de atenção do modelo sem comprometer informações contextuais potencialmente relevantes. Esses fatores evidenciam a existência de trade-offs entre interpretabilidade, desempenho preditivo e eficiência computacional.

Como perspectivas futuras, destacam-se a avaliação dessas abordagens em cenários com maior diversidade de dados e menor disponibilidade de amostras rotuladas, o desenvolvimento de métodos de alinhamento com menor custo computacional e

a investigação de mecanismos explicativos capazes de associar padrões aprendidos pela rede a conceitos semanticamente interpretáveis por especialistas. De modo geral, os estudos analisados indicam que a incorporação de sinais explicativos às funções de perda representa uma direção promissora para o desenvolvimento de modelos multitarefa mais transparentes e alinhados aos requisitos de aplicações críticas.

Agradecimentos

O presente trabalho foi realizado com o apoio do projeto IAIA, sendo financiado pelo CNPq (Processo nº 406417/2022-9) e pela FACEPE (Processo nº IBPG-0771-1.03/25).

Referências

Amyar, A., Modzelewski, R., Vera, P., Morard, V., and Ruan, S. (2022). Weakly supervised tumor detection in pet using class response for treatment outcome prediction. *Journal of Imaging*, 8(5):130.

Aravinda, C., Shaffi, N., Nair, S. S. K., Khan, M. S., Viswan, V., and Arjun, B. (2026). More-xai: Moment-based rigid registration and multi-perspective explainability for fracture localization in x-ray imaging. *IEEE Access*.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Dai, G., Wang, C., Tang, Q., Zhang, Y., Dai, D., Qiao, L., Yan, J., and Chen, H. (2025). Interpretable breast cancer diagnosis using histopathology and lesion mask as domain concepts conditional simulation ultrasonography. *Information Fusion*, 123:103343.

de Araujo Morais, L. R., de Melo Fragoso, G. A., Ludermir, T. B., and Monteiro, C. L. A. (2024). Explainable ai for the brazilian stock market index: A post-hoc approach to deep learning models in time-series forecasting. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 436–447. SBC.

de Araujo Morais, L. R. and Ludermir, T. (2026). Explainability and uncertainty for time-series classification in cryptocurrency data: a hybrid xai methodology based on comte and leftist. *International Journal of Data Science and Analytics*, 22(1):127.

Hu, W. (2025). Instance-level human parsing based on semantic-edge fusion in crowded scenes. In *2025 7th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, pages 1095–1098. IEEE.

Kitchenham, B., Charters, S., et al. (2007). Guidelines for performing systematic literature reviews in software engineering.

Li, W., Yang, W., Zhang, W., Liu, T., Liu, Y., and Liu, L. (2023). Hierarchical disentanglement-alignment network for robust sar vehicle recognition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:9661–9679.

Liu, L., Tong, Z., Cai, Z., Wu, H., Zhang, R., Le Bris, A., and Olteanu-Raimond, A.-M. (2024). Hieru-net: A hierarchical semantic segmentation method for land cover mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14.

Ludermir, T. B. (2021). Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos avançados*, 35:85–94.

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003.

Mukherjee, S., Templeton, K., Tindimwebwa, S., Lin, I., Sutin, J., Yu, M., Peterson, M., Truwit, C., Schiff, S. J., and Monga, V. (2025). Glapal-h: Global, local, and parts aware learner for hydrocephalus infection diagnosis in low-field mri. *IEEE Transactions on Biomedical Engineering*.

Rieger, L., Singh, C., Murdoch, W., and Yu, B. (2020). Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shi, L. (2025). Medically explainable deep learning-based glaucoma diagnosis. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–6. IEEE.

Xu, G., Feng, C., Guo, X., Zhu, Z., and Wang, W. (2025). A joint learning framework for bridging defect prediction and interpretation. *IEEE Transactions on Reliability*.

Yu, T. and Liu, J. (2024). Multi-task learning of visual attributes for image aesthetics assessment. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 350–354. IEEE.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Zhu, H., Salcudean, S., and Rohling, R. (2022). Gaze-guided class activation mapping: Leverage human visual attention for network attention in chest x-rays classification. In *Proceedings of the 15th International Symposium on Visual Information Communication and Interaction*, pages 1–8.