UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Bauru

Juliana da Costa Feitosa

Análise da Explicabilidade dos Métodos de Inteligência Artificial eXplicável Aplicados a Segmentação de Imagens de Peixes Pacu

Bauru, São Paulo, Brasil

16 de abril de 2024

Juliana da Costa Feitosa

Análise da Explicabilidade dos Métodos de Inteligência Artificial eXplicável Aplicados a Segmentação de Imagens de Peixes Pacu

Tese de Doutorado, para o curso de Pós-Graduação em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus de Bauru.

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. José Remo Ferreira Brega

Bauru, São Paulo, Brasil

16 de abril de 2024

F311a

Feitosa, Juliana da Costa
Análise da explicabilidade dos métodos de Inteligência Artificial eXplicável aplicados a segmentação de imagens de peixes pacu / Juliana da Costa Feitosa. -- Bauru, 2024
92 f. : il., tabs., fotos

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru
Orientador: José Remo Ferreira Brega

1. Inteligência Artificial. 2. Inteligência Artificial Explicável. 3. Avaliação. 4. Perturbação de Pixels. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Bauru

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE JULIANA DA COSTA FEITOSA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 07 dias do mês de março do ano de 2024, às 14:00 horas, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de JULIANA DA COSTA FEITOSA, intitulada **Análise da Explicabilidade dos Métodos de Inteligência Artificial eXplicável Aplicados a Segmentação de Imagens de Peixes Pacu**. A Comissão Examinadora foi constituida pelos seguintes membros: Prof. Dr. JOSE REMO FERREIRA BREGA (Orientador(a) - Participação Virtual) do(a) Departamento de Computação / UNESPCâmpus de Bauru, Prof. Dr. DANILO MEDEIROS ELER (Participação Virtual) do(a) Departamento de Matemática e Computação / Faculdade de Ciências e Tecnologia de Presidente Prudente, Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Participação Virtual) do(a) Departamento de Ciência da Computação / UNESPBauru, Profa. Dra. VALERIA FARINAZZO MARTINS (Participação Virtual) do(a) Faculdade de Computação e Informática / Universidade Presbiteriana Mackenzie, Prof. Dr. ALEXANDRE CARDOSO (Participação Virtual) do(a) Faculdade de Engenharia Elétrica / Universidade Federal de Uberlândia. Após a exposição pela doutoranda e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, a discente recebeu o conceito final:_ _ _ _ _ _ _ _ _ _ _ _ _ **APROVADA** _ _ _ . Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Assinado de forma digital por Jose Remo Ferreira Brega:09610599826
Dados: 2024.03.07 16:40:18 -03'00'

Prof. Dr. JOSE REMO FERREIRA BREGA

Juliana da Costa Feitosa

Análise da Explicabilidade dos Métodos de Inteligência Artificial eXplicável Aplicados a Segmentação de Imagens de Peixes Pacu

Tese de Doutorado, para o curso de Pós-Graduação em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus de Bauru.

Comissão Examinadora

Prof. Prof. Dr. José Remo Ferreira Brega
Universidade Estadual Paulista “Júlio de Mesquita Filho- Bauru
Orientador

Prof. Dr. Kelton Augusto Pontara da Costa
Universidade Estadual Paulista “Júlio de Mesquita Filho- Bauru

Prof. Dr. Danilo Medeiros Eler
Universidade Estadual Paulista “Júlio de Mesquita Filho- Presidente Prudente

Prof. Dr. Alexandre Cardoso
Universidade Federal de Uberlândia - Uberlândia

Profa. Dra. Valeria Farinazzo Martins
Universidade Presbiteriana Mackenzie - São Paulo

Bauru, São Paulo, Brasil
16 de abril de 2024

Agradecimentos

Primeiramente agradeço a Deus, que me deu forças e condições para continuar e seguir em frente, mesmo em meio às dificuldades. Esse trabalho é graças a Ele e para Ele.

Agradeço aos familiares e amigos por todo apoio e incentivo que me deram durante todo esse tempo, pois sei que sem eles esse trabalho não seria possível. Obrigada mãe e Amanda por serem minha base. Obrigada a todas as minhas amigas por sempre me apoiarem, e obrigada Ana Paula Borgo por ter acreditado em mim até o fim. Sei que você estaria muito feliz se estivesse aqui.

Agradeço ao Prof. José Remo Ferreira Brega por mais uma vez acreditar no meu trabalho. Sou grata por todos os conselhos e por ter me ensinado a ser a profissional que sou hoje. Obrigada por tudo professor, pois esse trabalho não seria possível sem sua orientação. Foi uma honra ser sua orientada mais uma vez.

Agradeço ao Prof. João Paulo Papa por todo apoio e ajuda durante esse processo. Agradeço também aos demais professores do Departamento de Computação da Faculdade de Ciências da Unesp de Bauru, por todo apoio e torcida.

Agradeço ao Fabrício Batista que teve paciência em me ensinar tudo aquilo que eu precisava aprender para chegar até aqui. Sem você esse trabalho não seria possível. Muito obrigada por tudo.

Agradeço aos meus colegas de laboratório e projeto por todo o trabalho que realizamos juntos. Vocês fazem parte dessa história.

Agradeço aos meus alunos, que me ensinaram a ser professora e sempre torceram por mim. Muito obrigada por todo apoio meus queridos. Vocês também fazem parte dessa conquista.

Agradeço ao IBILCE pelo apoio acadêmico proporcionado pelo Programa de Pós-Graduação em Ciência da Computação.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), à qual agradeço.

Este trabalho foi realizado com muito esforço e dedicação. Foram inúmeras as vezes que pensei em desistir. Mas hoje vejo que valeu a pena chegar até aqui.

"Não temas, porque eu sou contigo; não te assombres, porque eu sou teu Deus; eu te fortaleço, e te ajudo, e te sustento com a destra da minha justiça." (Bíblia Sagrada, Isaías 41. 10)

Resumo

A concepção envolvendo o termo Inteligência Artificial (IA) traz grandes preocupações em relação a transparência das informações fornecidas como saída, exigindo cada vez mais explicações para que os usuários entendam sobre esse tipo de abordagem. Esse campo de pesquisa é chamado de Inteligência Artificial eXplicável (XAI) e pode ser definido como um conjunto de técnicas que une métodos de IA com abordagens explicativas eficazes para gerar saídas explicáveis, tornando os modelos mais compreensíveis para o ser humano. Apesar da XAI ser apresentada como uma das soluções para a falta de transparência dos modelos de IA do tipo caixa-preta, contestar as explicações geradas por esses métodos também se faz necessário, visto que a explicabilidade é considerada uma tarefa necessária para a melhora dos resultados. Somado a isso, diante dos vários métodos existentes, estratégias de melhoria das explicações tornaram-se contribuições significativas para a área. Com objetivo de analisar e proporcionar melhores resultados aos métodos de XAI aplicados a segmentação de imagens de peixes da espécie Pacu, foram utilizadas 100 imagens de entrada para o presente trabalho. Portanto, dois experimentos foram conduzidos com o intuito de encontrar o melhor método explicável, a partir de técnicas de perturbação de pixels, e também melhorá-los por meio da combinação lógica entre o melhor método e os demais. Foi a primeira vez que essa metodologia foi aplicada no contexto da Aquicultura. A partir dos resultados obtidos, o Mapa de Saliência obteve os melhores resultados ao demonstrar pixels que de fato foram considerados como os de maior relevância pela Mask R-CNN utilizada. Além disso, foi possível observar que a metodologia implementada proporcionou não só a melhora dos métodos explicáveis Grad-CAM, CNN Filters e Layer Grad-CAM, como também foi possível obter resultados melhores do que os alcançados por meio do Mapa de Saliência. Por fim, conclui-se que a principal contribuição deste trabalho está na possibilidade de melhorar a qualidade das explicações de um método explicável de forma simples, bem como a implementação de uma avaliação eficaz baseada em perturbação de pixels.

Palavras-chave: Inteligência Artificial. Inteligência Artificial Explicável. Avaliação. Perturbação de Pixels.

Abstract

The concept surrounding the term Artificial Intelligence (AI) raises major concerns regarding the transparency of the information provided as output, requiring more and more explanations for users to understand this type of approach. This field of research is called eXplainable Artificial Intelligence (XAI) and can be defined as a set of techniques that combine AI methods with effective explanatory approaches to generate explainable outputs, making models more understandable to humans. Although XAI is presented as one of the solutions to the lack of transparency in black-box AI models, contesting the explanations generated by these methods is also necessary, since explainability is considered a necessary task to improve results. Furthermore, given the various existing methods, strategies for improving explanations have become significant contributions to the area. To analyze and provide better results for the XAI methods applied to the segmentation of images of fish of the Pacu species, 100 input images were used for the present work. Therefore, two experiments were conducted with the aim of finding the best explainable method, using pixel perturbation techniques, and also improving them through the logical combination between the best method and the others. It was the first time that this methodology was applied in the context of Aquaculture. From the results obtained, the Saliency Map obtained the best results by demonstrating pixels that were considered the most relevant by the Mask R-CNN used. Furthermore, it was possible to observe that the implemented methodology not only improved the explainable methods Grad-CAM, CNN Filters, and Layer Grad-CAM, but it was also possible to obtain better results than those achieved using the Saliency Map. Finally, it is concluded that the main contribution of this work is the possibility of improving the quality of explanations of a simply explainable method, as well as the implementation of an effective evaluation based on pixel perturbation.

Keywords: Artificial Intelligence. eXplainable Artificial Intelligence. Evaluation. Pixels Perturbation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACM	*Association of Computing Machinery*
API	Interface de Programação de Aplicações (do inglês, *Application Programming Interface*)
CAM	Mapas de Ativação de Classes (do inglês, *Class Activation Map*
CNN	*Convolutional Neural Network*
CVS	*Computer Vision System*
DARPA	Agência de Projetos de Pesquisa Avançada de Defesa (do inglês, *Defense Advanced Research Projects Agency*)
DL	*Deep Learning*
DNN	Redes Neurais Profundas (do inglês, *Deep Neural Network*)
EUA	Estados Unidos da América
FC	Totalmente Conectada (do inglês, *Fully Connected*
GAP	Agrupamento Médio Global (do inglês, *Global Average Pooling*
GDPR	Regulamento Geral de Proteção de Dados (do inglês, *General Data Protection Regulation*)
IA	Inteligência Artifical
IEEE	*Institute of Electrical and Eletronics*
IoU	Intersecção Sobre a União
LaGeAC	Laboratório de Genética em Aquicultura e Conservação
ML	*Machine Learning*
MoRF	*Most Relevant First*
PICOC	*População, Intervenção, Comparação, Saídas e Contexto (do inglês,*Population, Intervention, Comparison, Outcomes, Context
ReLU	*Rectified Linear Unit)*
ResNet	Redes Residuais Profundas (do inglês, *Residual Neural Network*

RSL	Revisão Sistemática da Literatura
SD	*Sorensen Dice*
UNESP	Universidade Estadual Paulista
XAI	Inteligência Artificial eXplicável (do inglês, *eXplainable Artificial Intelligence*)

Sumário

1 Introdução

Recentes estudos mostram que a Inteligência Artificial (IA) deixou de ser tema de filmes de ficção científica, (e.g., o filme "2001 – Uma odisseia no espaço"criado em 1968 por Stanley Kubrick), para ser o foco de grandes estudos e pesquisas (MUTHUKRISHNAN et al., 2017). Assim, a IA é comumente usada atualmente para descrever as mais novas experiências de interação entre sistemas computacionais e seus usuários (KAUFMAN, 2019). Por meio desta interação, pode-se afirmar que sistemas e conceitos de IA podem ser encontrados em inúmeras áreas do conhecimento, como por exemplo, Direito, Medicina, Engenharia e Matemática (RUSSELL; NORVIG, 2004). Sua história começou muito antes do surgimento da própria sigla. Foi em 1943 que o conceito de neurônio artificial (atualmente denominado *perceptron*) foi criado como um modelo para representar um neurônio matematicamente. Essa criação auxiliou no avanço dos estudos e pesquisas sobre IA, inclusive no desenvolvimento das famosos Redes Neurais (MCCULLOCH; PITTS, 1943a).

A visão completa da IA foi apresentada pela primeira vez no artigo intitulado *"Computing Machinery and Intelligence"* escrito em 1950 por Alan Mathison Turing (TURING, 2009). Considerado o pai da computação, Turing também pode ser considerado o pai da IA, apesar de não ter usado o termo em seu artigo. O autor escreveu seu trabalho em torno da seguinte pergunta: "As máquinas podem pensar?". Apesar de parecer utópico para a época, foi a partir desse artigo que novas pesquisas começaram a surgir com o intuito de responder a pergunta formulada por Turing (MOOR, 2003).

Foi apenas em 1956 que o campo "Inteligência Artificial" surgiu na chamada Conferência de Dartmouth, que reuniu pesquisadores famosos como Claude Shannon e John McCarthy (MCCARTHY et al., 2006). A partir dessa conferência, surgiram novas pesquisas relacionadas a área e assim, nos anos seguintes, surgiram os chamados sistemas especialistas, considerados os primeiros sistemas de IA, cujo objetivo era aplicar um conhecimento para solucionar problemas específicos (HARMON; MAUS; MORRISSEY, 1988).

O objetivo da IA está em fazer com que máquinas simulem parcialmente o funcionamento da mente humana (KISTAN; GARDI; SABATINI, 2018). Dessa forma, não existe ainda um sistema de IA que simule completamente o nosso cérebro, e que resolva todo e qualquer tipo de problema solucionado por um ser humano. Todavia, ainda não se tem conhecimento de todos os problemas que são capazes de serem solucionados por meio de sistemas inteligentes ou de sua total capacidade (TEIXEIRA, 2019). Apesar disso, sua amplitude permite resolver problemas para áreas como a Aquicultura, cuja busca

por alternativas rápidas e não invasivas para medição de características dos peixes, tem incentivado a implementação de soluções de IA (FREITAS et al., 2023). A exemplo disso, a segmentação de imagens, que na maioria da vezes é realizada manualmente, pode ser implementada a partir de um algoritmo inteligente que auxilia na extração automatizada de medidas biométricas de peixes vivos, possibilitando que o animal fique menos estressado do que na segmentação manual (FERNANDES et al., 2020).

A evolução da computação e das pesquisas científicas em torno da IA permitiram o surgimento de duas abordagens para essa área. A primeira delas é a IA Baseada em Conhecimento, cujo objetivo é traduzir o conhecimento obtido em código de sistema (HARMON; MAUS; MORRISSEY, 1988). A segunda abordagem é a de Aprendizado Estatístico, cuja finalidade é criar e aplicar métodos estatísticos para que a máquina aprenda sozinha suas funções. Essa abordagem, além de ser a mais utilizada, também é vista como uma subárea da IA chamada de Machine Learning (ML) (MICHIE et al., 1994).

O surgimento de um modelo de aprendizagem depende dos dados de entrada utilizados por algoritmos de ML para aplicar equações matemáticas pré-definidas (QUIONERO-CANDELA et al., 2009). Assim, dados rotulados possuem um parâmetro disponível para consulta, como por exemplo, dados de pacientes de um hospital. Nesse exemplo, é possível ter como parâmetro disponível a informação de que o paciente tem ou não diabetes. Dessa forma, segundo Hodge e Austin (2004), existem três abordagens de ML: supervisionado, cujos dados são todos rotulados; não supervisionado, cujos dados não são rotulados; e semi-supervisionado, cujos dados de entrada são mistos.

O conceito de Redes Neurais também é considerado uma subárea da ML e tem por objetivo simular os neurônios humanos (MCCULLOCH; PITTS, 1943b). Assim, o marco inicial das pesquisas relacionadas a ela está na criação de um neurônio artificial chamado de *perceptron* (ROSENBLATT, 1958). Uma rede neural é treinada e não programada, e por isso, depende de um algoritmo de aprendizagem. Segundo Rauber (2005), o algoritmo de aprendizagem apresentado inicialmente em um *perceptron* era capaz de adaptar os pesos internos do neurônio para resolver problemas em que a classificação não podia ser feita linearmente. Apesar do sucesso apresentado para classificação linear, esse algoritmo não conseguiu resolver o problema do XOR (ou exclusivo). Portanto, foi necessário acrescentar mais uma camada de neurônios artificiais, além de precisar desenvolver novos algoritmos de aprendizagem para que o problema fosse enfim solucionado (MINSKY; PAPERT, 2017).

A partir desse contexto, uma outra subárea da ML, chamada *Deep Learning* (DL) ou Aprendizagem Profunda, foi criada com a finalidade de configurar os parâmetros dos dados de entrada para que a máquina aprenda sozinha, por meio de reconhecimento de padrões, em várias camadas de neurônio artificial (GOODFELLOW et al., 2016). Dessa forma, DL é usado atualmente para reconhecimento de imagem, fala, detecção de objetos

e descrição de conteúdo (DENG; YU, 2014).

De acordo com (FELLOUS et al., 2019), existem duas classificações possíveis para os modelos de ML existentes. A primeira são os modelos caixas-pretas, cujas decisões realizadas pela máquina são dificilmente explicáveis para um ser humano (e.g., DL). As caixas-pretas são consideradas mais complexas e com maior desempenho. Em contrapartida, existem modelos caixas-brancas cujas decisões são explicáveis, e portanto, mais transparentes (e.g., árvores de decisão) (CAMACHO et al., 2018).

Todo esse avanço da IA fez com que a preocupação em torno da transparência das decisões também alavancasse. A exemplo disso, recentemente foi possível observar o surgimento do requisito legal prescrito pelo art. 22 do Regulamento Geral de Proteção de Dados (*General Data Protection Regulation* - GDPR) descrito pela jurisdição da União Europeia (WOLF; RINGLAND, 2020; ARNOUT et al., 2019) que assegura a transparência das decisões de um sistema IA. Também foi possível observar o surgimento da Lei de Proteção de Dados Pessoas (LGPD), que semelhantemente demonstra a preocupação em relação aos avanços da IA e proteção de dados em âmbito nacional (PINHEIRO, 2020). Sendo assim, a busca por alternativas seguras e transparentes tornou-se prioridade por pesquisadores na área.

Baseado no contexto apresentado é que surgiu a Inteligência Artificial eXplicável (*eXplainable Artificial Intelligence* - XAI), definida como um conjunto de técnicas que une métodos de IA com abordagens transparentes eficazes para gerar saídas explicáveis (FELLOUS et al., 2019). Ou seja, o termo XAI refere-se a técnicas que tornam modelos de IA compreensíveis para o ser humano (WOLF; RINGLAND, 2020). Apesar de ganhar a devida atenção recentemente, de acordo com Xu et al. (2019), os conceitos relacionados à XAI datam de 40 anos atrás, onde regras eram usadas para explicar o funcionamento de sistemas especialistas. Entretanto, a Agência de Projetos de Pesquisa Avançada de Defesa (*Defense Advanced Research Projects Agency* - DARPA) dos Estados Unidos da América (EUA), criou em 2017 um programa destinado à XAI, cujo objetivo era criar sistemas de IA capazes de explicar sua lógica para o ser humano, de forma a caracterizar seus pontos fortes e fracos, e transmitir informações comportamentais futuras (GUNNING; AHA, 2019).

De acordo com Wolf e Ringland (2020), as explicações podem ser classificadas em global, cujo objetivo é descrever as representações do modelo utilizado, e local, cujo objetivo é explicar os dados de entrada. Além disso, a explicabilidade ocorre conforme a necessidade de compreensão do ser humano em relação ao sistema (WOLF, 2019). Assim, baseadas no usuário, que pode ser um especialista ou não, as explicações fornecidas pela IA podem ser faladas ou criadas para serem visualizadas, conforme a necessidade (WEBER et al., 2018). Segundo as definições da DARPA, as explicações podem ser classificadas em quatro modos: declarações analíticas, visualizações, casos e rejeições de escolhas alternativas (GUNNING,

2017). Em ambos os casos, a explicabilidade é alcançada com base no processo de predição do modelo de IA analisado.

A classificação de métodos de XAI também pode ser definida segundo a metodologia utilizada para gerar as explicações. Segundo (IVANOVS; KADIKIS; OZOLS, 2021), as técnicas de perturbação de pixels, por exemplo, permitem analisar a entrada do modelo em relação a saída do mesmo. Com isso, existem métodos de XAI baseados em perturbação de pixels, como por exemplo LIME e Occlusion, cujo objetivo é entender da melhor forma possível o funcionamento dos modelos de IA (IVANOVS; KADIKIS; OZOLS, 2021). Assim, a imagem de entrada é repetidamente modificada por meio de desfoques ou cores aleatórias em regiões específicas da imagem (HENDRYCKS; DIETTERICH, 2019). Dessa forma, os resultados obtidos são comparados com os resultados da imagem de entrada original (sem perturbação). Portanto, a região da imagem é considerada significativa se a sua remoção resulta em uma mudança perceptível no resultado (GUPTA; KOUNDAL; MONGIA, 2023).

Apesar da XAI ser apresentada como uma das soluções para a falta de transparência dos modelos de IA, contestar as explicações geradas por esses métodos também se faz necessário, visto que essas explicações podem trazer diferentes resultados quando submetidas aos processos globais e locais dos modelos, por exemplo (GHASSEMI; OAKDEN-RAYNER; BEAM, 2021). Segundo Doshi-Velez e Kim (2017) é preciso ter cuidado com os métodos interpretáveis, evitando afirmações vagas, e considerando fatores relevantes às tarefas realizadas e ao método utilizado. De acordo com (GHASSEMI; OAKDEN-RAYNER; BEAM, 2021), apesar de serem atraentes devido a explicabilidade apresentada, os métodos de XAI podem ter suas explicações dificultadas pela presença de fatores de confusão não reconhecidos. Por isso, é preciso verificar também se os resultados obtidos por esses métodos não sofrem alterações quando o modelo de IA é submetido a fatores externos, como por exemplo, alterações na imagem de entrada.

Para a área de segmentação de imagens, cujo objetivo é destacar uma ou mais regiões da entrada, a necessidade de uma boa explicação é ainda maior, já que buscase entender quais foram os pixels mais relevantes para a tomada de decisão do modelo (MINAEE et al., 2021). Diante desse cenário, é válido afirmar a importância da utilização de métodos explicáveis que forneçam explicações fidedignas ao usuário, com base nos conceitos e estudos relacionados a transparência dos modelos de IA. Bem como a realização de experimentos que determinam qual técnica apresenta melhor resultado explicável, inclusive na combinação com outras técnicas existentes, mesmo diante de situações controversas propostas pela perturbação de pixels, por exemplo.

1.1 Problema

A busca pela transparência dos modelos de IA tem contribuído para os avanços na área de XAI. Entretanto, entender o funcionamento dessas técnicas é primordial para poder avaliar os seus resultados, sendo esse um dos problemas envolvendo IA explicável. Para Doshi-Velez e Kim (2017), a busca pela interpretabilidade faz com que possíveis soluções sejam criadas, sem ao menos serem avaliadas sobre sua eficácia. Dessa forma, os autores afirmam a importância em ter cuidado na busca pela explicabilidade com base em questões que ainda estão em aberto sobre o tema, como por exemplo, a falta de rigor durante a avaliação dos resultados explicáveis. Assim, espera-se que os avanços nas pesquisas relacionadas a XAI tragam contribuições focadas nas avaliações e métricas para as explicações obtidas.

Outro problema envolvendo o tema é a falta de garantia de que as explicações resultantes das técnicas de XAI, de fato, condizem com a realidade apresentada pelo processo de predição dos modelos de IA. Assim, de acordo com Ghassemi, Oakden-Rayner e Beam (2021) os métodos de explicabilidade ainda não podem fornecer garantias de que uma decisão individual está correta, principalmente no contexto da área da saúde no qual o estudo foi realizado. Além disso, os autores afirmam que há diferença de comportamento desses métodos quando submetidos a problemas globais e locais, e que atualmente, busca-se explicações compreensíveis, para seres humanos, que possam ser usadas com segurança na tomada de decisões voltadas para a saúde.

Diante desse contexto, Kaur et al. (2020) alegam que há pouca preocupação sobre até que ponto estas ferramentas atingem o objetivo de serem explicáveis. De acordo com o resultados apresentados pelos autores, as visualizações produzidas por métodos de XAI podem, na maioria das vezes, auxiliar cientistas de dados a descobrirem problemas em conjuntos de dados ou em modelos de IA. Entretanto, as visualizações dos resultados podem acarretar no excesso de confiança pelo usuário, juntamente com o uso indevido desses métodos. Assim, em ambos os trabalhos apresentados, os autores reforçam a importância de novas pesquisas que ajudem na regulamentação dos métodos explicáveis. Além disso, se faz necessário entender o funcionamento dessas ferramentas em diferentes cenários, e avaliar os seus resultados com rigor.

A partir dos trabalhos analisados, o principal problema a ser investigado é a falta de garantia na procedência das explicações em relação aos resultados obtidos durante a predição do modelo de IA aplicado a segmentação de imagens. Uma estratégia na busca em resolver esse problema envolve o entendimento de técnicas explicáveis, bem como a análise da influência de agentes externos sobre seus resultados e a melhora das explicações com base na análise realizada.

1.2 Hipóteses e Questões de Pesquisa

De acordo com a problemática apresentada anteriormente, foram levantadas as seguintes questões de pesquisa:

- *Q1: Como avaliar se os métodos de XAI explicam o que de fato aconteceu na predição do modelo de IA?*
- *Q2: Os métodos de XAI são influenciados pela perturbação de pixels das imagens de entrada do modelo?*
- *Q3: Como melhorar métodos de XAI para apresentarem explicações condizentes com a predição do modelo de IA?*

A partir dessas questões, as hipóteses de pesquisa foram formuladas e são apresentadas a seguir: é viável testar métodos de XAI para poder analisar sua explicabilidade em relação a predição de um modelo aplicado à segmentação de imagens. Além disso, vale analisar o resultado de cada método, submetendo-os a um cenário de perturbação de pixels das imagens de entrada do modelo. A partir dessa análise, é possível classificar qual técnica apresenta melhor desempenho, e assim, usar essa mesma técnica para aprimorar o resultado dos demais métodos, com base na combinação entre eles.

Para validar ou não essas hipóteses, todas as questões apresentadas são respondidas diante dos resultados alcançados durante a pesquisa. Para isso, uma Revisão Sistemática da Literatura (RSL) foi realizada para verificar trabalhos relacionados a essas questões, trazendo o embasamento teórico científico necessário. Além disso, foi considerado o desenvolvimento de uma metodologia baseado em experimentos, que é apresentada ao longo desse trabalho.

1.3 Objetivos e Contribuições

O objetivo dessa tese é baseado nas hipóteses de pesquisa apresentadas. Assim, objetiva-se analisar a explicabilidade de métodos de XAI aplicados ao modelo de segmentação de imagens de peixes Pacu. Os métodos explicáveis foram submetidos a situações adversas, como a perturbação de pixels, a fim de encontrar o melhor método, baseado na predição do modelo de IA. Dessa forma, objetiva-se realizar experimentos que comprovem essas hipóteses e busque melhorar métodos de XAI a partir da melhor técnica explicável encontrada.

Essa tese apresenta a perturbação de pixels como ferramenta de análise dos métodos explicáveis, cujo objetivo é verificar a influência do cenário adverso em relação a qualidade das explicações. Com isso, a perturbação de pixels aqui não é usada como base

dos métodos de XAI, mas sim, como uma ferramenta de avaliação dos mesmos. Para isso, os seguintes passos foram necessários para obtenção desse objetivo:

- Implementação do modelo de segmentação de imagens de peixes Pacu;
- Aplicação de métodos XAI de visualização;
- Entendimento do funcionamento dos métodos aplicados para análise das explicações geradas;
- Análise dos métodos explicáveis diante de um ambiente sujeito a perturbação de pixels das imagens de entrada;
- Busca pelo método explicável mais consistente, diante do ambiente criado; e
- Busca pela melhor estratégia, a partir do método mais consistente, para melhorar os resultados.

De acordo com o objetivo principal apresentado, os experimentos foram realizados na linguagem Python, a partir do uso do Google Colaboratory. Além disso, foram usadas magens disponibilizadas pelo Laboratório de Genética em Aquicultura e Conservação (LaGeAC) da Unesp de Jabotical. Essas imagens são de peixes da espécie *Piaractus mesopotamicus*, conhecida popularmente como Pacu. O código dos experimentos está disponível para consulta pública em um repositório do Github [1].

Essa tese apresenta contribuições definidas a partir das hipóteses de pesquisa desenvolvidas. São elas:

- Implementação de métodos explicáveis restritos a segmentação de imagens de peixes da espécie Pacu;
- Criação de uma técnica de avaliação baseada em perturbação de pixels para métodos de XAI;
- Localização do melhor método de XAI para o cenário proposto;
- Aprimoramento de métodos de XAI já existentes a partir do melhor método encontrado; e
- Criação de artigos para publicação dos resultados obtidos.

[1] https://github.com/jufeitosa10/XAI_Segmentacao_Pacu.git

1.4 Estrutura da Tese

A disposição desta tese de doutorado foi realizada da seguinte forma: além desta introdução, no Capítulo 2 Fundamentação Teórica serão apresentados os conceitos principais relacionados ao tema; já no Capítulo 3 Revisão Sistemática Literatura serão apresentados os trabalhos resultantes da RSL que servem como base para comparação da presente tese, enquanto que a metodologia utilizada é apresentada no Capítulo 4 Metodologia. No Capítulo 5 Experimentos e Resultados são apresentados os experimentos realizados de acordo com as hipóteses de pesquisa, juntamento com os resultados obtidos e a análise crítica sobre eles. E por último, no Capítulo 6 Conclusões e Trabalhos Futuros serão apresentadas as conclusões, discussões e trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo serão apresentados os principais conceitos relacionados a este estudo, como por exemplo, Rede Neural Convolucional (*Convolutional Neural Network* - CNN). Além disso, também são apresentados os métodos explicáveis Grad-CAM, Mapa de Saliência, CNN Filters e Layer Grad-CAM, implementados nos experimentos realizados para a presente tese. Por fim, o conceito de perturbação de pixels é explicado, bem como o conceito de segmentação de imagens para a área de Aquicultura.

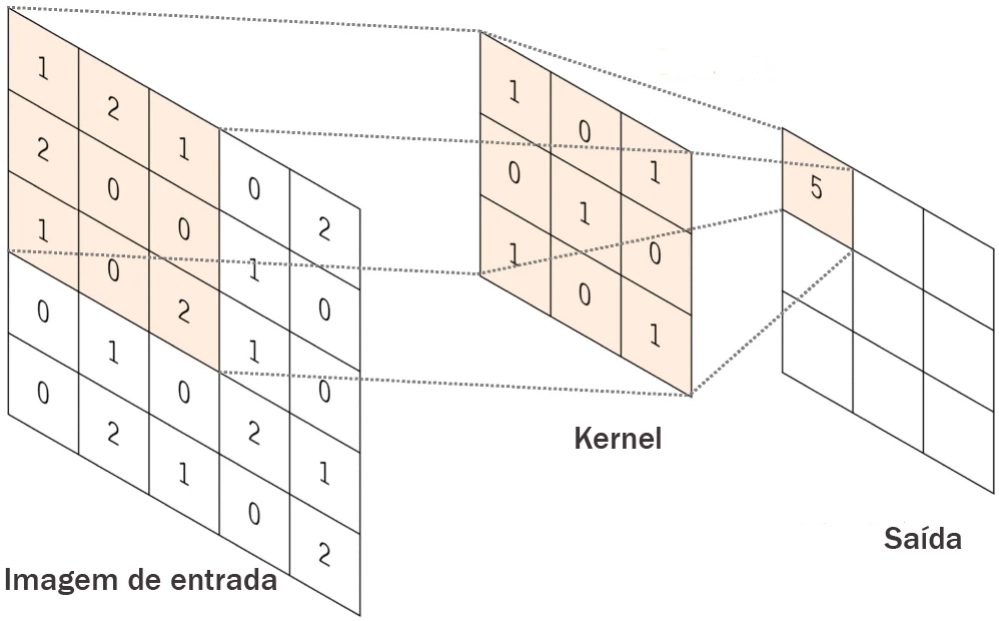
2.1 Redes Neurais Convolucionais

A CNN é definida como uma arquitetura de ML inspirada no mecanismo natural de percepção visual de seres vivos. Segundo Hubel e Wiesel (1962), as células do córtex visual animal são responsáveis pela detecção de luz em campos receptivos. A partir disso, Kunihiko Fukushima propôs em 1980 o Neocognitron, uma rede neural hierárquica de múltiplas camadas relacionada ao reconhecimento de padrões visuais (FUKUSHIMA, 1988). Já em 1990, o conceito de *backpropagation* foi implementado pela primeira vez em uma estrutura moderna chamada de LeNet-5 (LECUN et al., 1989), cuja finalidade estava em classificar dígitos manuscritos (*dataset* conhecido como MNIST). Além disso, Zhang et al. (1996) desenvolveram uma Rede Neural Artificial Invariante ao Deslocamento (SIANN) que tinha como objetivo reconhecer caracteres de uma imagem. Entretanto, a rede era incapaz de resolver problemas de maior complexidade, como por exemplo, classificação de imagens e vídeos de alta resolução. Por fim, devido a relação existente entre pixels mais próximos, as CNNs consistem em um poderoso conjunto de redes neurais que aprendem a partir dos dados da imagem, de forma a explorar algumas das estruturas mais conhecidas. Atualmente, essas redes estão presentes, principalmente, no campo da Visão Computacional e são caracterizadas pela sua eficiência em tarefas de classificação de imagens, por exemplo (ZHANG et al., 2023).

De acordo com Yamashita et al. (2018), a CNN é uma construção matemática normalmente composta por três tipos de camadas: convolução, *pooling* e camadas totalmente conectadas. A camada convolucional visa aprender representações de recursos das entradas. Dessa forma, ela é composta por vários núcleos de convolução (*kernel)* que são usados para calcular diferentes mapas de características (GU et al., 2018). Isto é, as convoluções funcionam como filtros que são posicionados em toda localização espacial da imagem captando os traços mais marcantes. Portanto, o tamanho dos mapas de características a serem construídos é determinado pelo tamanho do *kernel*, enquanto que a profundidade é definida pela quantidade de filtros (LEE; LEE; LEE, 2021). Na Figura 1 é possível

observar um exemplo de convolução para uma entrada bidimensional 5x5 e um *kernel* 3x3, resultando em uma saída 3x3.

Figura 1 – Camada de convolução de uma CNN.



Fonte - Adaptado de (YAMASHITA et al., 2018)

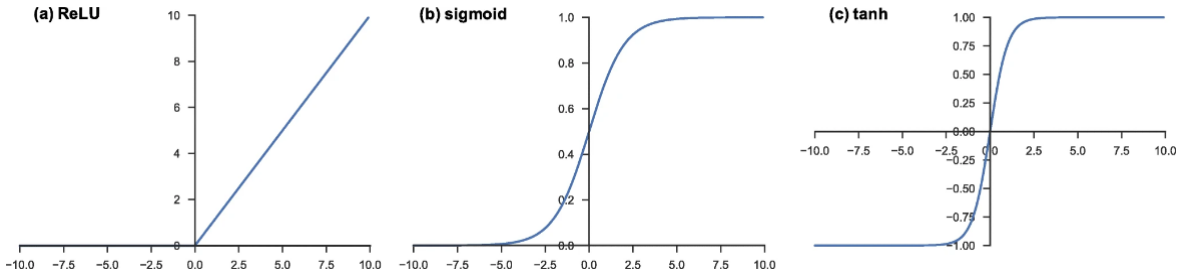
O resultado da convolução passa por uma função de ativação não linear, cujo objetivo é limitar a saída gerada. Com isso, a utilização dessas funções em Redes Neurais é motivada pela necessidade de introduzir aprendizado não linear ao modelo (ALBAWI; MOHAMMED; AL-ZAWI, 2017). Além disso, busca-se encontrar funções que sejam simples, pois uma função complexa reduz a velocidade do cálculo (HAO et al., 2020). Para CNNs, as funções Sigmoide e Tangente Hiperbólica (tanh) foram utilizadas com mais frequência, e são caracterizados por serem representações matemáticas do comportamento de um neurônio biológico (ALBAWI; MOHAMMED; AL-ZAWI, 2017). Atualmente, a função de ativação não linear mais comum é a Unidade Linear Retificada (ReLU) (NAIR; HINTON, 2010), que realiza o cálculo apresentado na Equação 2.1:

$$f(x) = max(0, x). \tag{2.1}$$

A função ReLU cria uma representação mais esparsa, já que o zero no gradiente leva à obtenção de um zero completo. Todavia, as funções de ativação anteriores sempre apresentaram resultados diferentes de zero no gradiente, o que pode dificultar o processo de treinamento do modelo (ALBAWI; MOHAMMED; AL-ZAWI, 2017). A Figura 2 ilustra o comportamento das três funções de ativação apresentadas.

A camada de *pooling* tem como função reduzir o tamanho espacial do mapa de características, a fim de diminuir a complexidade de outras camadas. Para a área de Processamento de Imagens, por exemplo, essa funcionalidade equivale à redução da

Figura 2 – Funções de ativação.

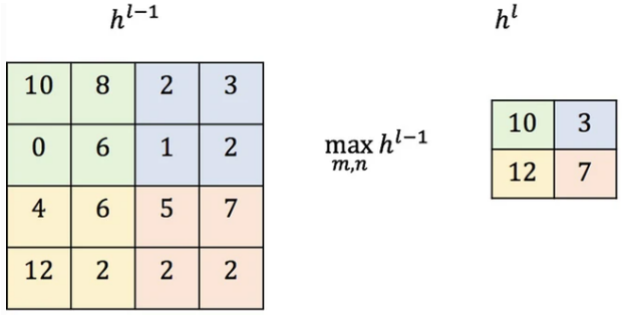


Fonte - (YAMASHITA et al., 2018)

resolução da imagem (ALBAWI; MOHAMMED; AL-ZAWI, 2017). Assim, da mesma forma que na camada de convolução, um filtro é utilizado para realizar o redimensionamento. Em contrapartida, nenhum peso é treinado para que essa camada possa ser implementada. A abordagem mais comum para essa camada é chamada de *pooling* máximo ou *max pooling*, cujo objetivo é escolher o maior valor a partir de um filtro que percorre e particiona a imagem (LEE; LEE; LEE, 2021). A partir dessa abordagem, é possível alcançar a invariância que permite a identificação de uma característica, independentemente de sua localização precisa (KAMATH et al., 2019). Dessa forma, uma camada de *pooling* máximo executa a transformação apresentada na Equação 2.2, cujos valores p e q indicam as coordenadas do neurônio em sua vizinhança local, e l representa a camada. A Figura 3 apresenta o funcionamento de uma camada de *pooling* máximo com um filtro de tamanho 2x2, na qual é possível observar a redução da dimensão no plano do mapa de características por um fator de 2.

$$h_{i,j}^{l} = \max_{p,q} h_{i+p,j+q}^{l-1}. \tag{2.2}$$

Figura 3 – Exemplo de camada de *pooling* máximo.



Fonte - Adaptado de (KAMATH et al., 2019)

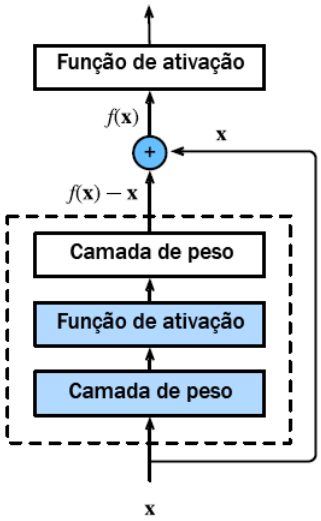
A última camada de uma CNN é denominada como Camada Totalmente Conectada ou *Fully Connected* (FC), cuja função é classificar a partir dos mapas de características extraídos das camadas implementadas anteriormente (LEE; LEE; LEE, 2021). Essa camada final normalmente possui o mesmo número de nós de saída que o número de classes. Além

disso, cada FC é seguida por uma função de ativação ReLU, por exemplo (YAMASHITA et al., 2018). Por fim, vale ressaltar que toda arquitetura CNN possui uma ou mais camadas FCs, conforme a profundidade do modelo implementado e o avanço dos estudos relacionados às CNNs (BASHA et al., 2020).

2.1.1 Redes Residuais Profundas

Conforme as redes ficam cada vez mais profundas, torna-se importante entender como a adição de camadas pode aumentar a complexidade e a expressividade da rede (ZHANG et al., 2023). Diante disso, em 2015 uma Rede Residual Profunda (ResNet) foi proposta por He et al. (2016) para reconhecimento de imagens, e essa rede consiste em um tipo de CNN, cuja entrada da camada anterior é adicionada à saída da camada atual. Essa característica contribui para o aprendizado da rede e na melhora do desempenho computacional (SHAFIQ; GU, 2022). Assim, a diferença entre uma ResNet e uma CNN típica pode ser observada em suas arquiteturas, na qual a ResNet possui um caminho de atalho conectando diretamente a entrada e a saída em um bloco de construção (WU; ZHONG; LIU, 2018). A Figura 4 ilustra o bloco residual utilizado para essa arquitetura, cuja linha sólida carregando a entrada da camada x para o operador de adição é chamada de conexão residual (ou conexão de atalho).

Figura 4 – Bloco residual.



Fonte - Adaptado de (ZHANG et al., 2023)

Nesse contexto, a função residual *F(x)* é aprendida treinando as camadas de peso, a partir do uso de dados rotulados. Além disso, as camadas de peso podem consistir em qualquer tipo de camada da Rede Neural, como por exemplo, camadas convolucionais ou FC (NIBALI; HE; WOLLERSHEIM, 2017). Por fim, o bloco residual permite que a passagem direta por meio da rede ignore seletivamente certas camadas, cuja função

F(x) é definida como sendo igual a zero. Dessa forma, a propagação acontece mais rápido devido as conexões residuais entre as camadas (ZHANG et al., 2023). Por isso, conforme a profundidade de uma rede aumenta, torna-se cada vez mais difícil para os gradientes realizarem a retropropagação (aumentar ou diminuir os pesos) da função de perda para as várias camadas, sem diminuir para zero ou ultrapassar o limite determinado. Entretanto, as ResNets permitem que gradientes passem sem atenuação pelas partes do modelo, usando as conexões de acordo com a função apresentada na Equação 2.3 (NIBALI; HE; WOLLERSHEIM, 2017). Dessa forma, esse tipo de rede torna-se mais fácil e simples de modificar, permitindo a utilização rápida e generalizada da mesma (ZHANG et al., 2023).

$$F(x) = H(x) - x. \tag{2.3}$$

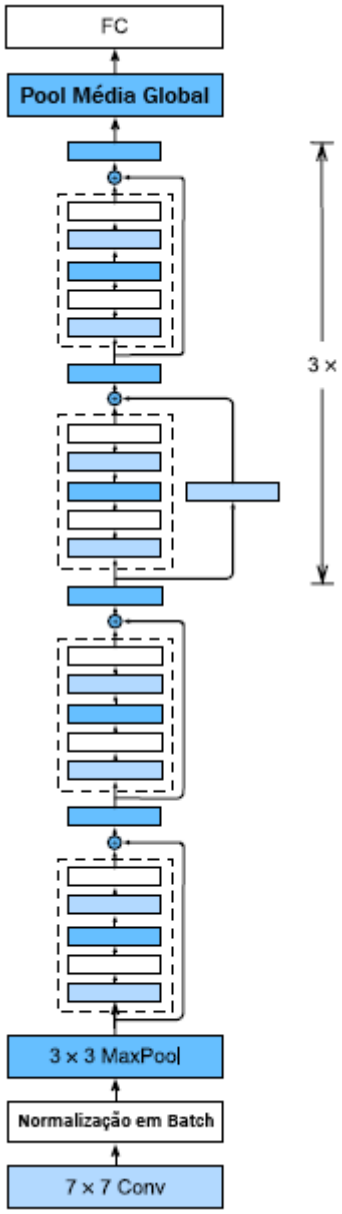
Combinando diferentes números de camadas e blocos residuais, é possível criar inúmeros modelos, como por exemplo, a ResNet com 18 camadas no total. Nesse contexto, a Figura 5 apresenta um exemplo de arquitetura de ResNet-18 que também contém uma camada de normalização de *batch*, cujo objetivo é acelerar a etapa de treinamento da rede por meio da regularização de pesos. Além disso, essa arquitetura também possui uma camada de *Pooling* de Média Global (GAP, do inglês *Global Average Pooling*) utilizada para obter a média de cada mapa de característica resultante do modelo. Com isso, em relação a imagem de entrada, a resolução diminui enquanto o número de canais aumenta até o ponto em que uma camada de GAP agrega todos os recursos (ZHANG et al., 2023). Por fim, a ResNet teve uma grande influência nos modelos subsequentes de Redes Neurais, e por isso, é considerada importante até hoje, tanto de natureza convolucional quanto sequencial.

2.1.2 Mask R-CNN

A tarefa de segmentação de instâncias tornou-se fundamental para o campo de Visão Computacional, pois consiste em localizar objetos em uma imagem de entrada, utilizando cada pixel da mesma. Essa tarefa contribui para aplicações, como por exemplo, direção autônoma, robótica e edição de imagens (CHENG et al., 2020). O principal desafio em sua implementação consiste na detecção correta de todos os objetos em uma imagem, ao mesmo tempo que segmenta com precisão cada instância. Dessa forma, a segmentação de instâncias combina elementos de detecção de objetos, cujo objetivo é utilizar uma caixa delimitadora para delimitação, e segmentação semântica, cuja função é classificar cada pixel em um conjunto fixo de categorias, sem diferenciar instâncias de objetos (HE et al., 2017).

Diante desse contexto, dentre os métodos existentes atualmente voltados para a tarefa de segmentação de instâncias, tem-se o Mask R-CNN desenvolvido por He et al.

Figura 5 – Arquitetura da ResNet-18.



Fonte - Adaptado de (ZHANG et al., 2023)

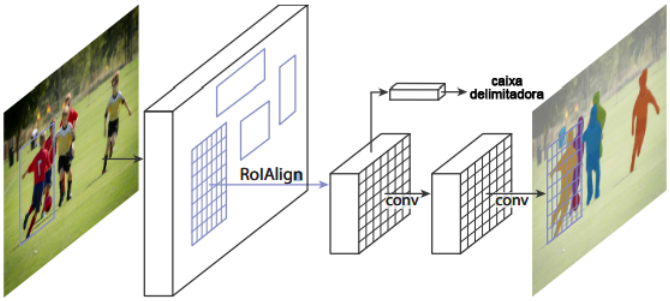
(2017), cujo objetivo é prever máscaras de segmentação em Regiões de Interesse (RoI). Assim, essa tarefa é realizada com base em previsões pixel a pixel, feitas com base nas informações locais fornecidas pelo modelo (CHENG et al., 2020). Essa abordagem é baseada no método Fast R-CNN, que combina extração de recursos, classificador e previsão de caixa delimitadora, juntamente com um método chamado *Pool* de Regiões de Interesse (RoIPool). O objetivo é reduzir o número de convoluções e o tempo de detecção para acelerar o processo. Para a abordagem Mask R-CNN, a vantagem está na segmentação automática da imagem, que é feita antes mesmo da classificação do objeto detectado, além de permitir delimitar o objeto e encontrar seus pontos-chaves (CHIAO et al., 2019).

Para o processo de treinamento, He et al. (2017) definiram uma perda multitarefa

em cada RoI amostrado, com base na Equação 2.4. Com isso, defini-se *Lcls* como perda de classificação e *Lbox* como a perda da caixa delimitadora. Já a máscara possui uma saída dimensional, que codifica K máscaras binárias de resolução $m \times m$, sendo uma para cada uma das K classes. Dessa forma, aplica-se a função Sigmoide para cada um dos pixels, cujo valor *Lmask* é definido como sendo a perda média da entropia cruzada binária. A definição dada a *Lmask* permite que máscaras sejam geradas para todas as classes sem que haja confusão entre elas. A Figura 6 apresenta a arquitetura utilizada pelo método.

$$L = Lcls + Lbox + Lmask. \tag{2.4}$$

Figura 6 – Arquitetura da Mask R-CNN.



Fonte - (HE et al., 2017)

A Mask R-CNN é comumente utilizada em combinação com as ResNets, cujos benefícios alcançados são principalmente em relação ao ganho de velocidade e a precisão da segmentação (HE et al., 2017). Os próprios autores do método utilizam uma ResNet-101 para apresentar os resultados promissores dessa abordagem, aplicados ao *dataset* COCO. Em Cheng et al. (2020), por exemplo, os autores utilizam a combinação entre a Mask R-CNN e uma ResNet-50 para comparar com outra abordagem de segmentação desenvolvida por eles. Além disso, o problema de segmentação de imagens é contexto para diversas áreas, inclusive a área da saúde, mais precisamente a análise de exames médicos (SHU et al., 2020).

2.2 Inteligência Artificial eXplicável

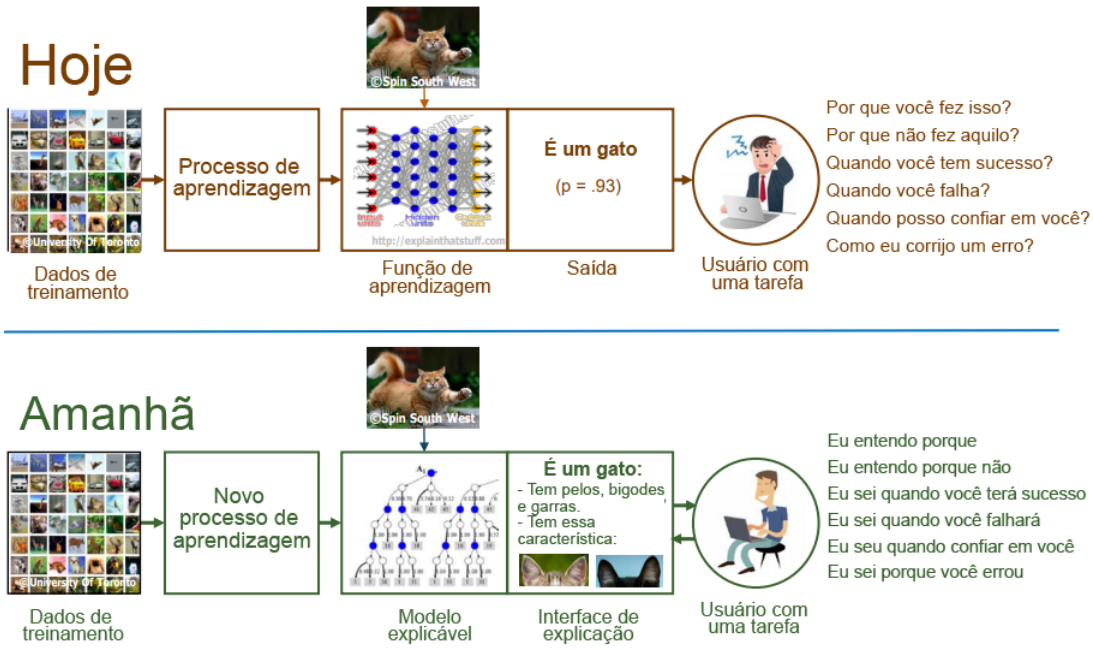
A falta de confiança do ser humano em torno da IA está relacionado a falta de transparência nas decisões desses sistemas. Assim, define-se transparência como sendo o acesso ao funcionamento interno de um modelo de IA (MITTELSTADT; RUSSELL; WACHTER, 2019). Para que um sistema de IA seja confiável, é essencial que o mesmo possua as seguintes características: competência, honestidade e alinhamento. Define-se competência como sendo a capacidade do sistema em realizar tarefas, tomar decisões ou

até mesmo fornecer informações. Já a honestidade está relacionada ao processo que leva a decisão a ser transparente e responsável. Por fim, o alinhamento pode ser definido como sendo a garantia de que o sistema não possui uma motivação oculta (WEBER et al., 2018).

Não existe transparência nos sistemas do tipo caixa-preta, o que os torna de difícil compreensão para o ser humano (KHAN; KHAN et al., 2012). Como estratégia para solucionar esse problema, é que surgiu o termo XAI, cuja finalidade está em reunir técnicas de IA com abordagens compreensíveis e saídas explicáveis ao usuário (FELLOUS et al., 2019). Diante desse contexto , o termo XAI é recente, porém seu conceito data de aproximadamente 40 anos atrás, onde um conjunto de regras era usado para explicar o funcionamento do sistemas especialistas da época (XU et al., 2019).

Esse conceito voltou a ganhar força em novembro de 2017 quando o Departamento de Defesa Americano (DARPA) criou um programa XAI, cuja finalidade é fazer com que as máquinas compreendam o contexto em que se encontram, e por meio disso, criem modelos explicativos que estejam relacionados ao mundo real (GUNNING, 2017). Ainda segundo o programa da DARPA, a XAI é necessária para que o ser humano entenda, confie adequadamente e gerencie com eficácia os sistemas de IA futuros. A Figura 7 demonstra a diferença entre os sistemas de IA atuais, cujo usuário questiona as tarefas realizadas, e o sistemas de IA futuros, cujo usuário terá suas perguntas respondidas por meio das explicações fornecidas. É importante observar que a proposta consiste na utilização de um novo processo de aprendizagem, e de modelo e interface explicativos.

Figura 7 – "O que estamos tentando fazer?".



Fonte - Adaptado de (GUNNING, 2017)

Para Wolf e Ringland (2020), transformar sistemas de IA imensamente complexos

em sistemas compreensíveis e explicativos exige investigar um modelo de IA treinado com muitos pontos de dados fictícios ou de teste (*dataset* de treinamento), usando suas saídas para produzir uma aproximação simplificada do modelo ou entendimento dos limites de decisão do mesmo. Além disso, dois tipos de explicação são apresentados Wolf e Ringland (2020):

- **Global:** cujo objetivo é fornecer uma explicação no nível do modelo. Assim, uma descrição do estado interno para cada classe interna do modelo é feita. Todavia, segundo Adadi e Berrada (2018) a interpretabilidade do modelo global é difícil de ser alcançado na prática; e
- **Local:** sua interpretabilidade é facilmente aplicável e atende as necessidades do ser humano em relação a compreensão do modelo, já que apenas uma parte do modelo é suficiente para a compreensão do todo. Uma explicação local é aquela que fornece uma explicação no nível de entrada. Isto é, como o modelo atribui um determinado rótulo para uma determinada entrada de dados.

Além dos tipos global e local, de acordo com Weber et al. (2018), um modelo XAI pode ser classificado de acordo com as explicações fornecidas, que podem ser faladas ou para serem visualizadas. Entretanto, de forma mais completa, no programa XAI da DARPA é apresentado quatro modos de explicação que facilitam a decisão do usuário (GUNNING, 2017):

- **Declarações analíticas:** são feitas em linguagem natural para descreverem os elementos e o contexto que dão suporte a uma decisão;
- **Visualizações:** que destacam diretamente as partes dos dados que servem de suporte a uma decisão e permitem que os usuários formem sua própria compreensão perceptiva;
- **Casos:** utilização de exemplos ou histórias específicas que auxiliam na tomada de decisão; e
- **Rejeições de escolhas alternativas:** argumentação contra determinadas respostas tendenciosas com base em análises, casos e dados.

Das classificações apresentadas, foi possível observar que a Visualização da Informação está presente em duas delas. Dada essa forte presença, bem como uma possível conversão de declarações analíticas em visualizações, essa área pode ser utilizada como ferramenta para gerar explicações para os usuários de um sistema IA (ELER et al., 2019).

2.2.1 Grad-CAM

Na tentativa de compreender melhor as CNNs, vários métodos surgiram na literatura para visualização da representação interna desse tipo de rede. Uma das abordagens criadas foram os Mapas de Ativação de Classe (CAM) (ZHOU et al., 2016), cujo objetivo é utilizar GAP nas camadas da CNN. Assim, esse método é baseado na utilização da GAP no mapa de ativação da última camada convolucional da rede, antes da FC. Dessa forma, CAM combina os mapas de ativação A da camada de convolução, que contém K filtros, além dos pesos $w_{k,c}$ de FC, cujos valores *(k,c)* representam a conexão ponderada específica entre a camada de convolução e a camada totalmente conectada utilizada para criar a pontuação do mapa de relevância, conforme a Equação 2.5 (RAS et al., 2022).

$$map_c = \sum_{k}^{K} w_{k,c} A_k. \tag{2.5}$$

O método Grad-CAM é uma variação do método CAM usando os gradientes da saída da rede em relação à última camada convolucional da CNN, a fim de obter o mapa de ativação de classe (SELVARAJU et al., 2017). Com isso, é possível que o Grad-CAM seja implementado em mais tipos de CNNs em comparação ao método CAM, exigindo apenas que a função de ativação final usada para previsão de rede seja uma função diferenciável, como por exemplo, a função softmax. Dessa forma, para cada mapa de característica na camada de convolução final, um gradiente da pontuação y_c (*logit*) da classe em relação a todo nó em A_k é calculado para obter uma pontuação de importância $\alpha_{k,c}$ para o mapa de características A_k. A partir disso, o método Grad-CAM combina linearmente essas pontuações de importância e as passa pela função ReLU para obter um mapa de relevância, conforme a Equação 2.6.

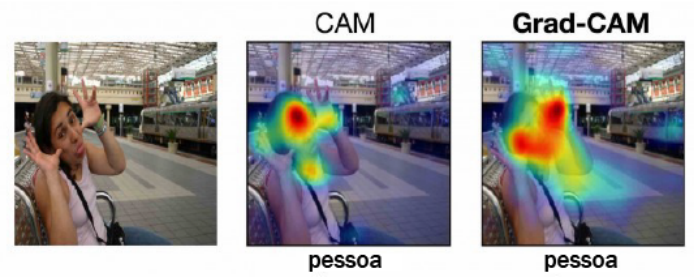
$$map_c = ReLU(\sum_{K}^{k} \alpha_{k,c} A_k). \tag{2.6}$$

A principal diferença entre CAM e Grad-CAM está na forma de gerar os pesos para os mapas de características. No CAM, os mapas de calor são gerados calculando a soma ponderada das ativações da última camada convolucional, a partir dos pesos da camada FC. Já no Grad-CAM, os gradientes de qualquer camada são usados para gerar esses pesos (MOHAMED; SIRLANTZIS; HOWELLS, 2022). Essa diferença é melhor ilustrada na Figura 8, cujos métodos foram aplicados para a classe pessoa e implementados em uma imagem.

2.2.2 Layer Grad-CAM

Quando o método Grad-CAM é aplicado a uma determinada camada, essa abordagem é chamada de Layer Grad-CAM (SELVARAJU et al., 2020). Portanto, segundo

Figura 8 – CAM x Grad-CAM.



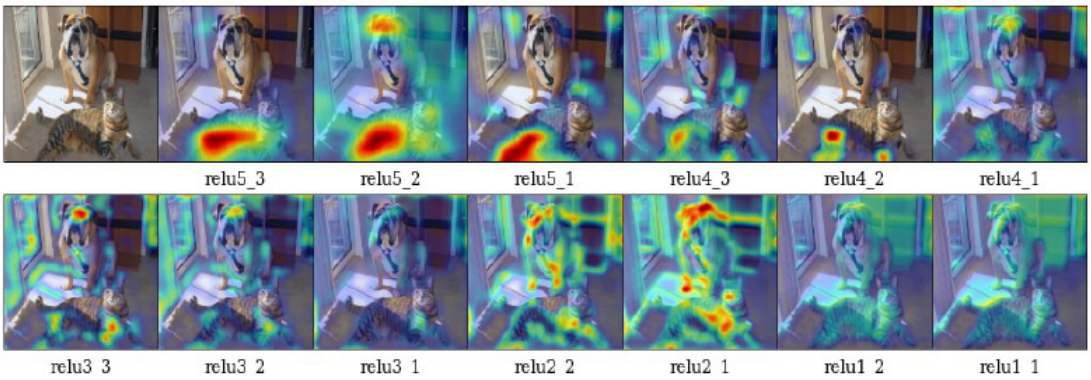
Fonte - Adaptado de (SELVARAJU et al., 2017)

(CHATTERJEE et al., 2022), os gradientes da saída são calculados em relação à camada escolhida cujos gradientes resultantes são calculados em média para cada canal de saída. Assim, o gradiente médio de cada canal é multiplicado pelas ativações da camada e em seguida, os resultados são somados em todos os canais. Para os pesos w referente a características da classe c, o GAP é realizado sobre os mapas de recursos A, de acordo com a Equação 2.7.

$$S^c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{i,j}^k. \tag{2.7}$$

A Figura 9 ilustra o comportamento do método Grad-CAM em relação aos diferentes mapa de características das camadas convolucionais de uma CNN para a classificação de "tigres"ou "gatos". É possível observar, inclusive, como as localizações mudam qualitativamente e que as melhores visualizações são frequentemente obtidas após a convolução mais profunda (SELVARAJU et al., 2020).

Figura 9 – Grad-CAM aplicado aos diferentes mapas de característica das camadas de uma CNN.



Fonte - Adaptado de (SELVARAJU et al., 2020)

2.2.3 Mapa de Saliência

Mapas de Saliência ou Mapas de Calor é um método de visualização comumente utilizado para explicar o processo de predição de uma rede. Dessa forma, muitas das vezes um mapa de calor é sobreposto à imagem de entrada original para que assim seja gerada a visualização. Além disso, esses mapas identificam características da entrada que são mais relevantes (salientes). Ou seja, destaca regiões que estimulam o máximo a rede, de forma a influenciar a saída do modelo (MOHAMED; SIRLANTZIS; HOWELLS, 2022).

Segundo Simonyan, Vedaldi e Zisserman (2013), a extração da saliência da classe ocorre dada uma imagem I_0 (com m linhas e n colunas) e uma classe c. Assim, o Mapa de Saliência da classe M é calculado a partir da derivada w que é encontrada por retropropagação. Depois disso, o Mapa de Saliência é obtido reorganizando os elementos do vetor w. Caso a imagem esteja em escala de cinza, o número de elementos em w é igual ao número de pixels em I_0, possibilitando que o mapa seja calculado como na Equação 2.8, cujo valor $h(i, j)$ equivale ao índice do elemento de w correspondente ao pixel da imagem na i-ésima linha e j-ésima coluna.

$$M_{i,j} = |w_{h(i,j)}|. \tag{2.8}$$

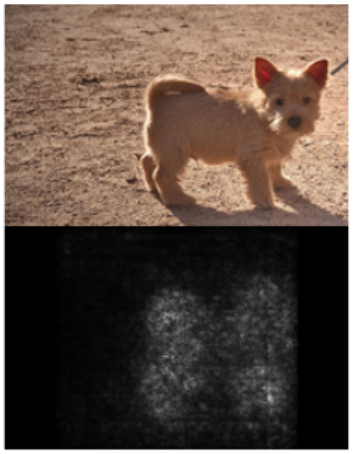
No caso da imagem colorida, assume-se, por exemplo, que o canal de cor c do pixel (i, j) da imagem I corresponde ao elemento de w com o índice $h(i, j, c)$. Dessa forma, para derivar um valor de saliência de classe exclusivo para cada pixel (i, j), é necessário alcançar a magnitude máxima de w em todos os canais de cores, conforme Equação 2.9. A Figura 10 ilustra um mapa extraído a partir de uma única passagem de retropropagação, por meio de uma classificação utilizando uma ConvNet. Assim, é possível observar que são destacadas as regiões equivalentes ao cachorro, conforme o treinamento da rede.

$$M_{i,j} = max_c|w_{h(i,j,c)}|. \tag{2.9}$$

2.2.4 CNN Filters

De acordo (ERHAN et al., 2009), uma forma qualitativa simples e comum de comparar recursos extraídos por uma primeira camada de uma CNN é observar os filtros aprendidos pelo modelo. Ou seja, é necessário analisar os pesos lineares na matriz de pesos representados no espaço da entrada. Assim, essa abordagem torna-se ainda mais importante quando as entradas são imagens que possam ser visualizadas. Dessa forma, esses filtros podem assumir a forma de detectores de traços, quando treinados em dados de dígitos, ou detectores de borda quando treinados em fragmentos de imagens naturais.

Figura 10 – Exemplo de Mapa de Saliência.



Fonte - Adaptado de (SIMONYAN; VEDALDI; ZISSERMAN, 2013)

De acordo com Hochuli et al. (2018), as camadas convolucionais de uma CNN consistem em filtros que são aplicados à entrada da camada e que produzem mapas que representam a presença de características locais específicas da entrada, e que foram aprendidas pela rede. Como os filtros são funções lineares da entrada, os pesos w_f do filtro e uma polarização aditiva b_f, determinam a contribuição da entrada x para a função de ativação σ, cuja a saída é representada pela Equação 2.10.

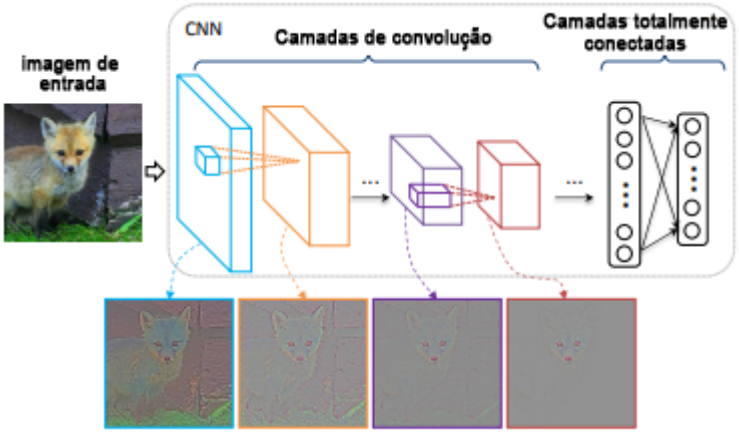
$$output_f = \sigma(w_f x + b_f). \tag{2.10}$$

Partindo do pressuposto de que uma unidade pode ser caracterizada pelos filtros da camada anterior à qual está mais fortemente conectada, ao obter uma combinação linear ponderada dos filtros da camada anterior, é possível observar que uma rede com restrições de expansão nas ativações, treinadas em imagens naturais, tenderá a aprender os chamados detectores de canto na segunda camada (ERHAN et al., 2009).

CNN Filters é uma técnica simples e eficiente, entretanto, é importante observar que existe uma ligação entre as atualizações do gradiente para maximizar a ativação de uma unidade e encontrar a combinação linear de pesos conforme descrito por Lee et al. (2009). Além disso, a visualização pode variar por camada, como por exemplo, as unidades da camada superior que representam características mais complexas, e que correspondem a combinações de características das camadas inferiores (ERHAN et al., 2009), conforme demonstrado na Figura 11.

Somado a isso, padrões multifacetados podem ser reconhecidos pelos filtros nas camadas intermediárias. Por fim, cada filtro em uma camada, exceto a primeira, define uma função não convexa com múltiplos máximos locais no espaço da imagem. Isso demonstra que a visualização é uma ferramenta útil para interpretar a caixa-preta das CNNs (XIE;

Figura 11 – Filtros de uma CNN.



Fonte - Adaptado de (YU et al., 2014)

YANG; LAI, 2019).

2.3 Segmentação de imagens na Aquicultura

Das espécies de peixe comuns no Brasil, o Pacu (cientificamente denominado *Piaractus mesopotamicus*) é um peixe de água doce e que pertence à mesma família das piranhas. Se alimenta, na maioria das vezes, de frutos e sementes, e é comumente encontrado em alguns países da América do Sul (FREITAS et al., 2021). Segundo FAO et al. (2016), a produção da espécie está em constante crescimento, inclusive em países como a China e o Vietnã. A Aquicultura destinada ao Pacu depende de sistemas distintos para produção e manutenção dessa produção. Com isso, determinar a relação de uma mesma característica em ambientes diferentes torna-se essencial para estabelecer melhorias. Com isso, o tamanho do peixe, a qualidade da carne e outras características tradicionais não são mais determinantes para o consumidor na hora da compra (FREITAS et al., 2021).

Existem dois tipos de morfotipos a serem considerados na Aquicultura: elíptico e arredondado. Na Truta arco-íris, por exemplo, o ganho de massa do peixe ocorre pelo aumento da largura e altura do corpo, tornando o peixe mais arredondado do que os demais (SAE-LIM et al., 2013). Para Freitas et al. (2021) o Pacu de morfotipo elíptico é indicado para quem pretende comprar o peixe inteiro. Entretanto, para consumo de subprodutos processados, como por exemplo o lombo, o morfotipo arredondado é o mais aconselhado. Apesar de ser uma característica importante, essa distinção nem sempre é algo simples de ser realizada.

A medição das características do animais é comumente realizada por meio da análise de imagens digitais, cujo objetivo é auxiliar no melhoramento do peixe e de suas medidas a partir de imagens obtidas manualmente. Esse processo pode ser demorado

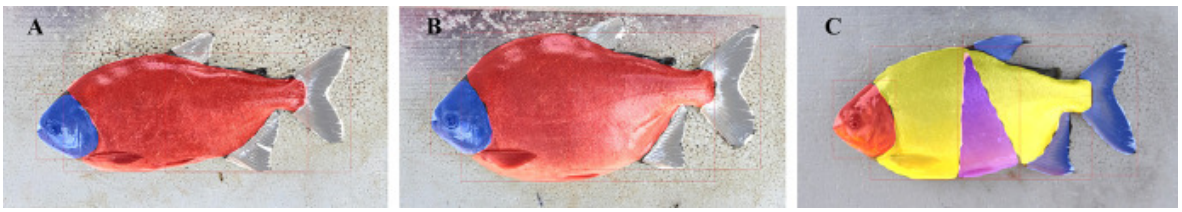
e exaustivo para o animal, principalmente quando implementado de forma rotineira (CARDOSO et al., 2021).

Com base nesse cenário, modelos de IA são utilizados para tornar essa indústria mais inteligente (YANG et al., 2021), permitindo inúmeras aplicações como por exemplo, a classificação de espécies e o reconhecimento dos animais (ZHAO et al., 2021). Além disso, o avanço na produção aquícola tem contribuído significativamente na busca por novas tecnologias, cujo objetivo é auxiliar na produção de peixes de forma sustentável e eficaz (FREITAS et al., 2023).

Segundo Fernandes et al. (2020), a busca por alternativas rápidas e não invasivas para medição de características dos peixes também tem sido crucial para a implementação de modelos de IA na área. Os autores inclusive utilizam da segmentação de imagens automatizada para extrair as medidas biométricas, a predição do peso corporal, do peso da carcaça e do rendimento do filé de peixes da espécie Tilápia do Nilo. A partir disso, foi criado o primeiro Sistema de Visão Computacional (*Computer Vision System* - CVS) para peixes vivos (FERNANDES et al., 2020).

Para a espécie de peixe Pacu, Freitas et al. (2023) apresentam um CVS que contém a combinação de estratégias de DL para segmentação das imagens dos animais. Dessa forma, uma Mask R-CNN foi utilizada para selecionar as regiões (corpo, cabeça, nadadeiras e pelve) e também para extrair as características da imagem original. A Figura 12 apresenta a segmentação da cabeça e corpo realizada para um peixe de morfotipo elíptico (A) e morfotipo arredondado (B). Também é possível observar todas as regiões de interesse segmentadas ao mesmo tempo (C).

Figura 12 – Segmentação de instâncias por aprendizado profundo das regiões do corpo.



Fonte - (FREITAS et al., 2023)

Em seguida, foi realizada a classificação de acordo com a etapa de treinamento realizada anteriormente. Com isso, as características do peixe podem ser medidas rotineiramente, sem causar grandes danos ao animal (FREITAS et al., 2023). Os resultados obtidos demonstram que criar CVS a partir de técnicas de IA é eficiente para estimar medidas morfométricas do Pacu, principalmente por conta da resiliência apresentada diante de diferentes condições de iluminação e *backgroud* da imagem (FREITAS et al., 2023). Apesar disso, de acordo com (FERNANDES et al., 2020), os modelos de IA utilizados podem ser considerados complexos, mesmo sendo de última geração.

2.4 Perturbação de pixels

De acordo com Szegedy et al. (2013), a estabilidade das redes neurais pode ser verificada quando as entradas são submetidas a perturbações, gerando as chamadas amostras adversárias. Dessa forma, em um contexto de explicação, essa abordagem pode ser usada para identificar quais partes da entrada levam a uma correta classificação, por exemplo. Além disso, é possível compreender a relação entre a entrada e a saída durante o treinamento da rede, a partir das classificações incorretas geradas.

A utilização das amostras adversárias expandiu-se de acordo com a evolução das redes neurais e suas aplicações. Com isso, a aprendizagem profunda adversária ganhou destaque na comunidade científica a fim de explorar a relação dos ataques adversários com os resultados fornecidos pelos modelos de IA existentes (PAPERNOT et al., 2016). Além disso, diversos métodos surgiram no intuito de verificar os diferentes tipos de entrada, sendo alguns deles não distinguíveis pelo ser humano (NGUYEN; YOSINSKI; CLUNE, 2015).

Diante do contexto apresentado, as perturbações podem ser ambientais, cujas transformações alteram a aparência do objeto em relação ao ambiente real. Dessa forma, tais transformações podem ser naturalmente obtidas a partir de precipitações e grafites, por exemplo. Em contrapartida, as perturbações também podem ser digitais, cujas alterações ocorrem a partir de imperfeições no *hardware*, na variação das configurações da câmera durante a captura, ou na edição da imagem capturada (STOCK; DOLAN; CAVEY, 2020).

Durante o processo de perturbação digital de uma imagem, é possível limitar as regiões e os pixels da imagem cuja transformação será aplicada, de acordo com a localização do objeto a ser detectado e/ou classificado (DONG et al., 2020). Além disso, dentre as técnicas de perturbação digital existentes, é possível alterar apenas um único pixel da imagem, bem como introduzir ruído ou desfocar a imagem com o objetivo de alterar a aparência do objeto (STOCK; DOLAN; CAVEY, 2020).

Segundo Gorokhovatskyi e Peredrii (2020), a construção de imagens perturbadas pode ser realizada a partir da alteração do pixel para uma cor determinada. Esse processo pode ser observado, inclusive, no método explicável LIME (RIBEIRO; SINGH; GUESTRIN, 2016), cujo objetivo é realizar o treinamento a partir das imagens perturbadas. Finalmente, é possível observar a utilização dos métodos de perturbação de pixels para identificar a parte da imagem que possui a maior responsabilidade em relação a decisão do processo de classificação. Em Fong e Vedaldi (2017), os autores usufruem de três diferentes métodos de perturbação para execução dessa tarefa: desfoque da imagem, substituição dos pixels por uma constante e aplicação de ruído (Figura 13).

Atualmente, a importância da perturbação de pixels é observada nos métodos explicáveis existentes, cuja transformação da imagem é vista tanto como ferramenta

Figura 13 – Métodos de perturbação de pixels

Fonte - Adaptado de (FONG; VEDALDI, 2017)

avaliativa, quanto como base para a explicação dos modelos de classificação e segmentação de imagens (GOROKHOVATSKYI; PEREDRII, 2020).

2.5 Considerações finais

Conceitos de IA estão cada vez mais presentes na sociedade, inclusive em áreas biológicos, como a Aquicultura. Apesar disso, sistemas XAI ainda são uma novidade para o ser humano, apesar de se mostrarem necessários, já que a busca por explicações em sistemas de IA deixou de ser uma característica adicional para tornar-se essencial. O avanço tecnológico trouxe benefícios para os usuários, porém também trouxe preocupações relacionadas às decisões que uma máquina pode tomar. Além disso, a grande quantidade de dados envolvidos nessas decisões trazem preocupação em relação a segurança dos usuários e seus respectivos dados. Assim, sistemas XAI surgiram com a intenção de sanar essas preocupações.

Apesar da grande expectativa gerada em torno da explicabilidade, a natureza dessas explicações também tem sido questionada. Com isso, dos métodos de XAI existentes, busca-se avaliar a qualidade dos resultados e também abordagens que tragam melhorias às explicações geradas. Assim, para avaliação desses métodos, a perturbação de pixels é considerada devido a sua ampla utilização e fácil aplicação no contexto de imagens digitais.

Para a presente tese, foram considerados apenas quatro métodos explicáveis, conforme descritos anteriormente: Grad-CAM, Layer Grad-CAM, Mapa de Saliência e CNN Filters. Todos esses métodos geram visualizações como resultado e são compatíveis com a Mask R-CNN combinada com uma ResNet-18, conforme utilizado nos experimentos descritos a seguir.

Nos demais Capítulos deste estudo, são aplicados os conceitos obtidos dos resultados da RSL para as questões de pesquisa apresentadas. A partir desses conceitos, dois experimentos foram conduzidos para determinar a análise e o aperfeiçoamento dos métodos de XAI, de acordo com as hipóteses de pesquisa.

3 Revisão Sistemática Literatura

Neste Capítulo serão apresentados os resultados obtidos por meio da Revisão Sistemática da Literatura em relação às questões de pesquisa apresentadas. Segundo Kitchenham e Charters (2007), a RSL pode ser definida como sendo um tipo de estudo secundário cujo objetivo é identificar, avaliar e interpretar trabalhos e estudos relevantes relacionados a um item de pesquisa. Dessa forma, para a presente tese, a RSL apresenta resultados relevantes que auxiliam na compreensão do tema proposto. Assim, os motivos pelos quais a RSL foi considerada neste trabalho são:

- Resumir evidências já existentes relacionadas às hipóteses e às questões de pesquisa apresentadas;
- Identificar lacunas no estado da arte a fim de sugerir uma pesquisa inédita; e
- Fornecer conteúdo base para a pesquisa apresentada.

A RSL também proporciona vantagens significativas para este presente estudo, como por exemplo, resultados consistentes, evidências mais confiáveis e garantia de maior conhecimento sobre determinado assunto (KITCHENHAM; CHARTERS, 2007). Com base nisso, serão apresentadas a seguir as etapas (planejamento, condução da revisão e documentação) que compõem uma RSL, cujo objetivo é auxiliar no processo de avaliação dos documentos obtidos durante o processo.

Para a etapa de planejamento, as questões de pesquisa foram determinadas de acordo com o modelo PICOC (População, Intervenção, Comparação, Saídas e Contexto) (SAMPAIO, 2015), cujo objetivo é determinar termos bases para o processo de revisão. Assim, os parâmetros do modelo para a presente tese foram definidos da seguinte forma:

- **População:** inteligência artificial explicável (XAI);
- **Intervenção:** conceitos, metodologias e tecnologias;
- **Comparação:** perturbação de pixels, perturbação;
- **Saídas:** explicabilidade, interpretabilidade; e
- **Contexto:** inteligência artificial (IA).

Todos os parâmetros foram utilizados em inglês para que a pesquisa por documentos alcançasse maior número de resultados, visto que os documentos mais relevantes

são escritos na língua inglesa. Além disso, de acordo com o PICOC determinado, foram elaboradas as questões de pesquisa da presente tese, conforme já apresentadas. São elas:

- *Q1: Como avaliar se os métodos de XAI explicam o que de fato aconteceu na predição do modelo de IA?*
- *Q2: Os métodos de XAI são influenciados pela perturbação de pixels das imagens de entrada do modelo?*
- *Q3: Como melhorar métodos de XAI para apresentarem explicações condizentes com a predição do modelo de IA?*

Conforme a condução da RSL, as questões têm por objetivo compreender as atuais técnicas de avaliação para métodos explicáveis, tal como a utilização da perturbação de pixels para análise da explicabilidade desses métodos, além de maneiras existentes para melhorar as explicações geradas. A partir disso, foram determinadas as palavras-chaves e a *string* utilizada para realizar a busca por estudos relacionados nas seguintes bibliotecas digitais: *IEEE Xplore, ACM Digital Library, ScienceDirect* e *Scopus*. Dessa forma, para auxiliar na pesquisa, foi utilizada a ferramenta *on-line* Persifal que foi desenvolvida para dar suporte às Revisões Sistemáticas nas áreas relacionadas a Engenharia de Software (PARSIFAL, 2020). Por meio dessa ferramenta, as palavras-chaves foram organizadas para compor possíveis *strings* de busca. Assim, das *strings* elaboradas, a que retornou uma quantidade considerável de estudos foi a apresentada a seguir:

("explainable artificial intelligence") AND ("concepts"OR "methodologies"OR "techniques") AND ("pixels perturbation"OR "perturbation") AND ("explicability"OR "interpretability") AND ("artificial intelligence")

Em relação aos métodos de avaliação dos documentos, foram realizadas duas etapas de acordo Kitchenham e Charters (2007). Assim, para a primeira etapa foi realizada a leitura dos títulos e resumos dos documentos selecionados a partir dos objetivos determinados anteriormente. Dessa forma, os documentos que não foram excluídos conforme os critérios estabelecidos, permaneceram para a etapa seguinte. Já a segunda etapa consiste na leitura completa dos documentos, com o objetivo de verificar o contexto no qual eram empregados os termos-chaves. A partir dessa leitura, foi determinado se o documento deveria ser excluído ou incluído, segundo critérios estabelecidos durante o planejamento. Assim, foram incluídos os documentos que respondem pelo menos uma das questões de pesquisa estabelecidas. Para a exclusão de documentos, foram considerados os seguintes critérios:

- Documentos publicados antes de 2019;
- O documento é um tutorial, sumário de evento, pôster ou qualquer outro documento incompleto;

- O documento apresenta título, *abstract* e/ou palavras-chaves que não condizem com o tema pesquisado;
- O documento é duplicado; e
- O documento não está disponível para *download*.

A quantidade de documentos selecionados a partir da *string* de pesquisa é apresentada na Tabela 1. Com isso, é possível observar que apenas 31 documentos respondem a pelo menos uma das questões criadas para essa tese. Esses documentos foram lidos e os principais pontos foram extraídos como resultados da RSL.

Tabela 1 – Quantidade de documentos selecionados durante a RSL.

Base científica	Documentos selecionados
IEEE Xplore	3
ACM Digital Library	63
ScienceDirect	243
Scopus	13
TOTAL	322
INCLUÍDOS (1ª etapa)	53
INCLUÍDOS (2ª etapa)	31

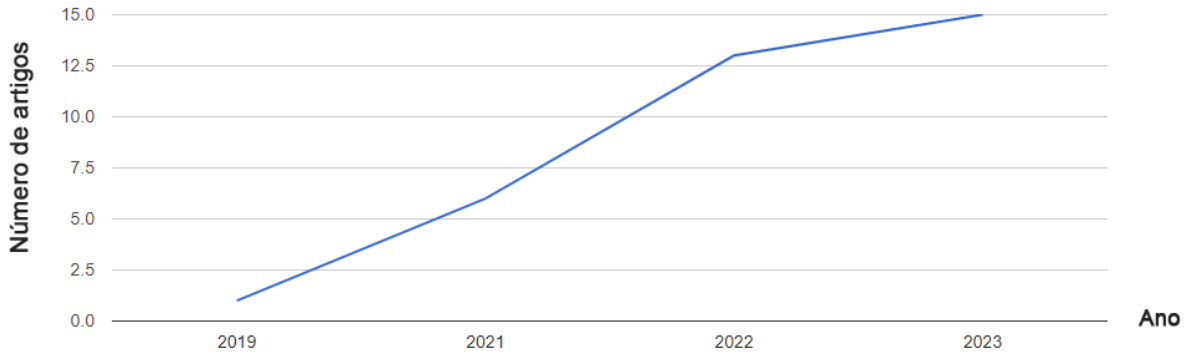
Fonte – Elaborada pela autora.

O gráfico ilustrado pela Figura 14 mostra a quantidade de artigos publicados ao longo dos últimos cinco anos, com base na *string* de pesquisa empregada. Dessa forma, é possível observar o crescimento no número de publicações a partir do ano de 2019. Essa crescente tem como base o início discreto da XAI, e o aumento do interesse pela área nos anos seguintes. Além disso, as pesquisas passaram de áreas mais específicas (e.g., Computação), para temáticas mais abrangentes, como por exemplo, a busca por métodos de avaliação de XAI para segmentação de imagens médicas. Assim, baseado nessa evolução, os resultados dessa revisão são apresentados na seção a seguir.

3.1 Resultados

Os resultados obtidos durante o processo de revisão são apresentados de acordo com às questões de pesquisa investigadas. Dessa forma, a principal análise realizada em cada um dos documentos foi baseada na quantidade de questões respondidas ao longo da pesquisa realizada pelos autores. Assim, a descrição dos documentos e seus respectivos resultados são apresentados a seguir.

Figura 14 – Gráfico de artigos da revisão por ano.



Elaborada pela autora.

3.1.1 Q1: Como avaliar se os métodos de XAI explicam o que de fato aconteceu na predição do modelo de IA?

Muito se fala da busca pela confiança em modelos de IA. Todavia, a necessidade em alinhar as explicações dos métodos de XAI com a explicação do ser humano demonstra a preocupação em também aumentar a confiança nos métodos de IA explicável (DÍAZ-RODRÍGUEZ et al., 2022). Com base nisso, a busca por métodos de avaliação da explicabilidade cresceu significativamente nos últimos anos, inclusive em cenários cujas explicações devem passar total confiança (e.g análise de imagens médicas) (SALEEM; SHAHID; RAZA, 2021) (JIN et al., 2023) (SALAHUDDIN et al., 2022). Dos documentos analisados nessa RSL, a maioria apresentou formas de avaliar a explicabilidade dos métodos de XAI. Assim, a preocupação em alcançar boas explicações se estendeu para outros áreas, proporcionando métodos avaliativos multidisciplinares (MOHSENI; ZAREI; RAGAN, 2021), incluindo campos como Psicologia, Filosofia e Ciências Sociais (VILONE; LONGO, 2021).

De acordo com Vieira e Digiampietri (2022), avaliar a explicabilidade depende de dimensões que determinam a qualidade da mesma. Essas dimensões podem ser descritas da seguinte forma:

- **Fidelidade:** refere-se a quão bem as explicações geradas se aproximam da predição do modelo de IA;
- **Compreensibilidade:** até que ponto as explicações são humanamente compreensíveis;
- **Robustez:** modelos robustos resistem a perturbação da entrada, ou seja, a explicação só é feita com base nos pontos de maior relevância, e não nos pontos perturbados; e
- **Complexidade:** descreve a complexidade computacional do método explicável.

Além das métricas computacionais, existem as métricas centradas no ser humano, cujo objetivo é avaliar a qualidade da explicação a partir da confiança de pessoas leigas no assunto. Com isso, pode-se classificar essas métricas em dois tipos: métricas subjetivas, que mediam a confiança humana nas explicações; e métricas objetivas, que mediam o estado comportamental dos humanos e o desempenho das tarefas (IBRAHIM; SHAFIQ, 2023).

Com base em diferentes critérios e perspectivas, Ding et al. (2022) apresentam uma tabela que ilustra diferentes métodos e métricas para avaliação da explicabilidade. A categorização foi realizada a partir do aspecto da avaliação, nas medidas utilizadas e no método de construção, permitindo uma visão geral das soluções avaliativas aplicadas a IA explicável. Dessa forma, o aspecto da avaliação pode ser classificada em:

- **Modelos mentais:** baseados em princípios da Psicologia racional, define como os indivíduos humanos interpretam as decisões da IA;
- **Eficácia e Satisfação:** análise da correlação entre a satisfação do usuário e a eficácia do método;
- **Confiança e Dependência:** a confiabilidade humana em modelos de IA pode ser considerada um aspecto importante que impacta experiências negativas ou positivas do sistema subjacentes;
- **Desempenho humano-IA:** análise do desempenho humano em suas tarefas relacionadas à sistemas de IA; e
- **Avaliação funcional:** avalia a aceitabilidade e abrangência das explicações geradas, por meio de medidas e métricas computacionais, como por exemplo, a perturbação da entrada do modelo.

Atualmente, todas os órgãos regulatórias voltados a IA concordam com a necessidade de avaliar cuidadosamente a qualidade das explicações automatizadas (ALI et al., 2023) (LISBOA et al., 2023). Assim, os resultados apresentados para essa questão sustentam a importância da avaliação dos métodos de XAI aplicados a essa presente pesquisa.

3.1.2 Q2: Os métodos de XAI são influenciados pela perturbação de pixels das imagens de entrada do modelo?

As perturbações permitem examinar a relação entre a entrada e saída de um modelo, permitindo observar qual parte da entrada um modelo atribui maior importância (IVANOVS; KADIKIS; OZOLS, 2021). A relação entre métodos explicáveis e as diferentes técnicas de perturbação pode ser descrita de várias formas. A primeira delas é por meio

da classificação dos métodos de XAI, que podem ser divididos em dois tipos: baseados em gradiente e baseados em perturbação. Dessa forma, os métodos baseados em perturbação, como por exemplo o Occlusion, possuem como principal característica a perturbação da entrada para definir se o modelo de IA identifica, durante o processo de inferência, os pontos de maior relevância (RAS et al., 2022). Com isso, grande parte dos trabalhos analisados nessa revisão consideram a perturbação de pixel como base de um método explicável.

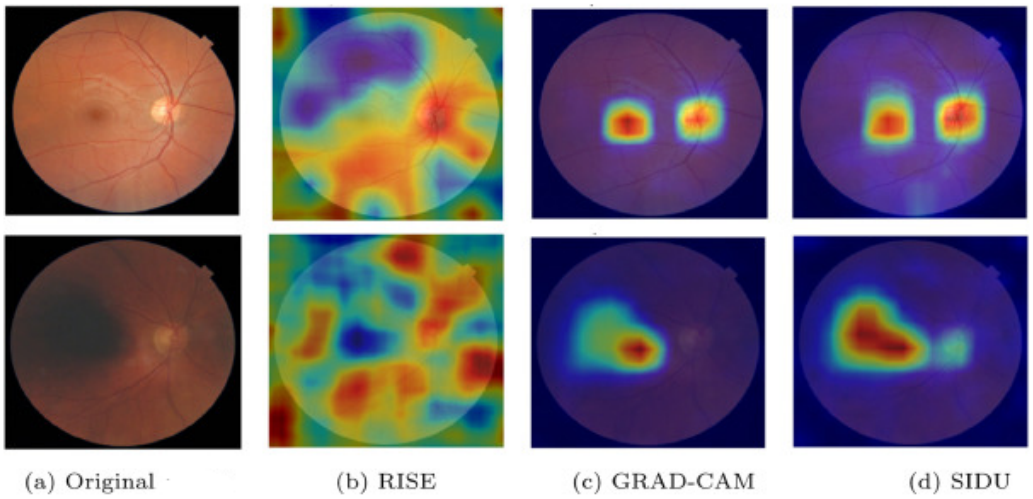
Outra alternativa foi encontrada em trabalhos que utilizam a perturbação da entrada como forma de avaliação, tanto da precisão dos modelos de IA (KADIR et al., 2023), como da explicabilidade dos métodos explicáveis (MOHAMED; SIRLANTZIS; HOWELLS, 2022). Em ambos os casos, quando há influência pelas alterações realizadas na entrada, o bom desempenho da ferramenta é comprovado. No caso dos métodos explicáveis, quando essa influência não é comprovada, é possível afirmar que as explicações geradas não condizem com a realidade do processo de inferência do modelo. A exemplo, para avaliar uma explicação em termos de descrição do comportamento do modelo, existe um método que substitui pixels ou regiões de pixels com base no processo MoRF (do inglês *Most Relevant Firts*), cuja substituição é feita em ordem decrescente com base na sua relevância média (GUMPFER et al., 2023).

A utilização da técnica de perturbação se estende para outros cenários que não seja classificação (LIN; LEE; CELIK, 2021) ou segmentação de imagens (GIPIŠKIS et al., 2023). Nos trabalhos realizados por Schlegel et al. (2019), Veerappa et al. (2022) e Abanda, Mori e Lozano (2022), por exemplo, a técnica é considerada para a entrada de modelos de séries temporais, sendo utilizada inclusive como forma de avaliar a explicabilidade dos métodos de XAI aplicadas a esse tipo de modelo. Dessa forma, é possível observar que a perturbação da entrada é uma técnica comumente usada e auxilia na detecção de falhas na explicabilidade, ou até mesmo no próprio modelo de IA utilizado.

Para modelos que utilizam imagem como entrada, a técnica de perturbação permite verificar diferentes tipos de ruídos em diferentes tipos de imagens, como por exemplo em Shi, Li e Yamaguchi (2023). Em cenários reais, tal como imagens médicas, esses ruídos podem ser gerados de diferentes formas durante a captura da imagem. A Figura 15 ilustra duas imagens de fundo de um olho sendo explicadas por três métodos visuais diferentes: RISE, Grad-CAM e SIDU (método desenvolvido por Muddamsetty et al. (2022)). Diante disso, é possível observar que a segunda imagem apresenta má qualidade devido a sombra gerada durante a captura da mesma. Isso fez com que os métodos explicáveis fossem imprecisos ao destacarem o disco óptico do olho. Sendo assim, muitas dessas situações não são consideradas durante o treinamento, permitindo que a predição do modelo seja realizada de forma errônea.

As técnicas de perturbação podem ajudar a detectar se o processo de predição

Figura 15 – Explicação visual de imagens de fundo de olho com e sem ruído.



Fonte - (MUDDAMSETTY et al., 2022)

será feito corretamente diante dessas alterações. Além disso, quando aplicados métodos explicáveis, o resultado apresentado deve ser de acordo com a perturbação dos pixels mais relevantes, conforme observado na figura apresentada. Ou seja, quando esses são modificados, a explicabilidade deve ser influenciada. Portanto, de acordo com as respostas encontradas para essa questão de pesquisa, pode-se afirmar que o método explicável com bom desempenho deve sim sofrer influência em relação a perturbação realizada na entrada do modelo de IA, inclusive para imagens e explicações visuais.

3.1.3 Q3: Como melhorar métodos de XAI para apresentarem explicações condizentes com a predição do modelo de IA?

A última questão de pesquisa investigada na revisão é referente a implementação de melhorias nos métodos de XAI, cujo objetivo é melhorar a qualidade da explicabilidade. Essa questão é a mais importante dessa presente tese e por isso, as respostas obtidas ajudaram a distinguir a diferença entre a metodologia aplicada e o que foi encontrado na literatura. Além disso, a busca por explicações mais confiáveis tem motivado pesquisadores de diversas áreas, como por exemplo Medicina, a buscarem por técnicas que melhorem as explicações fornecidas pelos métodos de XAI (SHOJAEI; ABADEH; MOMENI, 2023). Assim, apesar da questão ser considerada pertinente diante do cenário atual, apenas nove artigos dos 31 analisados apresentaram sugestões de melhorias.

Das soluções apresentadas, a mais comum encontrada está relacionada a modificações consideráveis nos métodos de XAI já existentes, conforme apresentado por Kucklick e Müller (2023). Nesse caso o método Grad-RAM foi criado para fornecer maior precisão em suas explicações por meio da utilização de Mapas de Ativação de Regressão (RAM)

ao invés de Mapas de Ativação de Classe (CAM), como ocorre no Grad-CAM. Outro exemplo de modificação de método de XAI é apresentado por Zafar e Khan (2021), cujo objetivo por trás do DLIME, método desenvolvido pelos autores, é utilizar o *Agglomerative Hierarchical Clustering* (HAC) por meio do KNN (K-*Nearest Neighbors*) para encontrar agrupamentos de pontos de dados semelhantes a uma instância de teste, ao invés de utilizar a perturbação aleatóriola proposta no método original LIME.

Algumas das pesquisas encontradas demonstram modificações realizadas em mais de um método, conforme ilustrado por Elguendouze et al. (2023). Assim, os autores apresentam uma abordagem de explicabilidade ponta a ponta para arquiteturas de legendagem de imagens baseadas em recursos visuais do tipo *Bottom-Up* (BU). Essa abordagem é combinada com os métodos LIME e LRP, que mesmo distintos em sua metodologia, foram modificados resultando em dois novos métodos denominados BU-LIME e BU-LRP. Outro exemplo de modificação em mais de um método explicável é apresentado por Gumpfer et al. (2023), que propõe uma nova técnica chamada SIGN (*Sign-based Improvement of Gradient-based explaNations*) que pode ser usada como substituta da entrada como fator de multiplicação, com o objetivo de eliminar a correlação e reduzir o viés em métodos como Gradient Input e LRP.

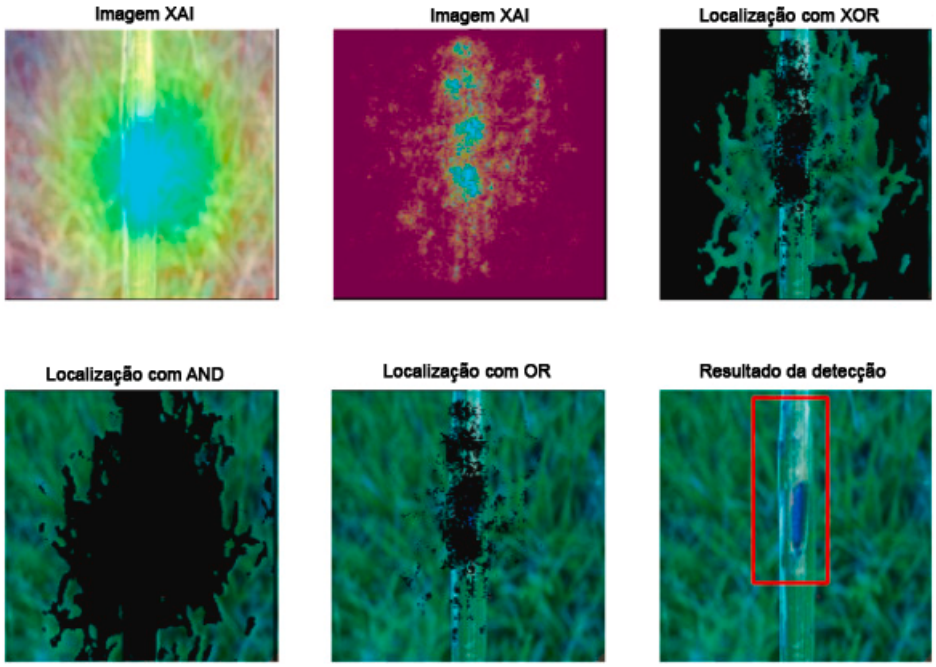
O método explicável LIME aparece em várias pesquisas relacionadas a melhorias da explicabilidade, como apresentado anteriormente. Em outro caso analisado, o método EvEx foi desenvolvido como uma nova IA explicável que utiliza explicações fornecidas pelo LIME e que foram combinadas com um algoritmo genético multiobjetivo (SOUSA; VELLASCO; SILVA, 2022). Em outro exemplo encontrado, os autores propõe um Explicador de Algoritmos Genéticos baseado em Ensemble (EGAE) automático, cuja a função é melhorar a explicabilidade do LIME, eliminando as intervenções do usuário na determinação do número de superpixels na imagem de entrada, bem como os principais recursos para a explicação automática (NEMATZADEH et al., 2023).

Um dos métodos mais modificados na literatura é o método CAM, que apresenta variantes como Grad-CAM e Score-CAM. Em um dos trabalhos analisados, os autores apresentam mais uma variação denominada SA-CAM, que utilizada em sua metodologia o mecanismo de auto-atenção. Ou seja, SA-CAM combina métodos CAM baseados em perturbações e baseados em gradiente por um termo de regularização de auto-atenção para gerar uma interpretação mais robusta (LIANG; LI; JIANG, 2022).

Outra solução de melhoria apresentada é a combinação entres métodos de XAI para melhorar a explicabilidade. Assim Coulibaly et al. (2022), apresentam múltiplos mapas visuais criados a partir da combinação dos métodos Grad-CAM, Integrated Gradient, Gradient Input e Occlusion. Seus desempenhos de visualização são comparados para detectar pragas em áreas rurais e a medição da sobreposição entre pares de imagens foram feitos por meio dos operadores AND, OR e XOR. A Figura 16 apresenta a combinação

dos métodos Grad-CAM e Integrated Gradient.

Figura 16 – Combinação entre métodos de XAI.



Fonte - (COULIBALY et al., 2022)

Os resultados apresentados fortalecem a ideologia de buscar melhorar métodos explicáveis. Além disso, é possível observar que de acordo com as pesquisas analisadas durante a revisão, a maior parte das melhorias sugeridas baseiam-se em modificações pontuais nos métodos já existentes. Entretanto, apenas um documento sugere a combinação entre métodos para melhorar a explicabilidade. Por fim, pode-se afirmar que a metodologia empregada nesta tese condiz com a ideologia proposta para os métodos explicáveis, e a questão proposta foi de fato esclarecida pela RSL.

3.1.4 Considerações finais

Dentre os documentos analisados durante o processo de revisão, apenas Gumpfer et al. (2023) responderam às três questões de pesquisa. Entretanto, apesar das semelhanças, o conjunto de respostas observadas induziu a uma proposta diferente, e portanto não invalidam as hipóteses de pesquisa propostas nesta tese. Grande parte dos documentos apresentaram respostas análogas ou semelhantes ao que foi proposto nesta tese para as questões 1 e 2. Diante disso, é possível observar que avaliar métodos de XAI utilizando a perturbação de pixel é uma abordagem correta e fundamentada pela literatura.

A diferença entre o presente trabalho e os demais documentos é observada na terceira questão de pesquisa que busca técnicas para melhorar os métodos explicáveis. As propostas de melhorias de explicabilidade apresentadas nos nove documentos são baseadas em diferentes abordagens, cujo foco é tornar a explicação mais eficiente. Entretanto,

em nossa tese, o objetivo foi melhorar métodos de XAI a partir da escolha do melhor método diante do cenário proposto, seguido da combinação desse com os demais métodos implementados. Diante disso, o trabalho que mais se assemelhou em relação a proposta apresentada foi elaborado por Coulibaly et al. (2022), que realizou combinações entre métodos de XAI locais e globais para melhorar a explicabilidade. Entretanto, apesar da avaliação feita pelos autores para identificar a qualidade da explicabilidade dos métodos, não foi realizada uma combinação do melhor método com os demais a fim de encontrar um bom resultado explicável. Ou seja, a combinação proposta pelos autores foi feita entre métodos escolhidos para a pesquisa, mas que não foram classificados como melhor ou pior.

Nos próximos capítulos, buscou-se observar se a metodologia proposta nesta tese vai além da melhoria dos métodos avaliados como inferiores. Ou seja, se o próprio método de XAI classificado como o melhor pode ser aperfeiçoado para alcançar melhores resultados.

4 Metodologia

Conforme apresentado no Capítulo 1, a presente tese objetiva analisar os resultados de métodos de XAI diante da perturbação de pixels aplicada às imagens de entrada de um modelo de IA voltado para a segmentação de imagens, a fim de encontrar o melhor método explicável e combiná-lo com outros métodos para melhorá-los. Dessa forma, foram realizados alguns passos para que esse objetivo fosse alcançado. São eles:

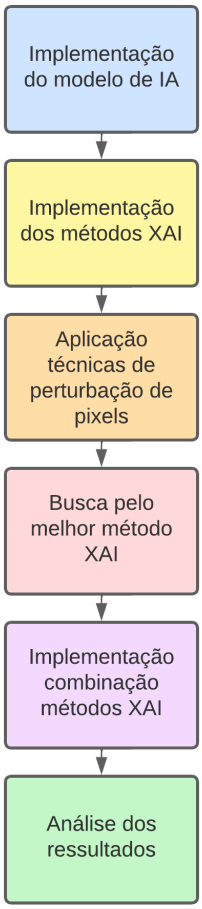
- Implementação do modelo de IA para segmentação de imagens de peixes da espécie Pacu;
- Implementação de quatro métodos de XAI compatíveis com o modelo aplicado;
- Aplicação de três tipos de perturbação de pixels para as imagens de entrada do modelo de IA;
- Busca do melhor método de XAI com base na análise dos resultados obtidos;
- Implementação da combinação entre métodos explicáveis a fim de melhorar os resultados; e
- Análise dos resultados obtidos.

Os passos foram realizados conforme a Figura 17, que demonstra a execução da metodologia dessa tese. As próximas seções desse capítulo destacam os detalhes de cada passo realizado.

4.1 Modelo de IA

O modelo de segmentação utilizado foi criado a partir de uma CNN, mais especificamente a Mask R-CNN utilizando como extrator de características uma variante da ResNet com 18 camadas. Inicialmente, o objetivo do modelo é segmentar regiões de interesse no peixe para fins de fenotipação, e posterior seleção genética. Dessa forma, a cada imagem de peixe utilizada como entrada, o modelo gera como saída as classes (cabeça, nadadeiras, corpo e *background*), as máscaras de cada região do peixe, uma caixa delimitadora e a pontuação de confiança na predição da classe, sendo a última utilizada como limiar para a segmentação. Esse modelo foi determinado com base em experimentos realizados previamente e que serviram para determinar a melhor abordagem para a presente tese.

Figura 17 – Sequência de atividades da metodologia da proposta.



Elaborada pela autora.

No total, foram 100 imagens de peixes da espécie Pacu fornecidas pelo LaGeAC para etapa de predição do modelo. Todas as imagens foram obtidas no modo retrato e no mesmo ambiente para que não houvesse alteração nos resultados do modelo, de acordo com a Figura 18. E apesar das imagens possuírem diferentes dimensões (pois não foram obtidas do mesmo dispositivo), isso não atrapalhou o desempenho da metodologia utilizada, já que cada método implementado seguiu a dimensão da imagem a ser analisada naquele momento. Com isso, para a etapa de treinamento, cada imagem foi segmentada manualmente usando uma ferramenta específica de segmentação chamada Labelbox [1]. Assim, o modelo foi treinado com o objetivo de destacar as partes do peixe e classificá-las corretamente conforme a segmentação realizada.

Para essa tese, o modelo de IA foi submetido ao processo de inferência para realizar as predições com base no que foi aprendido durante a etapa de treinamento. Além disso, a segmentação das imagens foi realizada com o intuito de delimitar o peixe da área de *background*, e não mais segmentar o animal em partes, conforme realizado anteriormente.

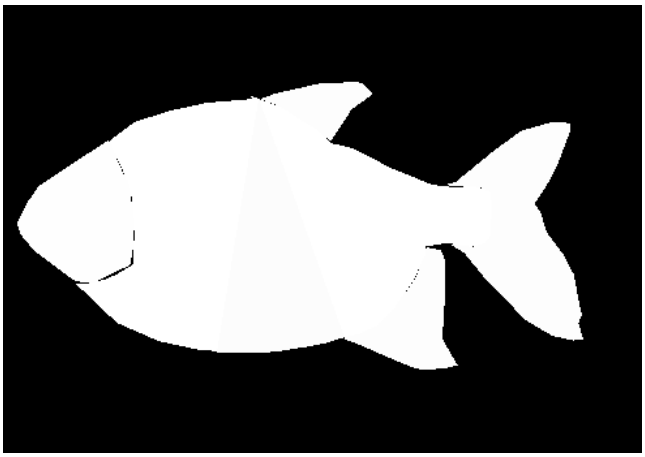
[1] Labelbox: https://labelbox.com/

Figura 18 – Exemplo de imagens de diferentes peixes Pacu fornecida pelo LaGeAC.

Elaborada pela autora.

Essa abordagem foi escolhida para uma primeira etapa, visando a segmentação em regiões como parte dos experimentos futuros. Assim, foi gerada uma máscara em preto e branco com base na segmentação manual, resultante da combinação das segmentações das regiões de interesse do peixe. Conforme a Figura 19, as áreas na cor branca representam a área do Pacu, enquanto que as áreas na cor preta representam o *background* e demais objetos da imagem.

Figura 19 – Exemplo de máscara original gerada a partir da segmentação manual realizada por meio da ferramenta Labelbox.



Elaborada pela autora.

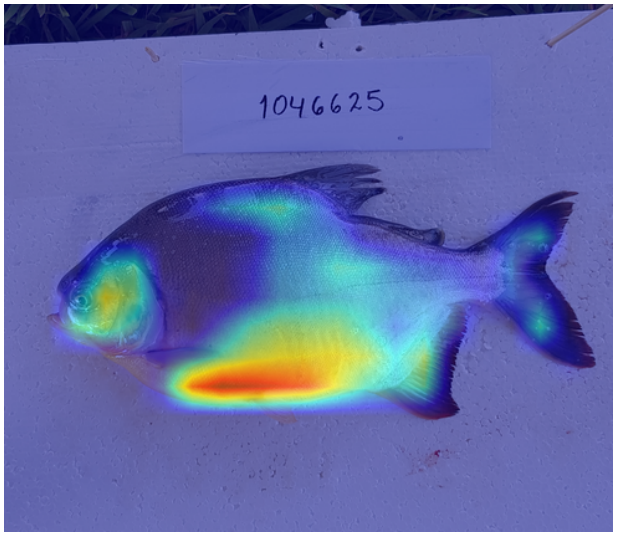
Cada uma das 100 imagens fornecidas tem sua respectiva máscara, possibilitando a comparação com as máscaras geradas pelo modelo durante os experimentos. Nesse contexto, as imagens foram submetidas às técnicas de perturbação de pixels e as máscaras (original e gerada durante a inferência) foram comparadas para analisar a diferença entre elas. Com isso, foi possível observar a influência das perturbações de pixels nos resultados dos métodos explicáveis. Essas análises são melhores discutidas ao longo desse trabalho.

4.2 Métodos de XAI

Os métodos de XAI foram aplicados com o objetivo de explicar as regiões mais importantes para o modelo de predição da Mask R-CNN. A escolha desses métodos foi realizada com base na compatibilidade com o modelo de IA utilizado nesta tese. Além disso, a maioria dos métodos explicáveis trazem resultados mediante ao processo de classificação. Todavia, para os experimentos dessa tese, o intuito foi obter resultados segundo o processo de segmentação, cujo objetivo foi separar o peixe do restante da imagem. Portanto, dentre os métodos estudados previamente, foram implementados quatro métodos explicáveis de visualização: Grad-CAM, Mapa de Saliência, CNN Filters e Layer Grad-CAM. Ambos destacam as regiões da imagem de maior relevância para o modelo identificar as regiões do peixe. Assim, é possível verificar se o resultado apresentado pela CNN condiz com as regiões importantes destacadas pelos métodos explicáveis.

Com base na fundamentação teórica apresentada, o método Grad-CAM é baseado em gradientes e utiliza cores para destacar as regiões mais relevantes da imagem durante o processo de segmentação. Dessa forma, as regiões apresentadas nas cores quentes (e.g., cor vermelha) são as mais importantes para o modelo durante o processo de inferência. Essa relevância diminui conforme as cores passam a ficar mais frias. Assim, as regiões na cor azul são as que apresentam a menor relevância na hora de separar o peixe das demais partes da imagem. A partir da Figura 20 é possível observar que a região do corpo e da cabeça do peixe Pacu são consideradas as mais importantes para o modelo, de acordo com o método Grad-CAM. Já as regiões que representam as nadadeiras superiores e traseiras não apresentam tanta importância para o processo de inferência.

Figura 20 – Exemplo de imagem de um peixe Pacu submetida ao método Grad-CAM.

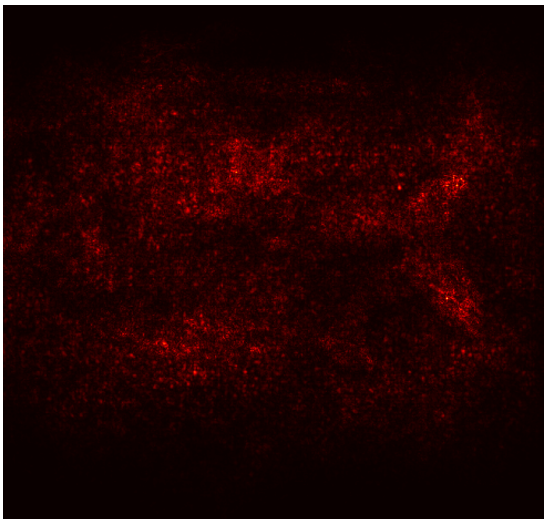


Elaborada pela autora.

O segundo método de XAI utilizado foi o Mapa de Saliência, cujo objetivo é destacar regiões relevantes da imagem a partir do brilho de cada pixel. Dessa forma, para

os experimentos dessa tese, a imagem de saída apresenta pixels nas cores preta e vermelha. Com isso, as regiões na cor preta representam o *background* da imagem, enquanto que as regiões na cor vermelha representam as áreas do Pacu, sendo que os pixels mais brilhantes são aqueles que apresentam maior importância para o processo de inferência. Conforme ilustrado na Figura 21, é possível observar que as regiões mais brilhantes estão localizadas nas nadadeiras do peixe. Alguns pixels de maior brilho também podem ser encontrados espelhados pelo corpo do animal.

Figura 21 – Exemplo de imagem de um peixe Pacu submetida ao método Mapa de Saliência.

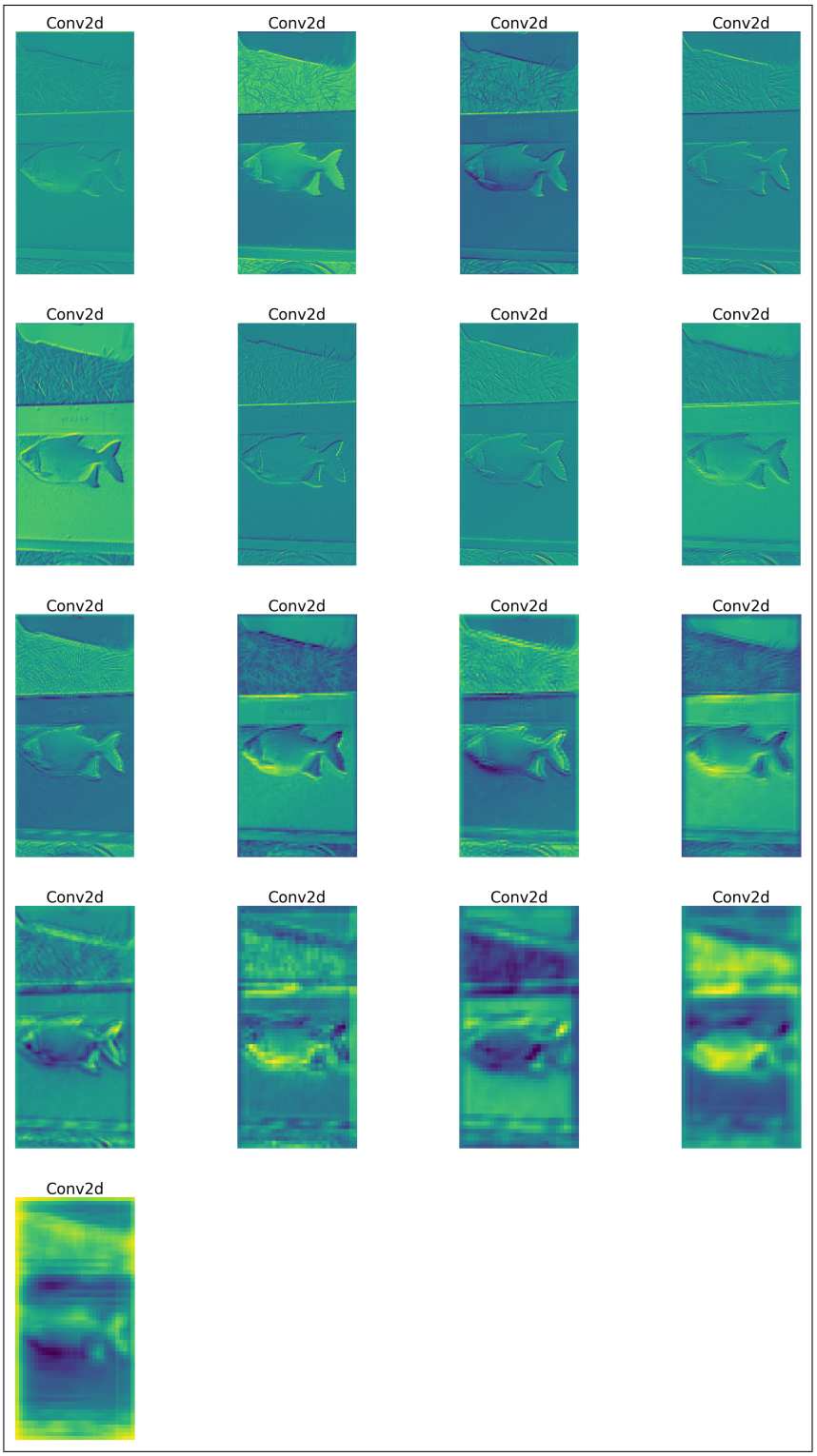


Elaborada pela autora.

O método CNN Filters reaproveita os filtros aplicados pela CNN para explicar o processo de inferência do próprio modelo, tornando-o também um método de XAI. Dessa forma, a quantidade de imagens geradas depende da quantidade de camadas de convolução utilizadas pela rede, o que para essa tese, foram 17 camadas no total (Figura 22). Cada imagem apresenta as regiões de maior relevância para a camada correspondente, fazendo com que os resultados sejam diferentes entre si, possibilitando diferentes explicações. As primeiras imagens são mais nítidas do que as últimas imagens, pois para cada camada, a imagem de entrada é a imagem de saída da camada anterior, sendo que em cada uma delas, alguns filtros são aplicados. O objetivo desses filtros é extrair informações relevantes, como por exemplo, regiões do peixe. Para esse tese, apenas o resultado da primeira camada da CNN foi utilizada para a explicabilidade do modelo, já que seria inviável usar todas as saídas para comparar com os demais métodos implementados.

O último método aplicado foi o Layer Grad-CAM, que baseia-se no cálculo do gradiente utilizado no método Grad-CAM, implementado na última camada de convolução da CNN. Com isso, para essa tese, a saída do método é uma imagem nas cores verde e vermelho conforme as classes consideradas para a segmentação durante o processo de inferência do modelo. Assim, de acordo com a Figura 23, a cor vermelha foi usada para

Figura 22 – 17 imagens de um peixe Pacu geradas pelo método CNN Filters.



Elaborada pela autora.

representar o *background* da imagem, enquanto a cor verde foi utilizada para representar a cabeça, o corpo e as nadadeiras do peixe, destacando assim o animal do restante da imagem. Dessa forma, é possível observar em verde quais pixels são consideradas relevantes para determinar as regiões do Pacu. Além disso, é importante ressaltar que a imagem de

saída gerada pelo método sofre uma diminuição em sua dimensão, já que é utilizada a saída da última camada da CNN como entrada para o método explicável, justificando a perda de qualidade da imagem.

Figura 23 – Exemplo de imagem de um peixe Pacu submetida ao método Layer Grad-CAM.

Elaborada pela autora.

Vale destacar que a visualização das explicações pode ser feita a partir de grupos de pixels ou pixel a pixel da imagem de entrada. Dessa forma, os métodos Grad-CAM e Layer Grad-CAM utilizam grupos de pixels para explicar o processo de inferência, enquanto que os métodos Mapa de Saliência e CNN Filters utilizam cada ponto da imagem para gerar a explicação. Com isso, o uso desses métodos mostrou-se importante devido a sua variabilidade e distinção. Ou seja, apesar da semelhança entre eles em relação ao tipo de visualização, e apesar de serem apenas quatro, os métodos de XAI implementados apresentam resultados explicáveis totalmente diferentes, possibilitando uma maior abrangência nos experimentos realizados, conforme apresentado nos capítulos seguintes.

4.3 Perturbação de Pixels

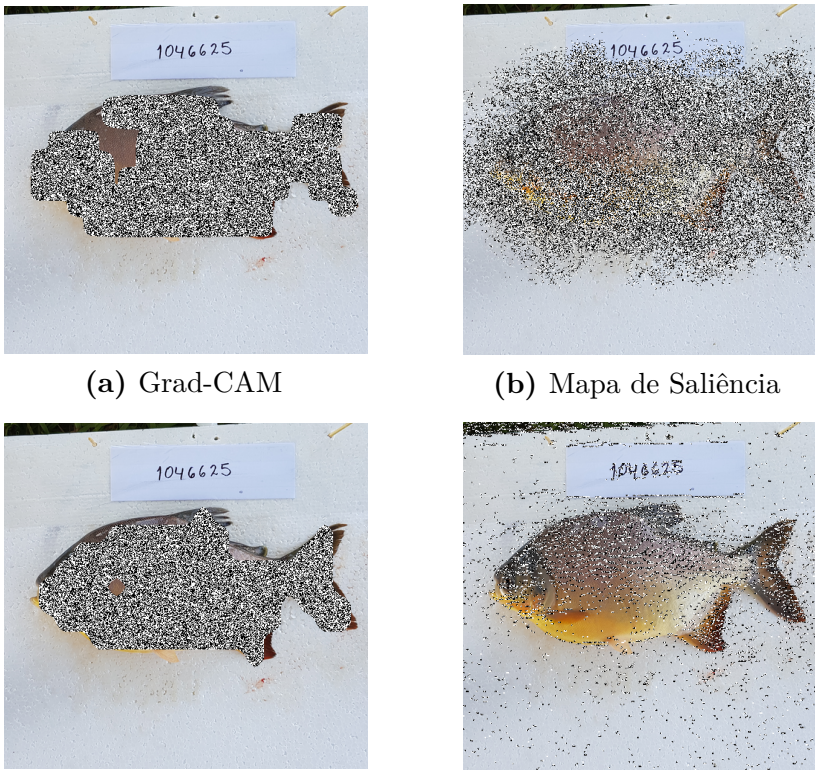
O desempenho dos modelos de IA pode ser analisado a partir de alterações nas imagens de entrada. Inclusive essa é a premissa utilizada em métodos explicáveis baseados em perturbação de pixels, cujo objetivo é explicar quais regiões ou pontos da imagem são considerados relevantes para o modelo. Dessa forma, para a presente tese, foram implementadas técnicas de perturbação para analisar a influência dessas alterações no desempenho dos modelos de IA, e como isso afeta a explicabilidade dos métodos de XAI aplicados. A escolha desse método de validação foi realizada com base em experimentos que ajudaram a determinar a melhor técnica para avaliação da explicabilidade.

As alterações promovidas pela perturbação de pixels podem acontecer de formas distintas, resultando em diferentes técnicas. Assim, três modelos de perturbação foram

considerados para essa tese. São eles: ruído branco, preto e aleatório. A implementação dessas técnicas foi realizada a partir das regiões de pixels destacadas pelos métodos explicáveis. Ou seja, considerando uma única imagem de entrada, para cada um dos métodos de XAI implementados, utilizou-se três técnicas de perturbação de pixels distintas, gerando três imagens de saída diferentes.

Para a técnica de ruído branco, as alterações são caracterizadas por pixels aleatoriamente pintados nas cores preta e branca. Dependendo do método explicável no qual essa perturbação foi implementada, o ruído branco pode aparecer em regiões mais densas, ou apenas em pontos da imagem. Isso ocorre também nas demais técnicas de perturbação citadas. A Figura 24 ilustra uma mesma imagem submetida ao ruído branco, de acordo com os diferentes métodos de XAI aplicados. Apesar de ser o mesmo tipo de perturbação, é possível observar que nos métodos Grad-CAM (Figura 24a) e Layer Grad-CAM (Figura 24c) as alterações ocorrem em regiões, conforme a característica do método. O mesmo ocorre com os métodos Mapa de Saliência (Figura 24b) e CNN Filters (Figura 24d), cujas alterações foram feitas pixel a pixel, de acordo também com a característica dos métodos explicáveis. Além disso, a perturbação de pixels só foi aplicada nas regiões ou pixels que os métodos explicáveis caracterizaram como relevantes para o modelo de IA.

Figura 24 – Exemplo de perturbação de pixels do tipo ruído branco para os quatro métodos de XAI.



(a) Grad-CAM **(b)** Mapa de Saliência

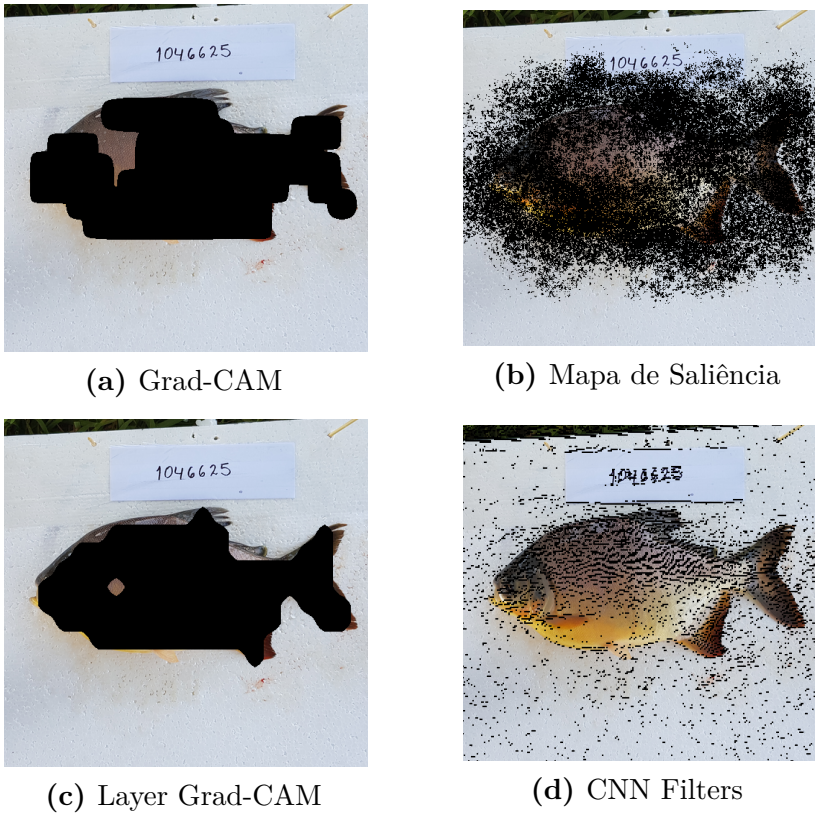
(c) Layer Grad-CAM **(d)** CNN Filters

Elaborada pela autora.

A segunda técnica de perturbação implementada é responsável por alterar os

pixels da imagem de entrada para a coloração preta. Com isso, com base em cada um dos métodos explicáveis aplicados, a cor preta é utilizada tanto em regiões como em pontos específicos da imagem, conforme sua relevância para o processo de inferência do modelo de IA. Assim, na Figura 25 é possível observar uma mesma imagem sendo submetida à perturbação de pixel na cor preta, conforme o método Grad-CAM (Figura 25a), Mapa de Saliência (Figura 25b), Layer Grad-CAM (Figura 25c) e CNN Filters (Figura 25d).

Figura 25 – Exemplo de perturbação de pixels na coloração preta para os quatro métodos de XAI.



(a) Grad-CAM **(b)** Mapa de Saliência

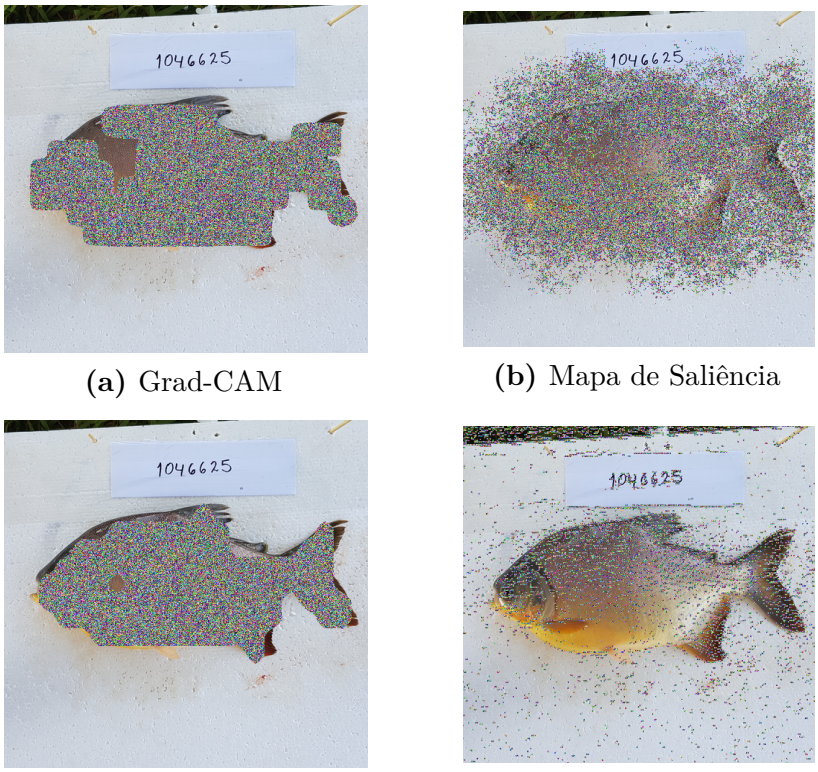
(c) Layer Grad-CAM **(d)** CNN Filters

Elaborada pela autora.

A última técnica de perturbação aplicada a essa tese altera os pixels de forma aleatória, com cores RGB entre 0 e 255. Assim, as regiões ou pontos mais relevantes para o processo de inferência são coloridos de acordo com o método XAI aplicado. Com isso, diferentemente das técnicas anteriores, mesmo para métodos explicáveis com visualização mais densa, cuja relevância é dada por regiões, ainda é possível observar as explicações pixel a pixel devido a grande variedade de cores. Conforme apresentado na Figura 26, cuja mesma imagem é submetida à perturbação de pixel aleatória, é possível verificar o comportamento dessa técnica nos quatro métodos explicáveis implementados (Figuras 26a, 26b, 26c e 26d).

A partir da implementação dessas técnicas de perturbação de pixels nas 100 diferentes imagens de peixes da espécie Pacu, foi possível analisar a relevância dos resultados dos métodos explicáveis adotados. Isto é, essas técnicas foram utilizadas como meio de

Figura 26 – Exemplo de perturbação de pixels do tipo aleatória para os quatro métodos de XAI.



(a) Grad-CAM **(b)** Mapa de Saliência

(c) Layer Grad-CAM **(d)** CNN Filters

Elaborada pela autora.

avaliação da explicabilidade dos métodos de XAI, possibilitando inclusive a escolha do melhor método explicável.

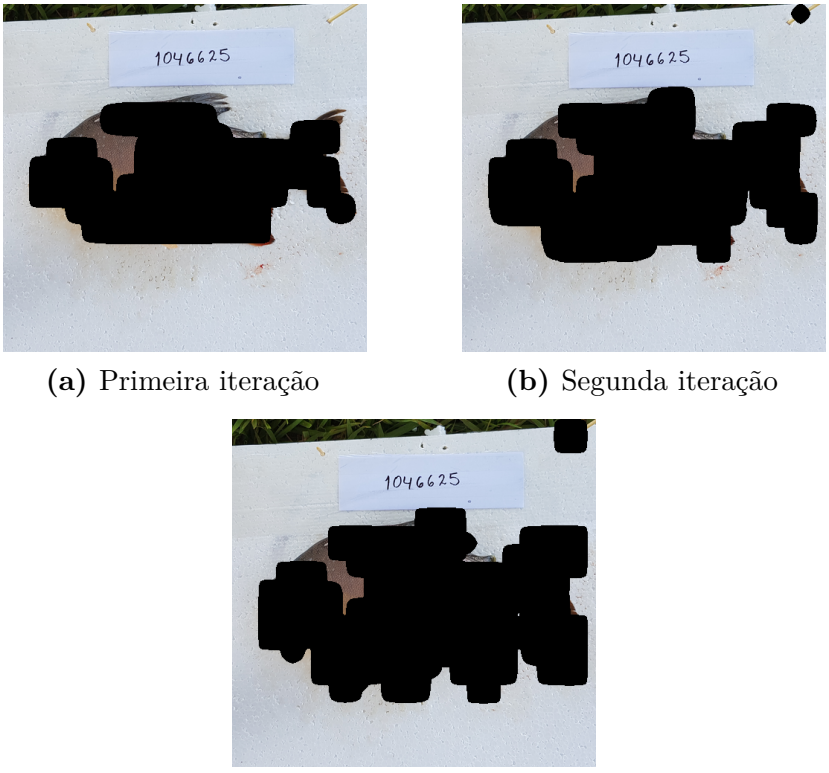
4.4 Melhor método de XAI

A busca pelo melhor método explicável é baseada nas explicações geradas em relação ao processo de inferência do modelo de IA implementado. Assim, antes mesmo de entender como foi feita a análise dos quatro métodos aplicados a essa tese, é necessário compreender o passo a passo que cada imagem foi submetida. Assim, para cada um dos métodos de XAI, as seguintes etapas foram executadas:

1. Imagem de entrada submetida ao modelo de IA;
2. Processo de inferência do modelo foi submetido ao método de XAI;
3. Pixels são perturbados conforme método explicável, gerando uma imagem de saída;
4. Imagem de saída é utilizada como imagem de entrada; e
5. Repetição das etapas até que ocorra um erro no processo de inferência.

A execução das etapas de 1 a 4 caracteriza o que chamamos de iteração. O número de iterações pode variar, dependendo da combinação entre método explicável e tipo de perturbação de pixels implementados. A cada iteração executada, mais perturbados são os pixels de maior relevância para o modelo, que são determinados de acordo com o método de XAI (Figura 27). Assim, a Figura 27a apresenta a primeira iteração para o método Grad-CAM sobre a influência da perturbação de pixels na cor preta. A Figura 27b, em relação a imagem anterior, apresenta mais regiões coloridas de preto, representando a segunda iteração. Já a última imagem (Figura 27c), que representa a terceira iteração, é a que mais possui regiões perturbadas, inclusive regiões que não fazem parte do corpo, cabeça ou nadadeiras do peixe. Dessa forma, é possível observar que a perturbação ocorre em cada iteração, para a mesma imagem, de forma cumulativa.

Figura 27 – Exemplo de iterações para o método Grad-CAM sobre a influência da perturbação de pixel na coloração preta.



(a) Primeira iteração **(b)** Segunda iteração

(c) Terceira iteração

Elaborada pela autora.

Quando se fala de um bom método explicável, isso dificulta o processo de inferência, já que as regiões relevantes são "cobertas"e não podem ser identificadas pelo modelo de IA. Dessa forma, quando o modelo indica erro durante a inferência, significa dizer que o método XAI utilizado é bom, pois não há mais nenhuma região do peixe que possa ser segmentada. Caso o modelo continue com suas iterações, é correto afirmar que essas regiões que foram perturbadas não são importantes para o modelo, indicando que o método XAI utilizado não é bom, pois destacou regiões que não são de fato relevantes.

A influência da perturbação de pixels aplicada ao modelo também determina o desempenho dos métodos de XAI. Ou seja, para cada método, é necessário analisar como o modelo de IA comporta-se diante de cada técnica de perturbação implementada. Dessa forma, o primeiro experimento realizado para esta tese, baseia-se na busca pelo melhor método, cujas as técnicas de perturbação foram testadas por meio das combinações entre eles. Maiores detalhes desse experimento são apresentados no capítulo seguinte.

4.5 Combinação de métodos de XAI

A partir da escolha do melhor método explicável, foi sugerida a combinação desse método com os demais métodos implementados. O objetivo desse processo foi identificar se a junção do melhor com o pior, pode melhorar ou não os resultados apresentados. Em caso de melhorias, isso responde uma das questões de pesquisa apresentada no início dessa tese: "Como melhorar métodos de XAI para apresentarem explicações condizentes com a predição do modelo de IA?".

Para analisar se a combinação foi eficiente de fato, foi necessário realizar as mesmas etapas apresentadas na busca pelo melhor método. Entretanto, a diferença está na aplicação do método explicável, que ao invés de ser utilizado apenas um método, foi realizada a combinação do melhor com cada um dos outros três métodos restantes. Dessa forma, as etapas são descritas da seguinte forma:

1. Imagem de entrada submetida ao modelo de IA;
2. Processo de inferência do modelo foi submetido ao melhor método de XAI combinado com outro método de XAI;
3. Pixels são perturbados conforme a combinação dos métodos explicáveis, gerando uma imagem de saída;
4. Imagem de saída é utilizada como imagem de entrada; e
5. Repetição das etapas até que ocorra um erro no processo de inferência.

A análise dos resultados foi realizada de forma semelhante ao processo de busca pelo melhor método de XAI. Ou seja, técnicas de perturbação de pixels foram implementadas para que fosse verificada a influência das mesmas sobre o processo de inferência e a explicabilidade da combinação dos métodos. Dessa forma, com base na quantidade de iterações obtidas para cada imagem, foi possível observar se ocorreu ou não a melhora na explicabilidade. Além disso, a combinação foi feita a partir das máscaras geradas por cada um dos método durante o processo de inferência. Com isso, a imagem de saída do modelo continha os pixels e as regiões mais importantes destacados conforme

a técnica de perturbação aplicada. A Figura 28 ilustra um exemplo de imagem de peixe que foi submetida a combinação dos métodos Grad-CAM e Mapa de Saliência, a partir da implementação da técnica de perturbação de coloração preta.

Figura 28 – Exemplo de imagem de um peixe Pacu submetida a combinação dos métodos Grad-CAM e Mapa de Saliência sobre efeito da perturbação de pixels de coloração preta.

Elaborada pela autora.

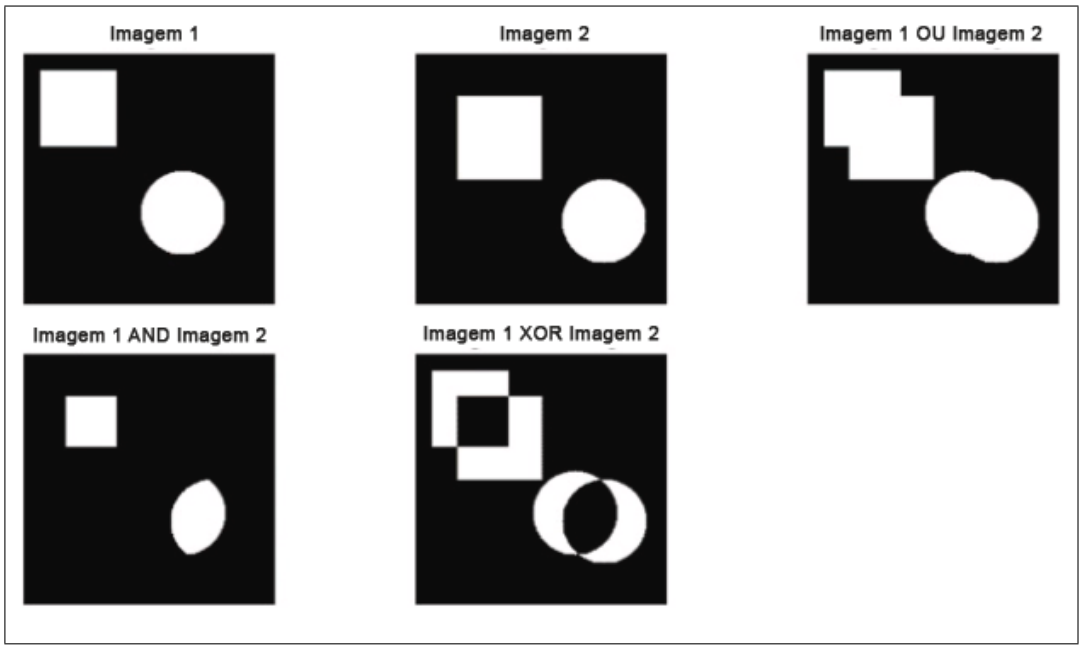
O processo de combinação entre os métodos explicáveis foi realizado com base nos operadores lógicos OR, AND E XOR (COULIBALY et al., 2022). A Figura 29 ilustra o resultado dessas combinações. Assim, é possível observar que o operador AND corresponde a intersecção entre as imagens resultantes dos métodos, enquanto que o operador OR corresponde a união das imagens e o operador XOR corresponde a disjunção exclusiva. Dessa forma, para a presente tese, o operador escolhido foi o AND devido a sua capacidade em determinar as áreas comuns entre as imagens.

O segundo experimento realizado foi baseado nessa etapa da metodologia, cujo objetivo é a combinação entre o melhor método de XAI (encontrado no primeiro experimento) com os demais métodos explicáveis implementados. Assim, foi possível analisar a melhora, ou piora, dos métodos combinados quando submetidos a influência das técnicas de perturbação de pixels. Mais informações sobre esses experimentos são apresentadas no capítulo seguinte.

4.6 Análise dos resultados

A última etapa da metodologia aplicada a essa tese foi destinada a análise dos resultados obtidos nos experimentos executados. Assim, métricas foram necessárias para melhor compreensão dos resultados obtidos, somada a criação de gráficos baseados nas

Figura 29 – Operadores lógicos para combinação de imagens resultantes de métodos de XAI.



Fonte - (COULIBALY et al., 2022)

iterações realizadas pelo modelo de IA. Essa abordagem foi necessária para compreender o desempenho dos métodos explicáveis, garantindo a escolha correta do melhor método dentre os quatro implementados. A análise da combinação entre os métodos também foi realizada de forma análoga, possibilitando compreender o comportamento dos mesmos a partir desse ajuste. Com isso, diante da influência de cada uma das técnicas de perturbação, os resultados foram separados da seguinte forma:

- Média da quantidade de iterações por método/combinação para cada técnica de perturbação aplicada;
- Média da quantidade de imagens de saída obtidas em cada método/combinação para cada técnica de perturbação aplicada;
- Média da quantidade de imagens de saída por técnica de perturbação de pixels aplicada;
- Média da quantidade de imagens de saída por método/combinação;
- Média da quantidade de iterações por método/combinação;
- Média dos índices Sorensen Dice e Intersecção Sobre União por método/combinação, para cada técnica de perturbação aplicada; e
- Média dos índices Sorensen Dice e Intersecção Sobre União por método/combinação.

O coeficiente Sorensen Dice é um método estatístico que visa analisar a similaridade entre duas amostras (SORENSEN, 1948; DICE, 1945). A fórmula original utilizada para esse método é apresentada na Equação 4.1, cujo resultado é apresentado no intervalo de [0,1]. Dessa forma, X e Y representam a quantidade de elementos nas duas amostras a serem comparadas, como por exemplo, imagens. Para essa tese, compara-se a similaridade entre a máscara original da segmentação do peixe, e a máscara gerada pelo modelo de IA a partir do método de XAI e da técnica de perturbação de pixel aplicados. Esse cálculo foi realizado a cada iteração, para cada um dos métodos explicáveis, sobre a influência de cada técnica de perturbação. Além disso, a intersecção entre as amostras indica quais são os pixels compartilhados pelas duas. Quanto mais próximo o resultado da equação é do valor 1, mais similar são as máscaras comparadas.

$$\frac{2|X \cap Y|}{|X| + |Y|}. \tag{4.1}$$

A Intersecção Sobre a União, também conhecida como índice de Jaccard (JACCARD, 1912), é uma métrica utilizada em algoritmos de DL para verificar a semelhança entre dois conjuntos. Neste caso, foi analisada a semelhança entre o conjunto de pixels da máscara original, em relação ao conjunto de pixels da máscara gerada pelo modelo. O cálculo dessa métrica é realizado a partir da divisão entre a sobreposição dos conjuntos em relação a união dos mesmos, conforme apresentado na Equação 4.2. Os conjuntos são representados por A e B, e o resultado obtido pela equação está no intervalo de [0,1].

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{4.2}$$

A partir desses valores, gráficos foram criados para melhor visualização dos resultados. Com isso, foi possível identificar o melhor método explicável e também a melhor combinação aplicada. Além disso, também foi possível observar qual técnica de perturbação de pixel foi mais influente em relação ao processo de inferência do modelo de IA e a explicabilidade do método de XAI. Por fim, também foi possível observar que em um mesmo método, as técnicas de perturbação podem influenciar de formas diferentes.

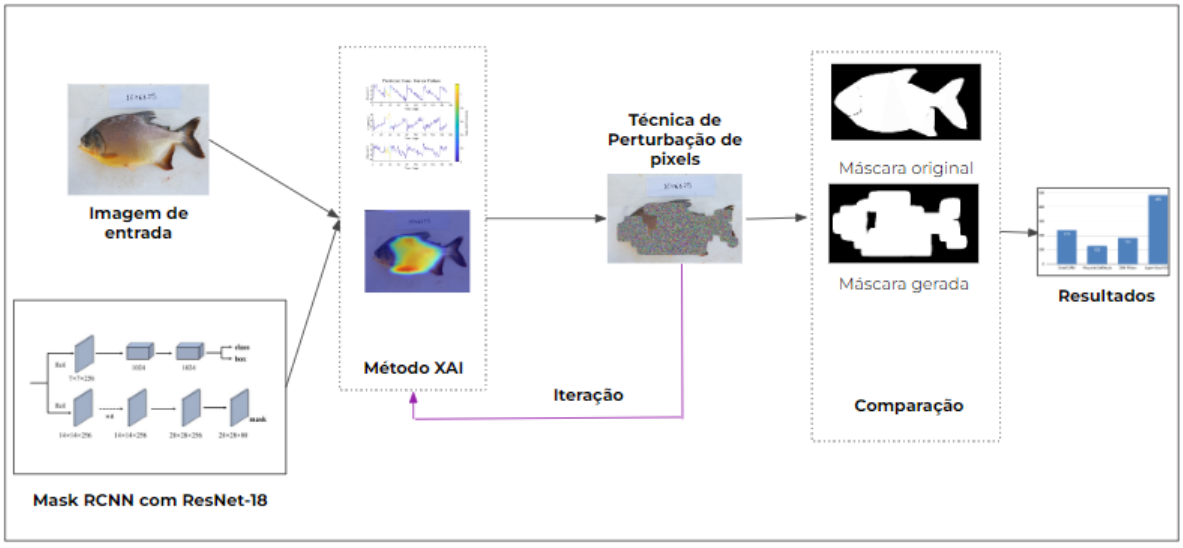
5 Experimentos e Resultados

O presente capítulo descreve os experimentos realizados e os resultados obtidos com base na metodologia apresentada anteriormente. Dessa forma, foram dois experimentos, sendo o primeiro relacionado a busca pelo melhor método explicável, e o segundo relacionado à melhora dos métodos de XAI por meio da combinação do melhor com os demais. Ambos foram implementados a partir de códigos criados com a biblioteca Pytorch da linguagem Python, e executados no Google Colaboratory. Esses experimentos são descritos em detalhes nas próximas seções.

5.1 Experimento I

A execução do primeiro experimento contou com um modelo já treinado para segmentação de peixes da espécie Pacu. O modelo foi criado a partir de uma Mask R-CNN utilizando como extrator de características uma variante da ResNet-18. Além disso, os métodos de XAI implementados (Grad-CAM, Mapa de Saliência, CNN Filters e Layer Grad-CAM) foram considerados durante o processo de predição para analisar quais pixels eram mais importantes. Por fim, as técnicas de perturbação implementadas (ruído branco, preto e aleatório) serviram como método de avaliação dos métodos explicáveis. As etapas desse primeiro experimento são apresentadas na Figura 30.

Figura 30 – Diagrama da metodologia utilizada para o primeiro experimento.



Elaborada pela autora.

O processo de perturbação das imagens de entrada ocorreu de acordo com as pontuações obtidas para cada pixel nos métodos de explicabilidade. Cada método considera

uma escala diferente para destacar as regiões mais relevantes, sendo que para cada um desses, foram considerados os pixels com os valores apresentados na Tabela 2. Para identificar valores limítrofes de cada um dos métodos, foi utilizada a inspeção visual em algumas das imagens presentes no *dataset*.

Tabela 2 – Valores dos pixels mais importantes para cada método de XAI.

Método XAI	**Valores**
Grad-CAM	Acima de 0,01
Mapa de Saliência	Acima de 0,045
Layer Grad-CAM	Acima de -0,5
CNN Filters	Abaixo de 0

Fonte – Elaborada pela autora.

Para determinar qual foi o melhor método explicável, foi necessário realizar a comparação da máscara original, gerada durante a segmentação manual, com a máscara gerada durante a predição do modelo após sucessivas perturbações nos pixels. Com isso, a comparação foi realizada com base nos índices IoU e SD explicados no capítulo anterior. Os resultados obtidos foram tabelados de acordo com a técnica de perturbação e o método de explicabilidade utilizado.

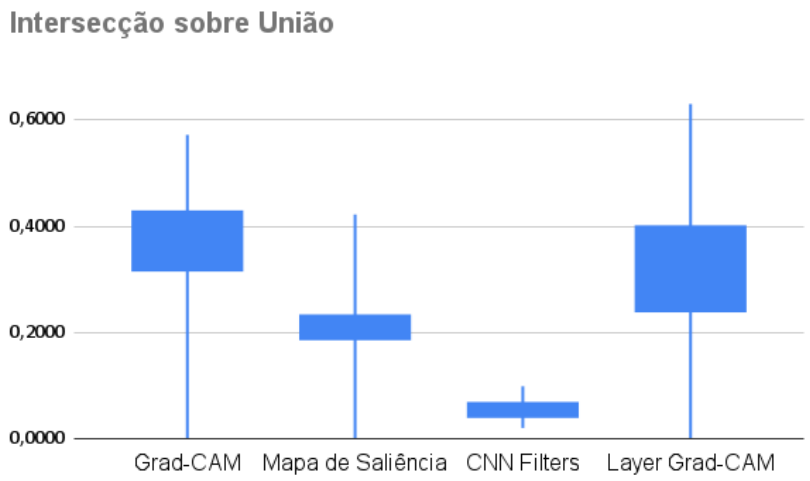
Todas as etapas foram feitas em no máximo cinco iterações por imagem. Em alguns casos, o próprio modelo interrompeu o processo antes mesmo de atingir o máximo. Nessas situações, o modelo não foi capaz de realizar a segmentação após a perturbação dos pixels, interrompendo assim o processo. Para cada iteração, a imagem utilizada como entrada era a imagem gerada como saída pela iteração anterior, exceto na primeira iteração cuja entrada era a imagem original do peixe. Dessa forma, a cada passagem, os pixels mais relevantes eram cada vez mais perturbados com o intuito de dificultar ainda mais o processo, e consequentemente, prejudicar a capacidade de predição do modelo. Com base nisso, os resultados obtidos desses experimentos foram computados e são apresentados a seguir.

5.1.1 Resultados

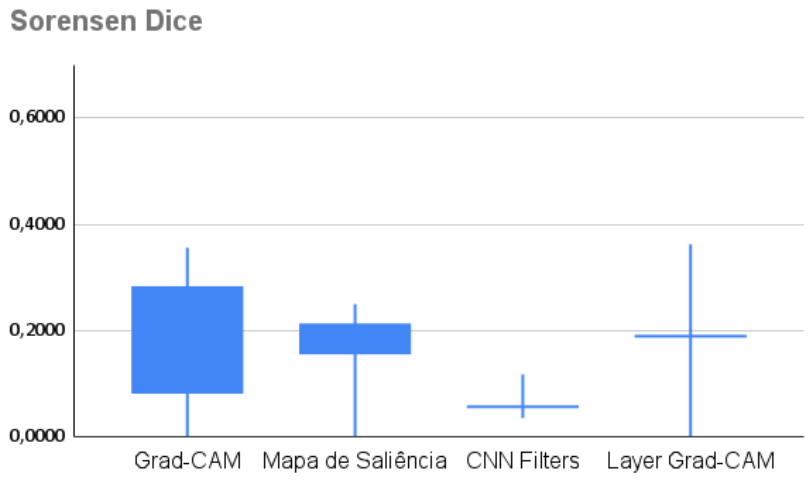
A partir do primeiro experimento, foi possível determinar a influência das técnicas de perturbação sobre os resultados apresentados pelos métodos de XAI, a fim de identificar o melhor método dentre os quatro apresentados. Portanto, ao observar os índices IoU e SD, foi possível verificar os diferentes comportamentos dos métodos durante o experimento. Os índices variam aproximadamente de 0 a 0,6, conforme os gráficos do tipo *boxplots* apresentados na Figura 31. Assim, é possível observar que o método Grad-CAM, em ambos os índices, teve uma variação considerável em relação aos outros métodos. Vale

ressaltar o comportamento do método Layer Grad-CAM que apresentou uma variação discreta no índice IoU (Figura 31a) e para o índice SD apresentou um comportamento contrário (Figura 31b), sendo o método com maior variação para esse índice. Essas variações mostram que a máscara da predição, devido a perturbação dos pixels da imagem, sofreu alterações significativas se comparada a máscara original. Esse processo aconteceu em uma única iteração ou ao longo das cinco iterações, a depender da disposição da técnica de perturbação com o método explicável utilizados.

Figura 31 – Gráficos *boxplots* do primeiro experimento referente aos índices IoU e SD dos métodos de XAI.



(a) IoU



(b) SD

Elaborada pela autora.

A partir da Tabela 3 é possível observar melhor os valores obtidos para os índices. Com esses resultados, conclui-se que as máscaras geradas para cada método de XAI destacam pixels ou regiões que vão além da cabeça, corpo e nadadeiras do peixe. O método CNN Filters, por exemplo, obteve os menores valores para ambos os índices, o

que indica que as regiões destacadas na cor branca, em sua maioria, não fazem parte do peixe Pacu. Todavia, para o método Layer Grad-CAM, é possível observar que os valores são os mais altos, o que indica que grande parte das áreas destacadas em branco pertencem ao peixe. Lembrando que a máscara gerada durante o processo de predição ignora características dos métodos, como cor ou brilho dos pixels. Essas características são de extrema importância para identificar as regiões mais relevantes da imagem e portanto, impedem que os métodos sejam avaliados corretamente apenas utilizando os índices IoU e SD. Por isso, para determinar os resultados, além desses valores, foram observadas a média da quantidade de iterações, e consequentemente, a média da quantidade de imagens geradas para cada um dos métodos. Esses valores foram considerados para determinar se o método foi ou não influenciado pelas técnicas de perturbação de pixels. Além disso, é preciso salientar que a quantidade de iterações determinou a quantidade de imagens geradas ao longo dos experimentos. Dessa forma, para cada imagem de entrada, gerou-se no máximo cinco imagens de saída, dependendo do método explicável e da técnica de perturbação aplicados em conjunto.

Tabela 3 – Valores obtidos para os índices IoU e SD em relação aos métodos de XAI

Método XAI	**Técnicas de Perturbação**	**IoU**	**SD**
Grad-CAM	Aleatório	0,2821	0,2102
	Preto	0,1654	0,1503
	Ruído branco	0,2864	0,2426
Mapa de Saliência	Aleatório	0,2000	0,1763
	Preto	0,1985	0,1705
	Ruído branco	0,2069	0,1816
Layer Grad-CAM	Aleatório	0,3688	0,2288
	Preto	0,2421	0,2040
	Ruído branco	0,3435	0,2468
CNN Filters	Aleatório	0,0544	0,0671
	Preto	0,0566	0,0721
	Ruído branco	0,0549	0,0685

Fonte – Elaborada pela autora.

A Tabela 4 apresenta a média da quantidade de iterações e a quantidade de imagens obtidas para cada um dos métodos de XAI e técnicas de perturbação implementados. Além disso, vale ressaltar que ambos os valores ajudam a determinar o desempenho de um método explicável diante das alterações dos pixels da imagem de entrada. Assim, os métodos que alcançaram o menor número de iterações e, consequentemente de imagens, foram os que apresentaram o melhor resultado. Dessa forma, observa-se que nem todos os métodos tiveram o mesmo comportamento, e que diferentemente do que foi observado com os valores dos índices IoU e SD, os métodos que apresentaram os melhores resultados, não necessariamente alcançaram esse mesmo feito diante da quantidade de iterações e imagens geradas.

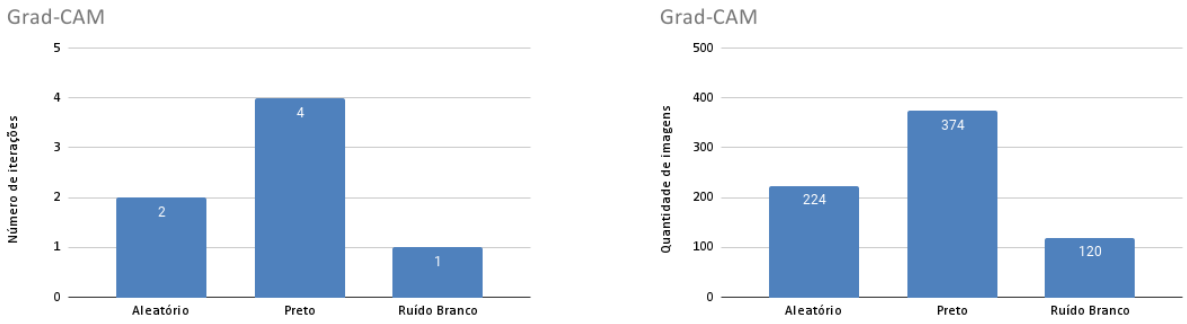
Tabela 4 – Média da quantidade de iterações e a quantidade de imagens por método de XAI em relação às técnicas de perturbação de pixels implementadas.

Método XAI	Técnicas de Perturbação	Iterações	Imagens
Grad-CAM	Aleatório	2	224
	Preto	4	374
	Ruído branco	1	120
Mapa de Saliência	Aleatório	1	123
	Preto	1	144
	Ruído branco	1	118
Layer Grad-CAM	Aleatório	5	493
	Preto	5	499
	Ruído branco	5	464
CNN Filters	Aleatório	2	188
	Preto	2	182
	Ruído branco	2	184

Fonte – Elaborada pela autora.

A Figura 32 apresenta a média da quantidade de iterações (32a) e a quantidade de imagens (32b) geradas para o método Grad-CAM. Diante disso, é possível observar que da mesma forma que o método apresentou variações nos valores dos índices IoU e SD em relação as diferentes técnicas de perturbação, o mesmo foi observado para a quantidade de imagens e iterações obtidas. Dessa forma, diante dos resultados apresentados, a técnica de cor preta foi a que menos afetou o método, pois as explicações apresentadas não destacaram os pixels de real importância para o modelo. Enquanto que a técnica de ruído branco foi a que mais influenciou o método Grad-CAM, resultando em uma quantidade menor de imagens, com média de uma única iteração por imagem de entrada.

Figura 32 – Média de iterações e quantidade de imagens obtidas por meio do método Grad-CAM.



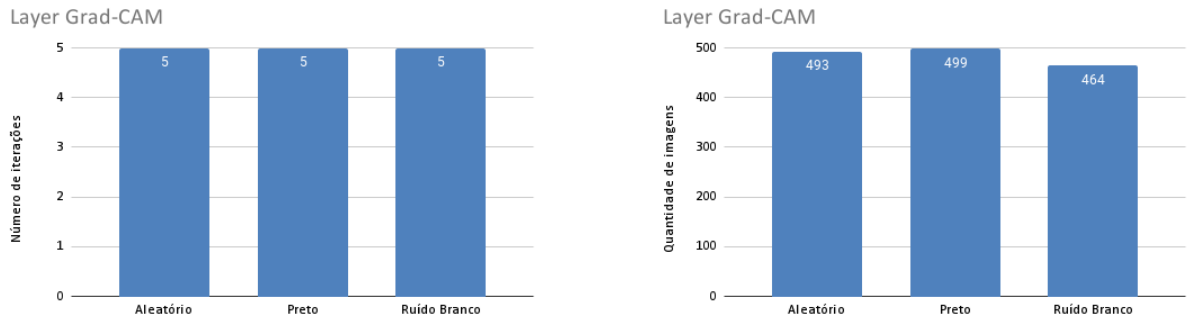
(a) Média da quantidade de iterações **(b)** Quantidade de imagens

Elaborada pela autora.

Sobre os métodos que atingiram o máximo do número de iterações (cinco), é possível afirmar que o modelo continuou sendo capaz de realizar a segmentação mesmo após a perturbação dos pixels mais relevantes para o método corrente. Isto é, a perturbação

não afetou o modelo e o processo de predição continuou normalmente. Com isso, é possível afirmar que o método explicável caracterizou como importantes pixels, ou regiões de pixels, de forma errônea. Assim, ao continuar a predição mesmo diante da perturbação, significa dizer que a alteração dos pixels não afetou significativamente o modelo de IA. A exemplo, tem-se o método Layer Grad-CAM que apresentou a média máxima de iterações, tornando-se o pior dentre os quatro métodos testados. Esses resultados podem ser observados por meio do gráfico da Figura 33 que mostra a média de iterações para cada tipo de perturbação (Figura 33a), juntamente com a quantidade de imagens resultantes (Figura 33b), cujo máximo possível seria 500 (máximo de cinco iterações para cada uma das 100 imagens de entrada).

Figura 33 – Média de iterações e quantidade de imagens obtidas por meio do método Layer Grad-CAM.



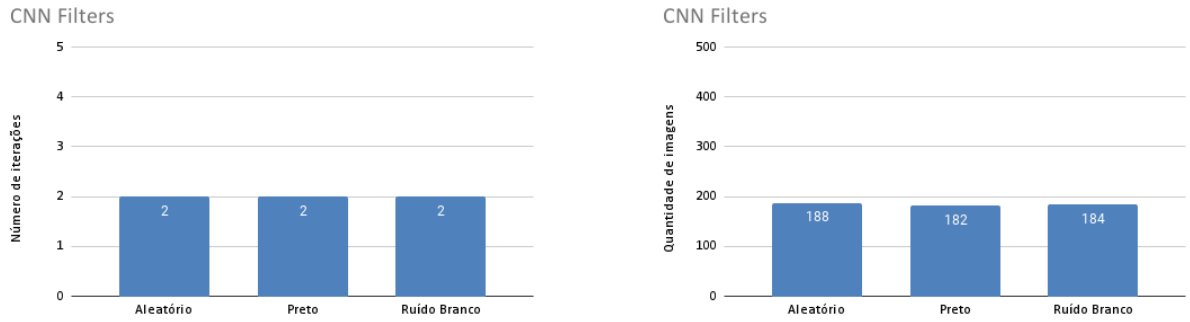
(a) Média da quantidade de iterações **(b)** Quantidade de imagens

Elaborada pela autora.

O método CNN Filters obteve resultados satisfatórios diante de todas as técnicas de perturbação de pixels aplicadas, se comparado ao método Layer Grad-CAM e Grad-CAM, conforme apresentado na Figura 34. Dessa forma, a média de iterações manteve-se estável diante das alterações realizadas nas imagens de entrada (Figura 34a), enquanto que a quantidade de imagens foi de duas por iteração (Figura 34b). Apesar disso, esses resultados demonstram que o modelo ainda sim realizou uma iteração a mais, indicando que os pixels destacados pelo método não eram todos relevantes para o processo de predição. Portanto, é possível observar que apenas após a segunda iteração é que o método conseguiu destacar todas as regiões que de fato eram significativas para o modelo.

Em relação aos métodos que interromperam o processo de predição antes mesmo de atingirem o máximo de cinco iterações, significa dizer que a implementação em conjunto do método XAI com a técnica de perturbação de pixels afetou o modelo, e portanto, impediu o processo de predição por completo. Dessa forma, é possível afirmar que o método explicável destacou os pixels verdadeiramente mais relevantes para o modelo de IA. O Mapa de Saliência alcançou esse resultado conforme apresentado na Figura 35, pois obteve a menor média de iterações (35a) ao mesmo tempo que demonstrou uma

Figura 34 – Média de iterações e quantidade de imagens obtidas por meio do método CNN Filters.

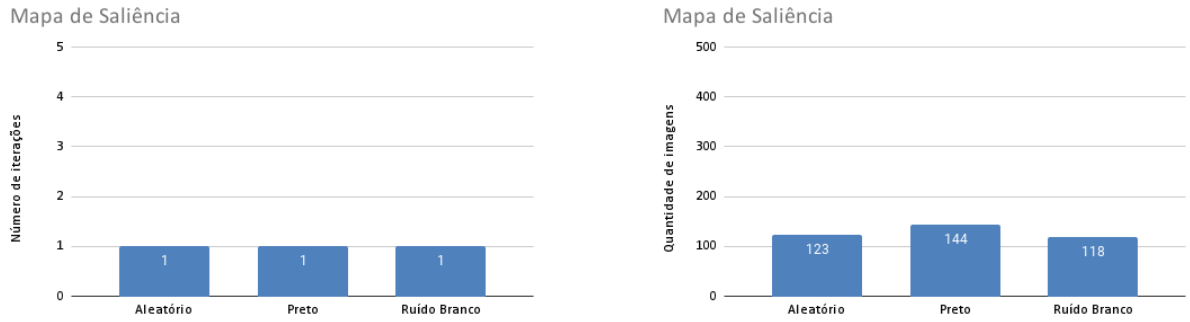


(a) Média da quantidade de iterações **(b)** Quantidade de imagens

Elaborada pela autora.

sensibilidade menor aos diferentes tipos de perturbação de pixels, tornando-se o melhor método entre os quatro. Isso inclusive refletiu na quantidade de imagens obtidas, conforme apresentado na Figura 35b.

Figura 35 – Média de iterações e quantidade de imagens obtidas por meio do método Mapa de Saliência.

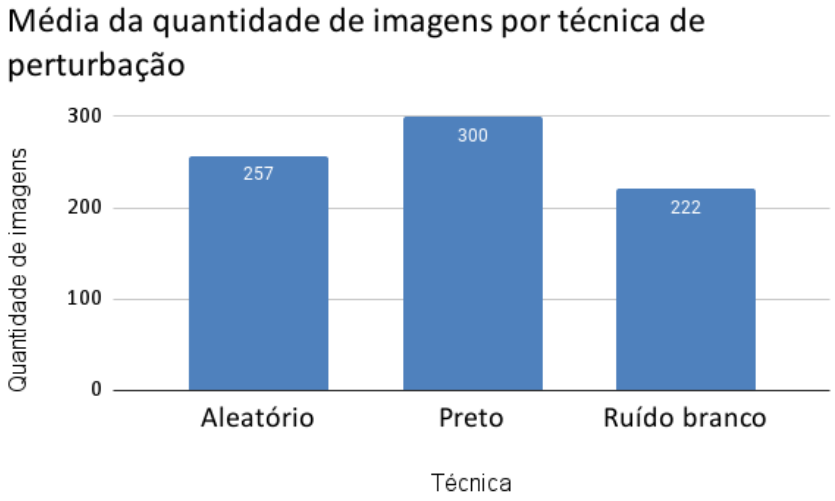


(a) Média da quantidade de iterações **(b)** Quantidade de imagens

Elaborada pela autora.

Das técnicas de perturbação de pixels implementadas, é possível observar que três dos quatro métodos XAI sofreram maior influência da técnica de coloração preta, se comparado as demais técnicas de rúido branco e aleatório, conforme ilustrado na Figura 36. É possível que esse fenômeno esteja relacionada ao uso de *padding* que geralmente são implementadas com pixels de valor zero, tornando assim o modelo mais resiliente a perturbações dessa natureza. Assim, para esses métodos, o número de iterações e imagens geradas foi maior para essa técnica de perturbação do que para as demais. Entretanto, o método CNN Filters foi o único a apresentar menor sensibilidade a essa técnica de perturbação, tornando-se o método com menor variação entre as técnicas, quando considerada a de quantidade de imagens resultantes.

Figura 36 – Média da quantidade de imagens por técnica de perturbação de pixels para o primeiro experimento.



Elaborada pela autora.

5.2 Experimento II

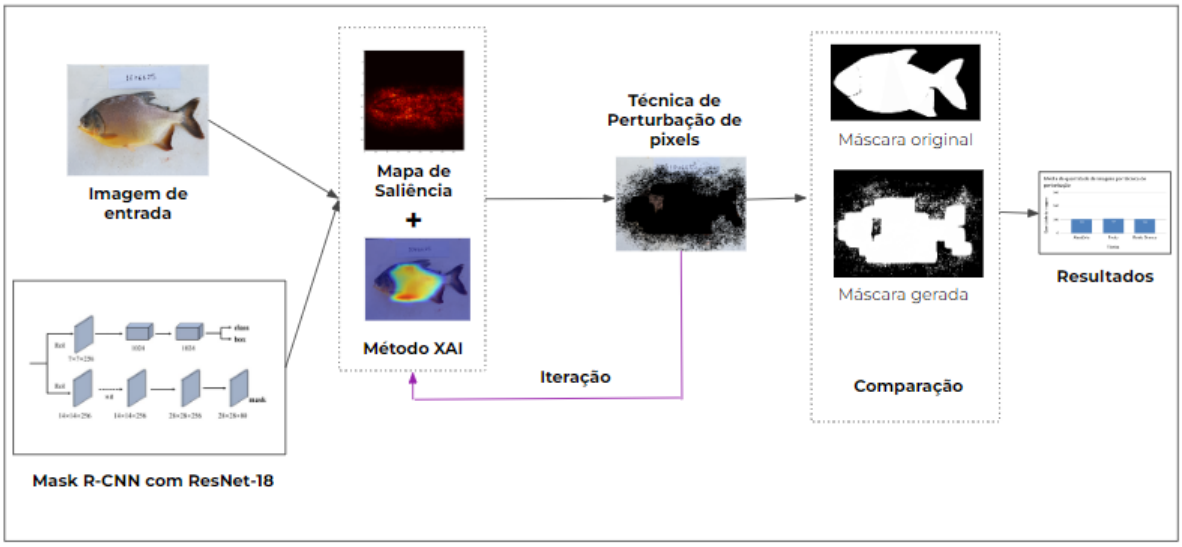
O segundo experimento foi realizado com base nos primeiros resultados obtidos nessa tese. Dessa forma, o objetivo foi melhorar os métodos de XAI a partir da combinação com o melhor método determinado anteriormente no primeiro experimento. Assim, com base nas três técnicas de perturbação de pixels implementadas, o melhor método explicável foi o Mapa de Saliência, que mostrou-se influenciável diante das modificações realizadas nas imagens do peixe Pacu. Ou seja, o método apresentou as regiões verdadeiramente relevantes para o processo de predição.

A combinação entre os métodos ocorreu conforme a metodologia apresentada, cuja operação lógica AND foi a escolhida para o experimento. Essa operação realiza a intersecção entre as imagens resultantes dos métodos de XAI, cujos pixels são destacados de acordo com sua relevância para ambos os métodos. Diante disso, as regiões mais importantes foram definidas com base na Tabela 2 apresentado no experimento I, que define os valores limítrofes utilizados em cada um dos métodos. Além disso, para avaliar os resultados das combinações, as mesmas técnicas de perturbação utilizadas anteriormente foram aplicadas.

O modelo de IA e o *dataset* utilizados foram os mesmos do primeiro experimento, cujo objetivo é realizar a segmentação das imagens de entrada com base no corpo, cabela e nadadeiras do peixe Pacu. Somado a isso, na Figura 37 é possível verificar as etapas metodológicas realizadas nessa implementação. Diferentemente do experimento anterior, as iterações são determinadas a partir da combinação dos métodos explicáveis com o Mapa de Saliência, e não mais da simples implementação de cada um deles. Com isso, os três métodos de perturbação de pixels foram aplicados, e a máscara resultante foi utilizada

para comparação, assim como realizado anteriormente.

Figura 37 – Diagrama da metodologia utilizada para o segundo experimento.



Elaborada pela autora.

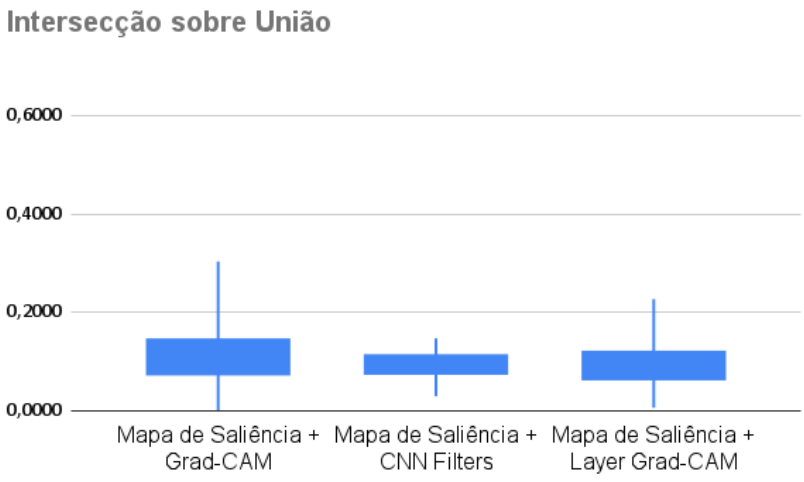
O máximo de cinco iterações por imagem também foi considerado para o presente experimento, a fim de avaliar se a combinação entre o método Mapa de Saliência e os demais métodos foi influenciada pela perturbação dos pixels classificados como significativos para o modelo. Além disso, os mesmos critérios foram considerados para análise dos resultados. Com isso, os valores obtidos para os índices IoU e SD foram observados, assim como a média da quantidade de iterações e a quantidade de imagens por método explicável. Por fim, a partir dos resultados foi possível observar o comportamento dos métodos diante da combinação, comparando inclusive com os resultados obtidos anteriormente, nos quais os métodos de XAI foram aplicados separadamente.

5.2.1 Resultados

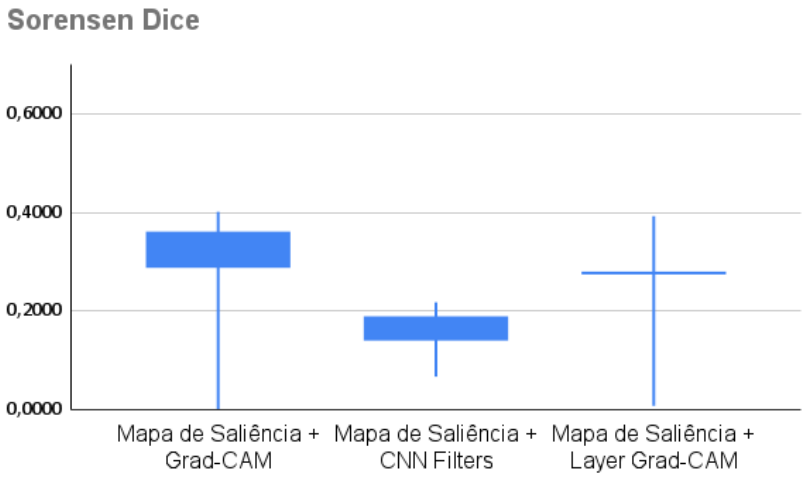
A avaliação dos métodos explicáveis combinados foi realizada de forma análoga ao experimento I. Com isso, a fim de identificar a melhor combinação dentre as três, os índices IoU e SD foram aplicados, e os resultados obtidos são apresentados nos gráficos do tipo *boxplot* da Figura 38. Assim, as variações foram mais sutis e o limite máximo foi de apenas 0,4. Esse comportamento pode ser justificado devido ao fato de que ao combinar os métodos Mapa de Saliência com os demais, mais pixels são destacados e mais diferenças são encontradas entre a máscara original e a gerada pelo modelo. Lembrando que nem sempre todos as regiões consideradas relevantes pelos métodos explicáveis pertencem as regiões do peixe. Com isso, a partir da Figura 38a, é possível observar que as três combinações alcançaram resultados semelhantes para o índice IoU, se observado o retângulo que indica a variação entre os valores. Entretanto, o valor máximo alcançado para o índice foi menor para a combinação Mapa de Saliência e CNN Filters, se comparada as demais.

Para o índice SD (Figura 38b), a combinação Mapa de Saliência e CNN Filters foi a que menos apresentou semelhanças entre as máscaras, de acordo com o valor máximo alcançado que foi de aproximadamente 0,2. Já a combinação entre Mapa de Saliência e Grad-CAM obteve menor variação entre os valores mínimo e máximo se comparado ao experimento anterior, indicando uma melhora nos resultados.

Figura 38 – Gráficos *boxplots* do segundo experimento referente aos índices IoU e SD das combinações entre os métodos de XAI.



(a) IoU



(b) SD

Elaborada pela autora.

A Tabela 5 apresenta a média dos índices obtida para cada combinação sobre a influência das três técnicas de perturbação de pixels. Todavia, apesar dos resultados alcançadas, os índices não podem ser observados de maneira isolada. Dessa forma, da mesma forma que no experimento anterior, foram considerados os valores obtidos para a média de iterações e a quantidade de imagens.

Tabela 5 – Valores obtidos para os índices IoU e SD em relação aos métodos de XAI combinados.

Método XAI combinados	**Técnicas de Perturbação**	**IoU**	**SD**
Mapa de Saliência + Grad-CAM	Aleatório	0,1125	0,3072
	Preto	0,1143	0,3048
	Ruído branco	0,1104	0,3015
Mapa de Saliência + Layer Grad-CAM	Aleatório	0,0969	0,3035
	Preto	0,0970	0,2972
	Ruído branco	0,0969	0,3021
Mapa de Saliêcia + CNN Filters	Aleatório	0,0931	0,1651
	Preto	0,0902	0,1619
	Ruído branco	0,0930	0,1654

Fonte – Elaborada pela autora.

Para uma análise mais aprofundada do presente experimento, também foi observada a média de iterações e a quantidade de imagens obtidas a partir do processo de predição explicada pela combinação dos métodos de XAI. Dessa forma, assim como no experimento anterior, foi considerada a quantidade máxima de cinco iterações. A partir da Tabela 6 é possível observar esses valores para cada uma das técnicas de perturbação de pixels aplicada, em cada um dos métodos de XAI implementados em combinação com o método Mapa de Saliência. Esses valores demonstram a melhora dos métodos em relação ao experimento anterior, no qual foi possível observar uma diferença significativa entre os métodos. Entretanto, para esse experimento, é possível observar que a operação lógica AND estabilizou os valores tornando-os melhores.

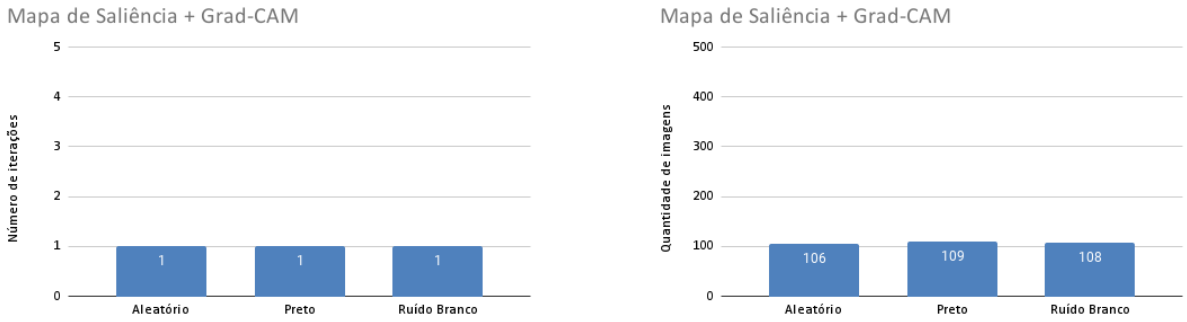
Tabela 6 – Média da quantidade de iterações e a quantidade de imagens por método de XAI combinado em relação às técnicas de perturbação de pixels implementadas.

Método XAI combinados	**Técnicas de Perturbação**	**Iterações**	**Imagens**
Mapa de Saliência + Grad-CAM	Aleatório	1	106
	Preto	1	109
	Ruído branco	1	108
Mapa de Saliência + Layer Grad-CAM	Aleatório	1	101
	Preto	1	106
	Ruído branco	1	102
Mapa de Saliência + CNN Filters	Aleatório	1	101
	Preto	1	107
	Ruído branco	1	101

Fonte – Elaborada pela autora.

O método Layer Grad-CAM, antes considerado o pior método explicável para o contexto de segmentação de imagens do peixe Pacu, foi o que obteve a mudança mais significativa em seus resultados, conforme observado na Figura 40. Portanto, em ambas as

Figura 39 – Média de iterações e quantidade de imagens obtidas por meio da combinação entre os métodos Mapa de Saliência e Grad-CAM.

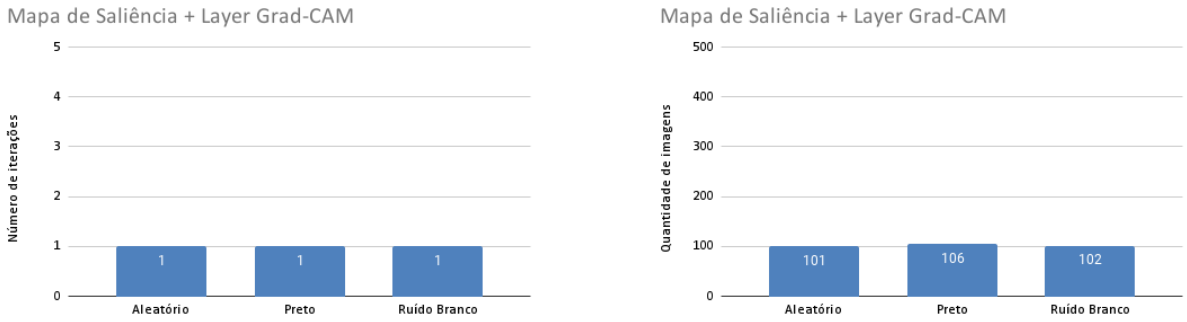


(a) Média da quantidade de iterações **(b)** Quantidade de imagens

Elaborada pela autora.

técnicas de perturbação o método havia apresentado valores acima de 400 para a quantidade de imagens resultantes, enquanto que para o experimento II, o método comportou-se muito melhor e com resultados bem abaixo, sendo o menor valor o de 101 imagens para a técnica de ruído aleatório (Figura 40b). Vale ressaltar que a quantidade mínima possível de ser obtida nesse modelo é de 100 imagens (sendo uma imagem resultante para cada uma das 100 imagens de entrada de peixes). Dessa forma, na Figura 40a é possível observar que para todos tipos de perturbação, o método realizou, em média, apenas uma iteração.

Figura 40 – Média de iterações e quantidade de imagens obtidas por meio da combinação entre os métodos Mapa de Saliência e Layer Grad-CAM.



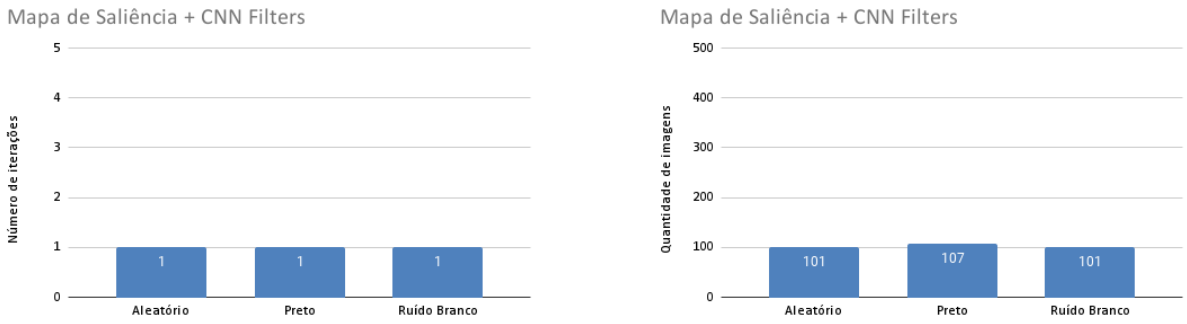
(a) Média da quantidade de iterações **(b)** Quantidade de imagens

Elaborada pela autora.

Acompanhando o mesmo desempenho obtido com os outros métodos, os resultados da combinação com o Mapa de Saliência para o método CNN Filters também foram satisfatórios. Assim, de acordo com a Figura 41, é possível observar que os valores foram melhores do que os alcançados no primeiro experimento. Assim, enquanto a menor quantidade de imagens obtidas pelo método foi de 182 para a perturbação de coloração preta no experimento I, a maior quantidade obtida no experimento II foi de 107 para o mesmo tipo de perturbação (Figura 41b). Isso demonstra que mesmo o pior valor obtido

supera o melhor resultado alcançado anteriormente. Além disso, a Figura 41a apresenta a média de iterações obtidas que de dois foi para um, confirmando o melhor desempenho do método ao ser combinado.

Figura 41 – Média de iterações e quantidade de imagens obtidas por meio da combinação entre os métodos Mapa de Saliência e CNN Filters.

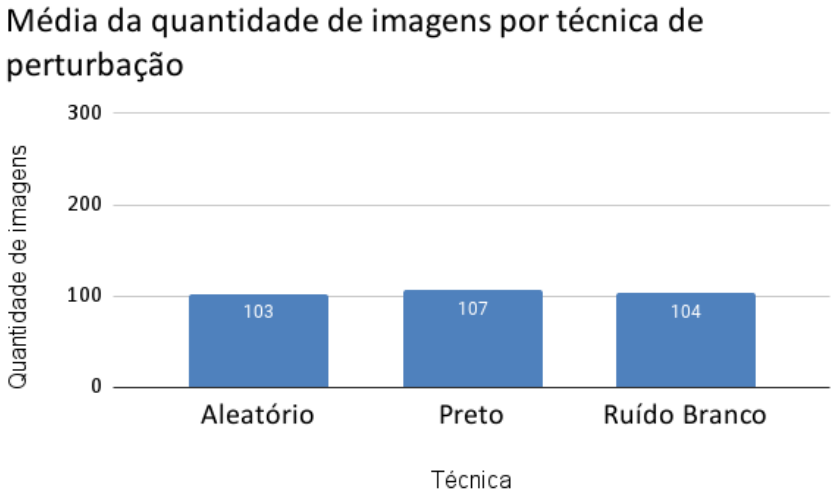


(a) Média da quantidade de iterações **(b)** Quantidade de imagens

Elaborada pela autora.

Apesar dos bons resultados, a técnica de perturbação de coloração preta manteve-se como sendo a pior dentre as três testadas. Em contrapartida, comparado ao primeiro experimento, essa técnica obteve uma melhora importante e que pode ser observada a partir do gráfico da Figura 42. Assim, enquanto a média da quantidade de imagens para o primeiro experimento foi de 300 imagens para ruído de cor preta, a média obtida no segundo experimento foi de 107. Além disso, foi observada anteriormente uma diferença significativa dessa técnica para as demais, enquanto que para o presente experimento, essa diferença foi sútil (diferença máxima de quatro imagens da média de cada técnica).

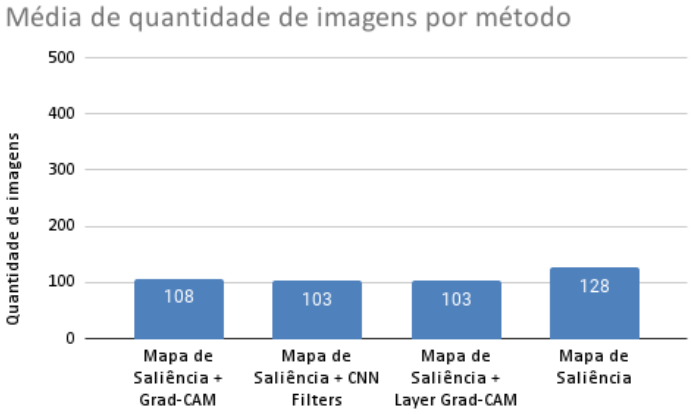
Figura 42 – Média da quantidade de imagens por técnica de perturbação de pixels para o segundo experimento.



Elaborada pela autora.

Em todos os casos apresentados, foi possível confirmar que a combinação serviu como fator determinante para a melhora dos três outros métodos explicáveis. Somado a isso, observou-se que mesmo os piores resultados desse experimento foram superiores aos resultados obtidos pelo próprio método Mapa de Saliência, considerado o melhor para o contexto apresentado. A Figura 43 apresenta um gráfico da média da quantidade de imagens obtida pelas combinações, em comparação com a média obtida pelo método Mapa de Saliência isoladamente. Observa-se que o valor de 128 imagens, considerado um bom resultado no primeiro experimento, foi superado pela outras três combinações. Esses resultados demonstram que as hipóteses da presente tese foram comprovadas, pois a partir da perturbação de pixels, o método Mapa de Saliência, avaliado como melhor, foi determinante para a melhora dos demais métodos de XAI implementados nesse experimento.

Figura 43 – Gráfico de barras da média da quantidade de imagens métodos.



Elaborada pela autora.

6 Conclusões e Trabalhos Futuros

A XAI vem sendo considerada como uma possível solução na busca pela transparência dos modelos caixa-preta. Entretanto, da mesma forma que é preciso verificar a procedência das predições dos modelos de IA, também é necessário analisar as explicações geradas pelos métodos explicáveis. Dessa forma, surgiu a necessidade de avaliar esses métodos para verificar a qualidade das explicações em relação ao resultados apresentados durante o processo de predição. Com isso, é possível determinar a melhor metodologia a ser considerada diante de um cenário específico, como por exemplo, a área da Saúde, cuja a explicação das imagens médicas deve ser realizada de maneira precisa.

Dentre as técnicas utilizadas para avaliação, para a presente tese foi considerada a técnica de perturbação de pixels, comumente utilizada para avaliar modelos que utilizam imagens para o conjunto de entrada. Dessa forma, a partir da alteração dos pixels mais importantes da imagem, foi possível verificar se a predição foi realizada normalmente e se as explicações obtidas foram influenciadas pelas alterações realizadas. Em caso positivo, é possível comprovar que o método explicável de fato apresenta as regiões mais importantes para o modelo durante o processo de predição. Em caso negativo, afirma-se que o método de XAI apresenta pixels diferentes do que aqueles que são realmente considerados pelos modelos de IA.

O primeiro experimento foi realizado a fim de avaliar os métodos Grad-CAM, Mapa de Saliência, CNN Filters, e Layer Grad-CAM no cenário de segmentação de imagens de peixes da espécie Pacu. De acordo com os resultados apresentados, diante das perturbações implementadas (cor preta, aleatório e ruído branco), o Mapa de Saliência obteve os melhores resultados ao demonstrar pixels que de fato foram considerados como os de maior relevância pela Mask R-CNN utilizada. A partir do resultado desse primeiro experimento, duas das três questões de pesquisa apresentadas nesse trabalho foram respondidas.

A técnica de perturbação de coloração preta, gerou mais iterações, e consequentemente mais imagens, em três dos quatro métodos XAI apresentados. Com isso, conclui-se que essa técnica possui a menor capacidade de gerar impacto ao modelo de segmentação dentre as técnicas testadas nessa tese. Também de acordo com a média de iterações e a média de imagens geradas, é possível concluir que o método Mapa de Saliência foi o menos sensível aos diferentes métodos de perturbação, enquanto que o método CNN Filters foi o menos sensível aos tipos de perturbação, demonstrando menor variação na média de quantidade de imagens. Por fim, o método Grad-CAM foi o mais sensível dentre os quatro, variando conforme a téncnica de perturbação utilizada.

Diante das investigações a cerca dos métodos de XAI, foi possível observar a grande

quantidade de metodologias explicáveis existentes para as diversas áreas de conhecimento. Essa variedade incentivou a busca por técnicas que proporcionam melhorias nos métodos de XAI, a fim de elevar a qualidade das explicações geradas. Com isso, o segundo experimento da presente tese teve como objetivo realizar a combinação, por meio da intersecção das máscaras resultantes, do melhor método de XAI (determinado no primeiro experimento) com os demais métodos implementados, a fim de melhorar suas explicações.

A partir dos resultados obtidos no segundo experimento, foi possível observar que a metodologia utilizada proporcionou não só a melhora dos métodos Grad-CAM, CNN Filters e Layer Grad-CAM, como também a melhora do próprio método Mapa de Saliência, já que os valores obtidos superaram a do melhor método. Esse resultado responde a última questão de pesquisa apresentada e corrobora as hipóteses apresentadas no início dessa tese. Dessa forma, conclui-se que é possível melhorar a qualidade das explicações de um método explicável aplicado a um modelo de segmentação de imagens de forma simples. Além disso, a combinação pode resultar na piora da explicabilidade do melhor método, quando esse é combinado com outro de baixa qualidade. O que não aconteceu no experimento apresentado nessa tese, já que o Layer Grad-CAM, considerado o pior método no experimento I, não influenciou a queda da eficácia do método do melhor método. Contudo, o Mapa de Saliência também teve uma melhora expressiva em seus resultados, indicando que a combinação dos métodos funcionou corretamente.

Para trabalhos futuros, sugere-se que os experimentos aqui apresentados sejam replicados em outros modelos de IA e em outros métodos de explicabilidade. Além disso, os valores limítrofes utilizados para determinar quais foram as regiões mais relevantes, podem ser estipulados utilizando outros mecanismos que não sejam a inspeção visual. Também é interessante testar outras técnicas de perturbação e também outras metodologias de avaliação existentes para verificar a qualidade das explicações. Dessa forma, é possível reunir maiores evidências sobre a eficácia dos métodos de XAI existentes na literatura. Sugere-se também realizar outros tipos de combinação lógica (e.g., OR E XOR) entre o Mapa de Saliência e os demais métodos, para obter novos resultados e verificar se houve melhorias na explicabilidade. Além disso, a aplicação do modelo de segmentação das imagens tal como realizada na etapa de treinamento do modelo (distinção entre corpo, cabeça, nadadeiras e *background*), também é vista como um experimento futuro. E por fim, também é possível aplicar essas hipóteses de pesquisa em outros contextos, como por exemplo classificação, e com *datasets* diferentes, a fim de ampliar os resultados obtidos nessa tese.

6.1 Publicações Derivadas da Tese

Com base na RSL, metodologia, experimentos e resultados apresentados nessa tese sobre a XAI, mais especificamente sobre a explicabilidade dos métodos de XAI aplicados a segmentação de imagens do peixe Pacu, foram publicados os trabalhos a seguir. Ambos são apresentados pelo nome, local de submissão e o estado de publicação até a defesa dessa tese:

- *eXplainable Artificial Intelligence in sentiment analysis of posts about Covid-19 vaccination on Twitter* (FEITOSA et al., 2023);
- *Influence of Pixel Perturbation on eXplainable Artificial Intelligence Methods* (VISAPP 2024 - Aceito);
- *Analysis of the explainability of eXplainable Artificial Intelligence methods applied to Pacu fish image segmentation* (*Applied Intelligence (APIN)* - Em revisão)

Referências

ABANDA, A.; MORI, U.; LOZANO, J. Ad-hoc explanation for time series classification. *Knowledge-Based Systems*, Elsevier, v. 252, p. 109366, 2022. Citado na página 46.

ADADI, A.; BERRADA, M. *Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160 (2018)*. 2018. Citado na página 32.

ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. Understanding of a convolutional neural network. In: IEEE. *2017 international conference on engineering and technology (ICET)*. [S.l.], 2017. p. 1–6. Citado 2 vezes nas páginas 25 e 26.

ALI, S. et al. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, Elsevier, v. 99, p. 101805, 2023. Citado na página 45.

ARNOUT, H. et al. Towards a rigorous evaluation of xai methods on time series. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. [S.l.: s.n.], 2019. p. 4197–4201. ISSN 2473-9936. Citado na página 18.

BASHA, S. S. et al. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, Elsevier, v. 378, p. 112–119, 2020. Citado na página 27.

CAMACHO, D. M. et al. Next-generation machine learning for biological networks. *Cell*, Elsevier, v. 173, n. 7, p. 1581–1592, 2018. Citado na página 18.

CARDOSO, A. J. d. S. et al. Estimation of genetic parameters for body areas in nile tilapia measured by digital image analysis. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 138, n. 6, p. 731–738, 2021. Citado na página 38.

CHATTERJEE, S. et al. Torchesegeta: Framework for interpretability and explainability of image-based deep learning models. *Applied Sciences*, MDPI, v. 12, n. 4, p. 1834, 2022. Citado na página 34.

CHENG, T. et al. Boundary-preserving mask r-cnn. In: SPRINGER. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. [S.l.], 2020. p. 660–676. Citado 3 vezes nas páginas 28, 29 e 30.

CHIAO, J.-Y. et al. Detection and classification the breast tumors using mask r-cnn on sonograms. *Medicine*, Wolters Kluwer Health, v. 98, n. 19, 2019. Citado na página 29.

COULIBALY, S. et al. Explainable deep convolutional neural networks for insect pest recognition. *Journal of Cleaner Production*, Elsevier, v. 371, p. 133638, 2022. Citado 5 vezes nas páginas 48, 49, 50, 63 e 64.

DENG, L.; YU, D. Deep learning: methods and applications. *Foundations and trends in signal processing*, Now Publishers Inc. Hanover, MA, USA, v. 7, n. 3–4, p. 197–387, 2014. Citado na página 18.

DÍAZ-RODRÍGUEZ, N. et al. Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Information Fusion*, Elsevier, v. 79, p. 58–83, 2022. Citado na página 44.

DICE, L. R. Measures of the amount of ecologic association between species. *Ecology*, JSTOR, v. 26, n. 3, p. 297–302, 1945. Citado na página 65.

DING, W. et al. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, Elsevier, 2022. Citado na página 45.

DONG, X. et al. Robust superpixel-guided attentional adversarial attack. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2020. p. 12895–12904. Citado na página 39.

DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. Citado 2 vezes nas páginas 19 e 20.

ELER, D. M. et al. Visual approach to support analysis of optimum-path forest classifier. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2019. Citado na página 32.

ELGUENDOUZE, S. et al. Explainability in image captioning based on the latent space. *Neurocomputing*, Elsevier, v. 546, p. 126319, 2023. Citado na página 48.

ERHAN, D. et al. Visualizing higher-layer features of a deep network. *University of Montreal*, v. 1341, n. 3, p. 1, 2009. Citado 2 vezes nas páginas 35 e 36.

FAO, I. et al. *The state of world fisheries and aquaculture 2016*. [S.l.]: Publications of Food and Agriculture Organization of the United Nations Rome, 2016. 200 p. Citado na página 37.

FEITOSA, J. D. C. et al. explainable artificial intelligence in sentiment analysis of posts about covid-19 vaccination on twitter. In: *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*. [S.l.: s.n.], 2023. p. 65–72. Citado na página 82.

FELLOUS, J.-M. et al. Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. *Frontiers in Neuroscience*, v. 13, 2019. Cited By 0. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077335140&doi=10.3389%2ffnins.2019.01346&partnerID=40&md5=e2389a13dd1c6f249124ddfaad761e07>. Citado 2 vezes nas páginas 18 e 31.

FERNANDES, A. F. et al. Deep learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in nile tilapia. *Computers and electronics in agriculture*, Elsevier, v. 170, p. 105274, 2020. Citado 2 vezes nas páginas 17 e 38.

FONG, R. C.; VEDALDI, A. Interpretable explanations of black boxes by meaningful perturbation. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 3429–3437. Citado 2 vezes nas páginas 39 e 40.

FREITAS, M. V. et al. High-throughput phenotyping by deep learning to include body shape in the breeding program of pacu (piaractus mesopotamicus). *Aquaculture*, Elsevier, v. 562, p. 738847, 2023. Citado 2 vezes nas páginas 17 e 38.

FREITAS, M. V. et al. Genotype by environment interaction and genetic parameters for growth traits in the neotropical fish pacu (piaractus mesopotamicus). *Aquaculture*, Elsevier, v. 530, p. 735933, 2021. Citado na página 37.

FUKUSHIMA, K. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, Elsevier, v. 1, n. 2, p. 119–130, 1988. Citado na página 24.

GHASSEMI, M.; OAKDEN-RAYNER, L.; BEAM, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, Elsevier, v. 3, n. 11, p. e745–e750, 2021. Citado 2 vezes nas páginas 19 e 20.

GIPIŠKIS, R. et al. The impact of adversarial attacks on interpretable semantic segmentation in cyber-physical systems. *IEEE Systems Journal*, IEEE, 2023. Citado na página 46.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1. Citado na página 17.

GOROKHOVATSKYI, O.; PEREDRII, O. Multiclass image classification explanation with the complement perturbation images. In: SPRINGER. *Data Stream Mining & Processing: Third International Conference, DSMP 2020, Lviv, Ukraine, August 21–25, 2020, Proceedings 3*. [S.l.], 2020. p. 275–287. Citado 2 vezes nas páginas 39 e 40.

GU, J. et al. Recent advances in convolutional neural networks. *Pattern recognition*, Elsevier, v. 77, p. 354–377, 2018. Citado na página 24.

GUMPFER, N. et al. Signed explanations: Unveiling relevant features by reducing bias. *Information Fusion*, Elsevier, p. 101883, 2023. Citado 3 vezes nas páginas 46, 48 e 49.

GUNNING, D. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, v. 2, p. 2, 2017. Citado 3 vezes nas páginas 19, 31 e 32.

GUNNING, D.; AHA, D. Darpa's explainable artificial intelligence program. *AI Magazine*, v. 40, n. 2, p. 44–58, 2019. Cited By 6. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069460356&doi=10.1609%2faimag.v40i2.2850&partnerID=40&md5=6e65468f94dffe77176516ab6d991363>. Citado na página 18.

GUPTA, L. K.; KOUNDAL, D.; MONGIA, S. Explainable methods for image-based deep learning: a review. *Archives of Computational Methods in Engineering*, Springer, v. 30, n. 4, p. 2651–2666, 2023. Citado na página 19.

HAO, W. et al. The role of activation function in cnn. In: IEEE. *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*. [S.l.], 2020. p. 429–432. Citado na página 25.

HARMON, P.; MAUS, R.; MORRISSEY, W. *Expert systems: tools and applications*. [S.l.]: John Wiley & Sons, Inc., 1988. Citado 2 vezes nas páginas 16 e 17.

HE, K. et al. Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 2961–2969. Citado 3 vezes nas páginas 28, 29 e 30.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado na página 27.

HENDRYCKS, D.; DIETTERICH, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. Citado na página 19.

HOCHULI, J. et al. Visualizing convolutional neural network protein-ligand scoring. *Journal of Molecular Graphics and Modelling*, Elsevier, v. 84, p. 96–108, 2018. Citado na página 36.

HODGE, V.; AUSTIN, J. A survey of outlier detection methodologies. *Artificial intelligence review*, Springer, v. 22, n. 2, p. 85–126, 2004. Citado na página 17.

HUBEL, D. H.; WIESEL, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, Wiley-Blackwell, v. 160, n. 1, p. 106, 1962. Citado na página 24.

IBRAHIM, R.; SHAFIQ, M. O. Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 10, p. 1–37, 2023. Citado na página 45.

IVANOVS, M.; KADIKIS, R.; OZOLS, K. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, Elsevier, v. 150, p. 228–234, 2021. Citado 2 vezes nas páginas 19 e 45.

JACCARD, P. The distribution of the flora in the alpine zone. 1. *New phytologist*, Wiley Online Library, v. 11, n. 2, p. 37–50, 1912. Citado na página 65.

JIN, W. et al. Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical Image Analysis*, Elsevier, v. 84, p. 102684, 2023. Citado na página 44.

KADIR, M. A. et al. A user interface for explaining machine learning model explanations. In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. [S.l.: s.n.], 2023. p. 59–63. Citado na página 46.

KAMATH, U. et al. Convolutional neural networks. *Deep learning for NLP and speech recognition*, Springer, p. 263–314, 2019. Citado na página 26.

KAUFMAN, D. *A inteligência artificial irá suplantar a inteligência humana?* [S.l.]: ESTAÇÃO DAS LETRAS E CORES EDI, 2019. Citado na página 16.

KAUR, H. et al. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. [S.l.: s.n.], 2020. p. 1–14. Citado na página 20.

KHAN, M. E.; KHAN, F. et al. A comparative study of white box, black box and grey box testing techniques. *Int. J. Adv. Comput. Sci. Appl*, Citeseer, v. 3, n. 6, 2012. Citado na página 31.

KISTAN, T.; GARDI, A.; SABATINI, R. Machine learning and cognitive ergonomics in air traffic management: Recent developments and considerations for certification. *Aerospace*, v. 5, n. 4, 2018. Cited By 3. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062152075&doi=10.3390%2faerospace5040103&partnerID=40&md5=d114d4dfecb745aaba6883aadc2cb081>. Citado na página 16.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. 2007. Citado 2 vezes nas páginas 41 e 42.

KUCKLICK, J.-P.; MÜLLER, O. Tackling the accuracy-interpretability trade-off: Interpretable deep learning models for satellite image-based real estate appraisal. *ACM Transactions on Management Information Systems*, ACM New York, NY, v. 14, n. 1, p. 1–24, 2023. Citado na página 47.

LECUN, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural computation*, MIT Press, v. 1, n. 4, p. 541–551, 1989. Citado na página 24.

LEE, G.; LEE, S. J.; LEE, C. A convolutional neural network model for abnormality diagnosis in a nuclear power plant. *Applied Soft Computing*, Elsevier, v. 99, p. 106874, 2021. Citado 2 vezes nas páginas 24 e 26.

LEE, H. et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: *Proceedings of the 26th annual international conference on machine learning*. [S.l.: s.n.], 2009. p. 609–616. Citado na página 36.

LIANG, Y.; LI, M.; JIANG, C. Generating self-attention activation maps for visual interpretations of convolutional neural networks. *Neurocomputing*, Elsevier, v. 490, p. 206–216, 2022. Citado na página 48.

LIN, Y.-S.; LEE, W.-C.; CELIK, Z. B. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. [S.l.: s.n.], 2021. p. 1027–1035. Citado na página 46.

LISBOA, P. et al. The coming of age of interpretable and explainable machine learning models. *Neurocomputing*, Elsevier, v. 535, p. 25–39, 2023. Citado na página 45.

MCCARTHY, J. et al. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, v. 27, n. 4, p. 12–12, 2006. Citado na página 16.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943. Citado na página 16.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943. Citado na página 17.

MICHIE, D. et al. Machine learning. *Neural and Statistical Classification*, Technometrics, v. 13, n. 1994, p. 1–298, 1994. Citado na página 17.

MINAEE, S. et al. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 44, n. 7, p. 3523–3542, 2021. Citado na página 19.

MINSKY, M.; PAPERT, S. A. *Perceptrons, reissue of the 1988 expanded edition with a new foreword by Léon Bottou: an introduction to computational geometry*. [S.l.]: MIT press, 2017. Citado na página 17.

MITTELSTADT, B.; RUSSELL, C.; WACHTER, S. Explaining explanations in ai. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2019. (FAT* '19), p. 279–288. ISBN 9781450361255. Disponível em: <https://doi.org/10.1145/3287560.3287574>. Citado na página 30.

MOHAMED, E.; SIRLANTZIS, K.; HOWELLS, G. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. *Displays*, Elsevier, v. 73, p. 102239, 2022. Citado 3 vezes nas páginas 33, 35 e 46.

MOHSENI, S.; ZAREI, N.; RAGAN, E. D. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, ACM New York, NY, v. 11, n. 3-4, p. 1–45, 2021. Citado na página 44.

MOOR, J. *The Turing test: the elusive standard of artificial intelligence*. [S.l.]: Springer Science & Business Media, 2003. v. 30. Citado na página 16.

MUDDAMSETTY, S. M. et al. Visual explanation of black-box model: Similarity difference and uniqueness (sidu) method. *Pattern recognition*, Elsevier, v. 127, p. 108604, 2022. Citado 2 vezes nas páginas 46 e 47.

MUTHUKRISHNAN, M. et al. The future of artificially intelligent assistants. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2017. p. 33–34. Citado na página 16.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. [S.l.: s.n.], 2010. p. 807–814. Citado na página 25.

NEMATZADEH, H. et al. Ensemble-based genetic algorithm explainer with automized image segmentation: A case study on melanoma detection dataset. *Computers in Biology and Medicine*, Elsevier, v. 155, p. 106613, 2023. Citado na página 48.

NGUYEN, A.; YOSINSKI, J.; CLUNE, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 427–436. Citado na página 39.

NIBALI, A.; HE, Z.; WOLLERSHEIM, D. Pulmonary nodule classification with deep residual networks. *International journal of computer assisted radiology and surgery*, Springer, v. 12, p. 1799–1808, 2017. Citado 2 vezes nas páginas 27 e 28.

PAPERNOT, N. et al. The limitations of deep learning in adversarial settings. In: IEEE. *2016 IEEE European symposium on security and privacy (EuroS&P)*. [S.l.], 2016. p. 372–387. Citado na página 39.

PARSIFAL. *Perform Systematic Literature Reviews*. 2020. Https://parsif.al/about/. Acesso 01-10-2020. Citado na página 42.

PINHEIRO, P. P. *Proteção de dados pessoais: Comentários à lei n. 13.709/2018-lgpd*. [S.l.]: Saraiva Educação SA, 2020. Citado na página 18.

QUIONERO-CANDELA, J. et al. *Dataset shift in machine learning*. [S.l.]: The MIT Press, 2009. Citado na página 17.

RAS, G. et al. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, v. 73, p. 329–396, 2022. Citado 2 vezes nas páginas 33 e 46.

RAUBER, T. W. Redes neurais artificiais. *Universidade Federal do Espírito Santo*, p. 29, 2005. Citado na página 17.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144. Citado na página 39.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958. Citado na página 17.

RUSSELL, S. J.; NORVIG, P. *Inteligência artificial*. [S.l.]: Elsevier, 2004. Citado na página 16.

SAE-LIM, P. et al. Genotype-by-environment interaction of growth traits in rainbow trout (oncorhynchus mykiss): a continental scale study. *Journal of Animal Science*, Oxford University Press, v. 91, n. 12, p. 5572–5581, 2013. Citado na página 37.

SALAHUDDIN, Z. et al. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, Elsevier, v. 140, p. 105111, 2022. Citado na página 44.

SALEEM, H.; SHAHID, A. R.; RAZA, B. Visual interpretability in 3d brain tumor segmentation network. *Computers in Biology and Medicine*, Elsevier, v. 133, p. 104410, 2021. Citado na página 44.

SAMPAIO, A. Improving systematic mapping reviews. *ACM SIGSOFT Software Engineering Notes*, ACM New York, NY, USA, v. 40, n. 6, p. 1–8, 2015. Citado na página 41.

SCHLEGEL, U. et al. Towards a rigorous evaluation of xai methods on time series. In: IEEE. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. [S.l.], 2019. p. 4197–4201. Citado na página 46.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 618–626. Citado 2 vezes nas páginas 33 e 34.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, Springer US New York, v. 128, n. 2, p. 336–359, 2020. Citado 2 vezes nas páginas 33 e 34.

SHAFIQ, M.; GU, Z. Deep residual learning for image recognition: A survey. *Applied Sciences*, MDPI, v. 12, n. 18, p. 8972, 2022. Citado na página 27.

SHI, R.; LI, T.; YAMAGUCHI, Y. Understanding contributing neurons via attribution visualization. *Neurocomputing*, Elsevier, p. 126492, 2023. Citado na página 46.

SHOJAEI, S.; ABADEH, M. S.; MOMENI, Z. An evolutionary explainable deep learning approach for alzheimer's mri classification. *Expert Systems with Applications*, Elsevier, v. 220, p. 119709, 2023. Citado na página 47.

SHU, J.-H. et al. An improved mask r-cnn model for multiorgan segmentation. *Mathematical Problems in Engineering*, Hindawi Limited, v. 2020, p. 1–11, 2020. Citado na página 30.

SIMONYAN, K.; VEDALDI, A.; ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. Citado 2 vezes nas páginas 35 e 36.

SORENSEN, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, v. 5, p. 1–34, 1948. Citado na página 65.

SOUSA, I. P. de; VELLASCO, M. M.; SILVA, E. C. da. Evolved explainable classifications for lymph node metastases. *Neural Networks*, Elsevier, v. 148, p. 1–12, 2022. Citado na página 48.

STOCK, J.; DOLAN, A.; CAVEY, T. Strategies for robust image classification. *arXiv e-prints*, p. arXiv–2004, 2020. Citado na página 39.

SZEGEDY, C. et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. Citado na página 39.

TEIXEIRA, J. *O que é inteligência artificial*. [S.l.]: E-Galáxia, 2019. Citado na página 16.

TURING, A. M. *Computing machinery and intelligence*. [S.l.]: Springer, 2009. Citado na página 16.

VEERAPPA, M. et al. Validation of xai explanations for multivariate time series classification in the maritime domain. *Journal of Computational Science*, Elsevier, v. 58, p. 101539, 2022. Citado na página 46.

VIEIRA, C. P.; DIGIAMPIETRI, L. A. Machine learning post-hoc interpretability: a systematic mapping study. In: *XVIII Brazilian Symposium on Information Systems*. [S.l.: s.n.], 2022. p. 1–8. Citado na página 44.

VILONE, G.; LONGO, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, Elsevier, v. 76, p. 89–106, 2021. Citado na página 44.

WEBER, R. O. et al. Investigating textual case-based XAI. In: COX, M. T.; FUNK, P.; BEGUM, S. (Ed.). *Case-Based Reasoning Research and Development*. Springer International Publishing, 2018. v. 11156, p. 431–447. ISBN 978-3-030-01080-5 978-3-030-01081-2. Series Title: Lecture Notes in Computer Science. Disponível em: <http://link.springer.com/10.1007/978-3-030-01081-2_29>. Citado 3 vezes nas páginas 18, 31 e 32.

WOLF, C. T. Explainability scenarios: Towards scenario-based xai design. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. New York, NY, USA: Association for Computing Machinery, 2019. (IUI '19), p. 252–257. ISBN 9781450362726. Disponível em: <https://doi.org/10.1145/3301275.3302317>. Citado na página 18.

WOLF, C. T.; RINGLAND, K. E. Designing accessible, explainable ai (xai) experiences. *SIGACCESS Access. Comput.*, Association for Computing Machinery, New York, NY, USA, n. 125, mar. 2020. ISSN 1558-2337. Disponível em: <https://doi.org/10.1145/3386296.3386302>. Citado 3 vezes nas páginas 18, 31 e 32.

WU, S.; ZHONG, S.; LIU, Y. Deep residual learning for image steganalysis. *Multimedia tools and applications*, Springer, v. 77, p. 10437–10453, 2018. Citado na página 27.

XIE, G.; YANG, K.; LAI, J. Filter-in-filter: low cost cnn improvement by sub-filter parameter sharing. *Pattern Recognition*, Elsevier, v. 91, p. 391–403, 2019. Citado na página 37.

XU, F. et al. Explainable AI: A brief survey on history, research areas, approaches and challenges. In: TANG, J. et al. (Ed.). *Natural Language Processing and Chinese Computing*. Springer International Publishing, 2019. v. 11839, p. 563–574. ISBN 978-3-030-32235-9 978-3-030-32236-6. Series Title: Lecture Notes in Computer Science. Disponível em: <http://link.springer.com/10.1007/978-3-030-32236-6_51>. Citado 2 vezes nas páginas 18 e 31.

YAMASHITA, R. et al. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, Springer, v. 9, p. 611–629, 2018. Citado 4 vezes nas páginas 24, 25, 26 e 27.

YANG, X. et al. Deep learning for smart fish farming: applications, opportunities and challenges. *Reviews in Aquaculture*, Wiley Online Library, v. 13, n. 1, p. 66–90, 2021. Citado na página 38.

YU, W. et al. Visualizing and comparing convolutional neural networks. *arXiv preprint arXiv:1412.6631*, 2014. Citado na página 37.

ZAFAR, M. R.; KHAN, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, MDPI, v. 3, n. 3, p. 525–541, 2021. Citado na página 48.

ZHANG, A. et al. *Dive into deep learning*. [S.l.]: Cambridge University Press, 2023. Citado 4 vezes nas páginas 24, 27, 28 e 29.

ZHANG, W. et al. An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. *Medical Physics*, Wiley Online Library, v. 23, n. 4, p. 595–601, 1996. Citado na página 24.

ZHAO, S. et al. Application of machine learning in intelligent fish aquaculture: A review. *Aquaculture*, Elsevier, v. 540, p. 736724, 2021. Citado na página 38.

ZHOU, B. et al. Learning deep features for discriminative localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 2921–2929. Citado na página 33.