UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

FABRÍCIO STEINLE AMOROSO

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL COM LIME E SHAP APLICADA À REDE NEURAL CONVOLUCIONAL

BAURU
Janeiro/2023

FABRÍCIO STEINLE AMOROSO

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL COM LIME E SHAP APLICADA À REDE NEURAL CONVOLUCIONAL

Trabalho de Conclusão de Curso do Curso de Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.
Orientador: Prof. Dr. Clayton Reginaldo Pereira

BAURU
Janeiro/2023

A524i

Amoroso, Fabrício Steinle
Inteligência Artificial Explicável com LIME e SHAP aplicada à Rede Neural Convolucional / Fabrício Steinle Amoroso. -- Bauru, 2023
48 p. : il.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru
Orientador: Clayton Reginaldo Pereira

1. Inteligência Artificial Explicável. 2. Rede Neural Convolucional. 3. Parkinson. 4. SHAP. 5. LIME. I. Título.

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

Essa ficha não pode ser modificada.

Fabrício Steinle Amoroso

Inteligência Artificial Explicável com LIME e SHAP aplicada à Rede Neural Convolucional

Trabalho de Conclusão de Curso do Curso de Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Dr. Clayton Reginaldo Pereira
Orientador
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Prof. Dra. Simone Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Prof. Dr. Douglas Rodrigues
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Ciência da Computação

Bauru, _____ de _____ de _____.

Dedico o trabalho aos meus pais, que certamente olham para este com muito orgulho. O mérito também é de vocês.

Agradecimentos

Agradeço primeiramente meus pais, Irene e João, pelo amor e apoio contínuo que foram essenciais em todas as etapas da minha vida.

Agradeço Beatriz, minha namorada e parceira, que esteve ao meu lado em todos os momentos. Espero que possamos continuar juntos por muito mais tempo.

Agradeço também aos meus colegas e amigos que tornaram a experiência na faculdade e na vida muito mais feliz e proveitosa, que a amizade seja mantida nas próximas etapas de nossas jornadas.

Por fim, agradeço todos os professores que participaram da minha formação e que despertaram a vontade continuar sempre aprendendo.

"Quem caminha sozinho pode até chegar mais rápido, mas aquele que vai acompanhado, com certeza vai mais longe."

Clarice Lispector

Resumo

Modelos de inteligência artificial estão presentes na vida cotidiana nos mais diversos contextos, como sistemas médicos para auxílio na detecção de doenças e motores de busca, estando, por vezes, presentes até de maneira transparente aos usuários como no caso de algoritmos de recomendação de produtos. Ao passo que a adoção de IA cresce, a complexidade dos sistemas de inteligência artificial também aumenta, tornando mais desafiadora a tarefa de compreender como foi obtido determinado resultado. Refere-se a estes modelos complexos como caixa-preta, devido à sua dificuldade de interpretação. Inteligência artificial explicável pode ser utilizada para compreender como os modelos complexos, como redes neurais convolucionais, que são amplamente aplicados, chegam a seus resultados. É proposto neste projeto de conclusão de curso, implementar técnicas de inteligência artificial explicável utilizando duas das ferramentas mais populares neste contexto: LIME e SHAP, ambas aplicadas a um modelo de rede neural convolucional utilizado para classificar imagens de exames médicos de escrita, pertencentes a um grupo de indivíduos saudáveis e outro grupo de pacientes de Parkinson. Através dos resultados obtidos foi possível obter explicações sobre o modelo descrito que podem ser interpretadas por seres humanos.

Palavras-chave: Inteligência Artificial Explicável, LIME, Parkinson, Rede Neural Convolucional, SHAP.

Abstract

Artificial intelligence models are present in everyday life in the most diverse contexts, such as medical systems to aid in the detection of diseases and search engines, and are sometimes even transparently present to users, as in the case of product recommendation algorithms. As the adoption of AI grows, the complexity of artificial intelligence systems also increases, making the task of understanding how a given result was achieved more challenging. These complex models are referred to as 'black boxes', due to their difficulty in interpretation. Explainable artificial intelligence can be used to understand how complex models, such as convolutional neural networks, which are widely applied, arrive at their results. It is proposed in this thesis, to implement explainable artificial intelligence techniques using two of the most popular tools in this context: LIME and SHAP, both applied to a convolutional neural network model used to classify images of medical handwriting exams, belonging to a group of healthy individuals and another group of Parkinson's patients. Through the obtained results it was possible to obtain explanations about the described model that can be interpreted by human beings.
Keywords: Convolutional Neural Network, Explainable Artificial Intelligence, LIME, Parkinson, SHAP.

Lista de figuras

Sumário

1 Introdução

Aprendizado de máquina, do inglês *Machine Learning* (ML), é um ramo de inteligência artificial que utiliza dados e algoritmos para tentar imitar a forma que humanos aprendem.(IBM, 2022d) Um modelo de ML é treinado através de dados para reconhecer padrões entre os dados utilizados para seu treino, esperando que seja capaz de inferir resultados corretamente ao se deparar com dados fora de seu treinamento. (MICROSOFT, 2022)

Pode-se entender como 'modelo' a representação matemática dos objetos presentes em um conjunto de dados e seus relacionamentos, buscando entender de que forma cada um de seus atributos afetam e pesam nos resultados. Entende-se como atributos qualquer coisa, desde números de curtidas ou compartilhamentos em redes sociais, tamanho em metros e quantidade de quartos de uma casa e até moléculas em experimentos de laboratório. Os atributos existentes em cada modelo dependem principalmente do objetivo do mesmo e do contexto dos dados sendo utilizados. (PARSONS, 2021)

As soluções utilizando aprendizado de máquina são aplicadas de diversas formas e impactam diretamente na maneira que a humanidade vive e evolui. Algoritmos de recomendação de produtos, motores de busca, propagandas, redes sociais, assistentes virtuais e sistemas médicos para auxílio na detecção de doenças são exemplos de aplicações utilizadas extensivamente, mesmo que de forma imperceptível pelos usuários. (THUNG et al., 2012)

Como afirma Zitnik et al. (2019), as tecnologias de aprendizado de máquina possibilitaram pesquisas biológicas e de saúde humana em uma escala sem precedentes. Nem sempre é possível relacionar precisamente todos os fatores envolvidos nesses temas por meio de modelagens matemáticas simples. Métodos capazes de trazer sentido e encontrar relações de alta complexidade em meio aos dados possibilitam o entendimento de fenômenos complexos, como doenças, que podem depender de diversos fatores para ocorrerem.

Vale ressaltar ainda, que técnicas de ML podem ser utilizadas para análise de créditos em sistemas bancários, em sistemas de prevenção a fraude, analisando características e tendências que indiquem transações fraudulentas dentro de serviços e até mesmo na produção de outros sistemas de *software*, potencialmente sendo empregadas para detectar erros de escrita, sugerir correções, chegando ainda a estarem aptas a criar integralmente um código para determinado fim.

Todas as utilidades citadas acima reforçam o fato de que sistemas de ML já são amplamente utilizados e tendem a continuar crescendo em uso, aplicabilidade e desempenho. De acordo com Dwarakanath et al. (2018), espera-se que em um futuro próximo, a maioria das aplicações estarão utilizando e se beneficiarão de alguma forma de técnicas de ML.

Ao passo que a adoção de inteligência artificial (IA) cresce, a complexidade dos sistemas de inteligência artificial também aumenta, tornando mais desafiadora a tarefa de compreender como foi obtido determinado resultado. Refere-se a estes modelos complexos como caixa-preta, devido à sua dificuldade de interpretação. (IBM, 2022b)

Inteligência Artificial Explicável, do inglês *Explainable AI* (XAI), diz respeito aos métodos e técnicas aplicados à sistemas de inteligência artificial, com o objetivo de melhorar a compreensão para seres humanos dos resultados e decisões obtidas por esses sistemas. (ANKARSTAD, 2020)

> "Existem muitas vantagens em entender como um sistema de IA chegou a uma saída específica. Explicabilidade pode ajudar desenvolvedores a garantir que o sistema esteja funcionando como se espera, pode ser necessário alcançar padrões regulatórios, ou pode ser importante possibilitar que aqueles que são afetados por uma decisão confrontem ou mudem aquele resultado."[1] (IBM, 2022b)

Para alcançar a explicabilidade, existem ferramentas como LIME e SHAP, que são agnósticos a modelos e capazes de gerar explicações locais, bem como globais; sendo essas, as mais utilizadas para o fim de obter explicações.

Um problema que pode ser abordado nesse sentido, seria explicar os resultados de um sistema de inteligência artificial complexo, como um modelo de classificação de imagens, com o objetivo de predizer se determinado exame médico pertence a um indivíduo saudável, ou a um paciente acometido por Parkinson, que é uma doença neurológica que afeta os movimentos, causando alterações na escrita.

1.1 Problema

> "À medida que IA se torna mais avançada, os humanos são desafiados a compreender e refazer o caminho de como o algoritmo chegou a um determinado resultado. Todo o processo de cálculo se torna no que é comumente chamado de "caixa-preta": impossível de interpretar. Esses modelos de 'caixa preta' são criados diretamente a partir de dados. E nem mesmo os engenheiros ou cientistas de dados que criam o algoritmo conseguem entender ou explicar o que exatamente está acontecendo dentro deles ou como o algoritmo de IA chegou a um resultado específico."[2] (IBM, 2022b)

[1] Tradução realizada pelo autor. Texto original: "There are many advantages to understanding how an AI-enabled system has led to a specific output. Explainability can help developers ensure that the system is working as expected, it might be necessary to meet regulatory standards, or it might be important in allowing those affected by a decision to challenge or change that outcome.".

[2] Tradução realizada pelo autor. Texto original: "As AI becomes more advanced, humans are challenged to comprehend and retrace how the algorithm came to a result. The whole calculation process is turned into what is commonly referred to as a “black box"that is impossible to interpret. These black box models are created directly from the data. And, not even the engineers or data scientists who create the algorithm can understand or explain what exactly is happening inside them or how the AI algorithm arrived at a specific result.".

O elevado poder de tomada de decisões de forma automatizada de inteligências artificiais é um grande atrativo para sua aplicação nos negócios; porém, existindo a possibilidade de erros, dar total poder de escolha aos modelos, pode causar grandes impactos inesperados. As mesmas empresas que buscam os benefícios que a implantação de sistemas de IA podem trazer, ainda não se sentem confortáveis em deixá-los tomar grandes decisões, tendo em vista a dificuldade de compreender todos os fatores envolvidos no julgamento daquele modelo de IA, portanto não tem total confiança em seus resultados. (ANKARSTAD, 2020)

Em outros contextos, como aplicações de inteligência artificial na área médica, é vital que seja possível compreender os resultados de um modelo. Quando uma decisão da IA necessitar ser confrontada deve ser possível ter entendimento dos fatores que carregaram determinado resultado. Bem como, ao apresentar determinado resultado obtido através de tais sistemas, não seria possível transmitir confiança caso não haja transparência quanto às motivações que embasaram aquela decisão.

1.2 Justificativa

Sabendo das dificuldades envolvidas na aplicação de IA, bem como o fato de que grande parte das empresas buscam incorporá-la em seus negócios, já utilizam, ou até mesmo estão aprimorando e substituindo processos já existentes com ajuda dessas tecnologias, se faz necessário compreender como os modelos tomam decisões e ter transparência quanto aos fatores envolvidos.

Inteligência Artificial Explicável é uma solução para alcançar essa transparência e compreensão de modelos, fatores que se fazem necessários para continuar aplicando IA com maior confiança. (ANKARSTAD, 2020)

XAI pode ser utilizada por empresas e desenvolvedores de inteligência artificial para ajudar as partes interessadas a entender o comportamento da IA, auxiliando no processo de solucionar problemas existentes e melhorar o desempenho de modelos. (IBM, 2022b). Desta forma, justifica-se a importância de pesquisas que buscam explorar o assunto.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo deste trabalho consiste em aplicar técnicas de explicabilidade de inteligência artificial em um modelo de classificação de imagens, a fim de analisar os fatores que têm maior contribuição para as decisões daquele modelo e, a partir disso, ser capaz de compreender as decisões e resultados obtidos. Espera-se, juntamente a isso, ser possível pontuar a importância e reforçar a capacidade das técnicas de XAI de garantir a auditabilidade e auxiliar a plena

utilização de inteligência artificial nas suas diversas áreas de aplicação.

Para isto, foi utilizada a linguagem de programação Python, as bibliotecas Keras, Numpy, Matplotlib e Scikit-image e as ferramentas de explicabilidade LIME e SHAP. A modelagem foi feita com uma rede neural convolucional (CNN).

O foco desta pesquisa não é a performance dos modelos utilizados ou implementados, mas sim explorar a capacidade de ferramentas como LIME e SHAP de explicar suas decisões e de auxiliar na sua interpretação.

1.3.2 Objetivos Específicos

- Realizar o levantamento das principais ferramentas de XAI e abordar seu funcionamento.
- Expor e analisar os possíveis benefícios da aplicação de técnicas de explicabilidade em modelos.
- Implementar as técnicas de XAI no modelo de classificação de exames médicos de indivíduos saudáveis e pacientes com Parkinson.
- Avaliar os resultados obtidos com a aplicação.

1.4 Organização do Trabalho

O presente trabalho segue a seguinte estrutura: No capítulo 2 está a fundamentação teórica necessária para o entendimento do trabalho, apresentando brevemente sobre Parkinson e informações sobre o que é aprendizado de máquina e seus principais conceitos, explicações sobre modelos de ML e, mais ao final do capítulo, sobre inteligência artificial explicável. A fundamentação expõe com mais detalhes os modelos e ferramentas de XAI utilizados neste trabalho.

O capítulo 3 trata da metodologia utilizada para o desenvolvimento do trabalho, apresentando a base de dados utilizada e a metodologia para os experimentos.

O capítulo 4 contém o detalhamento do experimento conduzido para obter a explicabilidade, bem como os resultados obtidos e as discussões referentes.

Por fim, no capítulo 5 são apresentadas as conclusões e as possibilidades de trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo serão apresentados os conceitos fundamentais e embasamento teórico para a compreensão do projeto.

2.1 Doença de Parkinson

Parkinson é uma doença neurológica crônica e progressiva que afeta os movimentos dos indivíduos, causando tremores, lentidão de movimentos, rigidez muscular, desequilíbrio, alterações na fala e na escrita. (BRASIL, 2022a)

A Doença de Parkinson ocorre por causa da degeneração das células situadas na região do cérebro chamada substância negra. As células dessa região produzem dopamina, que é um neurotransmissor, ou seja, conduz as correntes entre as células nervosas do corpo. A falta ou diminuição da dopamina afeta o controle motor, resultando nos sintomas da doença. (EINSTEIN, 2022)

Sabendo que a causa do Parkinson é a morte de células do cérebro, o início e progressão da doença muito comumente estão relacionados à idade do indivíduo, tendo em vista que, conforme envelhecem, todos os indivíduos saudáveis apresentam morte progressiva das células nervosas que produzem dopamina, a doença raramente ocorre em crianças ou adolescentes. (GONZALEZ-USIGLI, 2022)

Existem ainda fatores que podem influenciar na degeneração acelerada das células, diminuindo muito mais rapidamente os níveis de dopamina. Como resultado disso, existem indivíduos que podem vir a manifestar os sintomas da doença antes de alcançar idade avançada. Ainda não existe confirmação exata dos motivos para degeneração das células dessa região do cérebro, mas acredita-se que existam diversos fatores envolvidos neste processo, podendo ser tanto de natureza genética quanto ambiental. (EINSTEIN, 2022)

Estima-se que existam aproximadamente 4 milhões de pessoas no mundo com a Doença de Parkinson, representando 1% da população mundial a partir dos 65 anos. (BRASIL, 2022b) Estatísticamente, a doença de Parkinson apresenta os primeiros sintomas entre os 40 e 80 anos, sendo que a partir dos 40 anos ela afeta cerca de 1 pessoa a cada 250, progredindo para 1 a cada 100 a partir dos 65 anos e chegando a afetar 1 entre 10 pessoas acima dos 80 anos de idade. (GONZALEZ-USIGLI, 2022)

O seu diagnóstico é feito com base na história clínica do paciente e no exame neurológico. Entretanto, devido aos tremores que comumente afetam os dedos ou as mãos dos pacientes, os sintomas podem ser utilizados para identificar os indivíduos que possuem Parkinson.

A doença não tem cura, entretanto é tratável com medicações que devem ser usadas por toda a vida. Esses medicamentos repõem parcialmente a dopamina que falta para o indivíduo, melhorando os sintomas da doença.

O tratamento com medicamentos deve ser acompanhado por uma equipe composta por terapeutas ocupacionais, fisioterapeutas e outros profíssionais da saúde que atuem nas áreas da vida afetadas pela doença, buscando restaurar ou manter as funções do corpo e incentivar a realização das atividades de vida diária de forma independente. (BRASIL, 2022b)

2.2 Aprendizado de Máquina e Aprendizado Profundo

Aprendizado de Máquina é uma ramificação da área de inteligência artificial, que busca utilizar conjuntos de dados e algoritmos para tentar imitar a maneira dos seres humanos de aprender, de forma que os algoritmos possam melhorar sua performance gradualmente. Pode ser definido ainda como "o campo de estudo que dá aos computadores a habilidade de aprender sem serem programados explicitamente."(BROWN, 2021) Esses algoritmos são, geralmente, treinados para realizar predições ou classificações por meio de métodos estatísticos e dados de entrada.

Os modelos de ML podem ser classificados entre três categorias principais: aprendizado de máquina supervisionado, aprendizado de máquina não supervisionado e aprendizado semi-supervisionado. Em adição, há ainda uma outra categoria relevante, que é o aprendizado por reforço. Alguns exemplos de modelos de aprendizado de máquina são redes neurais, regressão linear e logística, agrupamento (*clustering*), árvores de decisão e florestas aleatórias. (IBM, 2022d)

Aprendizado Profundo, ou *Deep Learning*, é uma subárea de aprendizado de máquina, assim como este também pertence ao conjunto maior de inteligência artificial. Redes neurais que contém três ou mais camadas podem ser consideradas profundas. Estas Redes Neurais Profundas (DNN) tentam simular o comportamento do cérebro humano e possibilitam a criação de modelos que são capazes de aprender a partir de grandes quantidades de dados. Comumente apresentam maior acurácia que os modelos não profundos e são mais otimizados para problemas de alta complexidade.

Além disso, modelos de aprendizado profundo têm maior capacidade de utilizar dados não estruturados e sem rótulos, podendo receber imagens ou textos como entrada sem grandes necessidades de pré-processamento ou interferência de humanos. Isso se deve a um processo de extração de atributos feito de forma automatizada. (IBM, 2022c)

Alguns exemplos de modelos de aprendizado profundo que podem ser pontuados são Redes Neurais Convolucionais (CNNs), Redes Neurais Recorrentes (RNNs), Redes Adversárias Generativas (GANs) e *Autoencoders*.

2.2.1 Rede Neural Convolucional (CNN)

Redes Neurais Convolucionais, ou, *Convolutional Neural Networks* são um tipo de rede neural artificial, portanto são formadas por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Entretanto, CNNs possuem performance superior a redes neurais artificiais (ANNs) ao trabalhar com dados de entrada como imagens, áudio ou fala.

Redes convolucionais são formadas principalmente por três tipos de camadas:

- Camada de convolução
- Camada de agrupamento, ou pooling
- Camada totalmente conectada (FC, do inglês *Fully Connected Layer*)

A primeira camada de uma CNN é a camada de convolução, geralmente seguidas por camadas de agrupamento e outras convoluções e ao final, a camada totalmente conectada.

As camadas de convolução são as principais em CNNs, nesta camada é realizada, por exemplo, a detecção de padrões na imagem através da aplicação de filtros, geralmente de dimensão 3x3, que podem ser treinados para detectar os atributos desejados. Esse filtro deve varrer a imagem em todas as camadas de cor (3 camadas para RGB, por exemplo) e identificar as características desejadas. Nas primeiras camadas serão detectados apenas bordas de imagem ou cores e formas simples, enquanto nas camadas mais avançadas já será possível identificar formas mais complexas e de diferentes formatos e, ao final, espera-se que detecte objetos ou partes do corpo em grandes porções da imagem.

Depois que o filtro passa por toda a imagem, calculando o produto escalar dos *pixels* em sua área, é obtido um produto final chamado de mapa de características. Para finalizar uma etapa de convolução é aplicada a função de ativação ReLU (*Rectified Linear Unit*) no mapa de características.

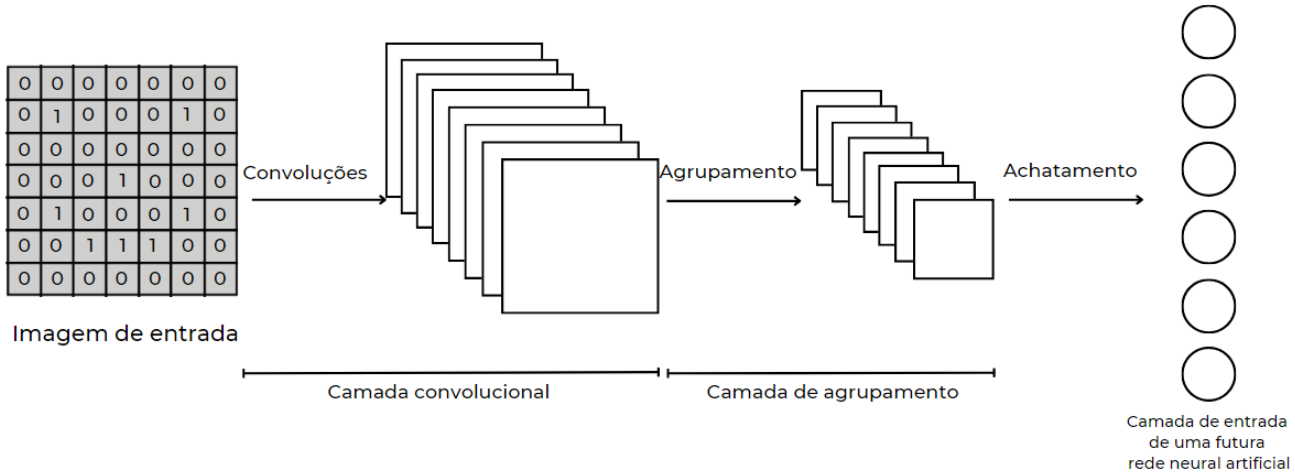
As camadas de agrupamento reduzem o número de parâmetros de entrada, reduzindo o tamanho de uma imagem, por exemplo, com o objetivo de diminuir a complexidade e melhorar a eficiência. Isso ocorre através de um filtro que varre a imagem enquanto aplica uma função de agregação que pode ser de *max pooling*, onde esse filtro irá selecionar os pixels, ou atributos, com maior valor dentro de uma área próxima, geralmente é uma camada em duas dimensões de 2x2. Alternativamente pode ser utilizado também *average pooling*, onde ao invés de selecionar o maior valor da área, é utilizado um valor médio.

Após o agrupamento é comum aplicar achatamento, ou *flattening*, na matriz existente. Isto é feito apenas para transformar as matrizes resultantes em um vetor linear mais longo.

A camada totalmente conectada serve basicamente como passagem entre os valores das camadas intermediárias para a camada de saída. A camada FC utiliza uma função de ativação

sigmoide para transformar os *inputs* em uma saída no formato de probabilidade que vai de 0 a 1. Uma visão geral do processo das CNNs pode ser observada na Figura 1.

Figura 1 – Representação geral do processo de uma rede convolucional



Fonte: Adaptado de Ali (2019)

Conforme os dados passam pelas camadas, a CNN aumenta a sua complexidade. Nas primeiras camadas de convolução apenas pequenos atributos são detectados, como bordas e cores e gradualmente são alcançados maiores níveis de abstração, chegando ao ponto onde, em uma imagem por exemplo, já poderiam ser identificadas grandes seções de uma vez, como janelas, objetos e membros do corpo. (IBM, 2022a)

2.3 Inteligência Artificial Explicável

Inteligência Artificial Explicável se refere ao conjunto de ferramentas e técnicas utilizadas para auxiliar no processo de compreensão dos resultados obtidos através de uma Inteligência Artificial.

XAI contrasta diretamente com o conceito de modelos caixa-preta, buscando explicar e tornar mais transparente o resultado do modelo. Isto ocorre principalmente através da pontuação dos fatores que mais influenciam para chegar até certa conclusão. Quando fala-se desses fatores, costuma-se referir às variáveis de entrada do sistema, ou seja, as características, ou atributos do modelo. (ANKARSTAD, 2020)

Modelos caixa-preta, do inglês, *black-box*, seriam todos os modelos que podem ser considerados complexos no seu funcionamento (CHU, 2020), se referindo principalmente a modelos que dificilmente seria possível até mesmo para o engenheiro ou cientista de dados que criou o algoritmo compreender ou explicar todos os passos tomados pela IA até chegar em determinado resultado, isso porque são modelos criados diretamente a partir dos dados. (IBM, 2022b). Observa-se uma representação visual de um modelo caixa-preta na Figura 2.

Figura 2 – Modelo black-box

Fonte: Elaborada pelo autor

Exemplos de modelos *black-box* são redes neurais artificiais, redes neurais profundas, algoritmos de boosting, máquinas de vetor de suporte e florestas aleatórias. (RIBEIRO SAMEER SINGH, 2016).

> "Embora os primeiros sistemas de IA fossem facilmente interpretáveis, os últimos anos testemunharam o surgimento de sistemas de decisão menos transparentes, como DNNs. À medida que IA se torna mais avançada, o processo de compreensão de como os algoritmos chegaram a determinado resultado passa a ser mais desafiador. O sucesso empírico dos modelos de DL, como as DNNs, decorre de uma combinação de algoritmos de aprendizado eficientes e seu enorme espaço paramétrico. Este último espaço compreende centenas de camadas e milhões de parâmetros, o que faz com que as DNNs sejam consideradas modelos complexos de caixa-preta. O oposto da caixa-preta é a transparência, ou seja, a busca por uma compreensão direta do mecanismo pelo qual um modelo funciona."[1] (ARRIETA, 2019)"

Quase todas as empresas utilizam ou têm planos de incorporar técnicas ou produtos IA a seus negócios. Bem como IA está inserida de diversas formas no cotidiano através de aplicativos, aparelhos eletrônicos, algoritmos de busca, até mesmo na medicina de precisão, onde os especialistas exigem muito mais informações do modelo do que uma simples previsão binária para apoiar seu diagnóstico. Outros exemplos incluem veículos autônomos em transporte, segurança e finanças, entre outros. (ARRIETA, 2019) Devido a isso, se faz ainda mais necessário compreender como esses sistemas tomam decisões, portanto XAI é uma área de grande relevância.

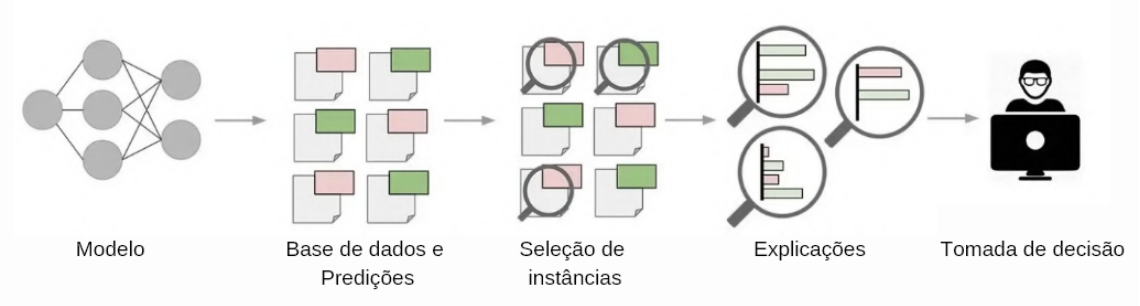
[1] Tradução realizada pelo autor. Texto original: "While the very first AI systems were easily interpretable, the last years have witnessed the rise of opaque decision systems such as Deep Neural Networks (DNNs). The empirical success of Deep Learning (DL) models such as DNNs stems from a combination of efficient learning algorithms and their huge parametric space. The latter space comprises hundreds of layers and millions of parameters, which makes DNNs be considered as complex black-box models. The opposite of black-boxness is transparency, i.e., the search for a direct understanding of the mechanism by which a model works.".

Dentre as vantagens de entender como modelos de inteligência artificial chegaram a seus resultados, pode-se citar que o entendimento pode auxiliar a garantir o pleno funcionamento desses sistemas, bem como ajudar as pessoas interessadas a compreender e confiar nos resultados, possibilitando até mesmo que esses resultados sejam confrontados ou sejam feitas mudanças nos modelos para alterá-los quando necessário. (ROYALSOCIETY.ORG, 2019)

É importante lembrar afinal que esses modelos são utilizados por seres humanos que devem ser capazes de confiar neles, e isso só pode se dar através da compreensão de suas predições e entendendo a motivação dos erros obtidos. (THORN, 2020)

Observa-se na Figura 3 o processo de análise dos resultados de um modelo utilizando explicações.

Figura 3 – Processo partindo do modelo até as decisões tomadas por humanos, que se baseiam nas explicações obtidas

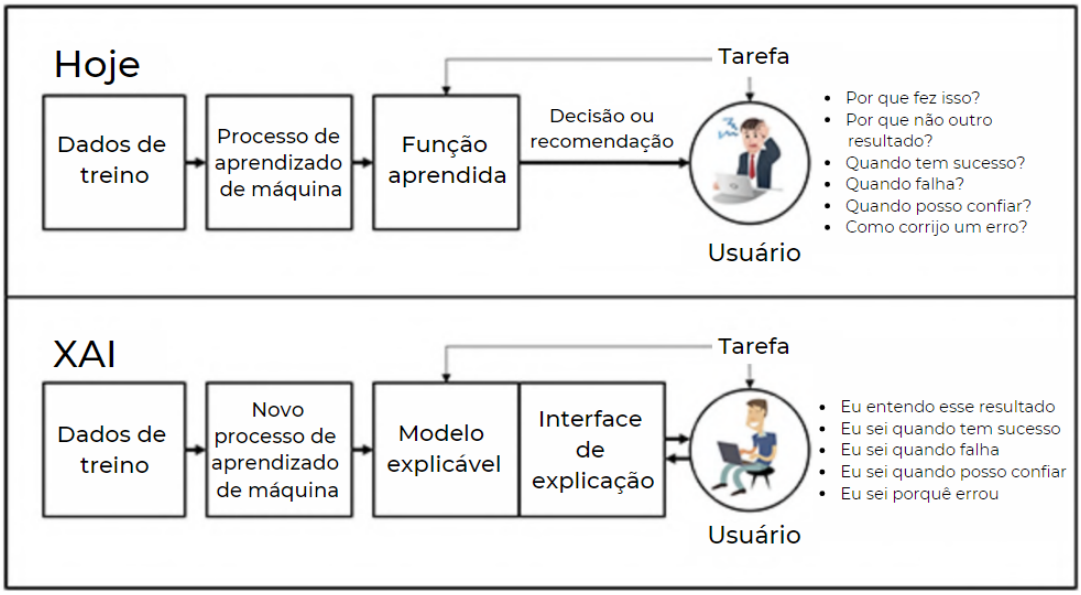


Fonte: Adaptado de Hulstaert (2018)

Redes neurais utilizadas em aprendizado profundo, como exemplo de caixa-preta, são extremamente difíceis de serem compreendidas por humanos. Existem fatores que nem sempre são explícitos, como vieses inesperados que podem se basear em raça, idade, gênero ou outras características e que podem causar diversos problemas de grandes magnitudes. Além disso, problemas de qualidade não encontrados no desenvolvimento e teste do modelo podem ser encontrados através da análise da explicação dos modelos. (IBM, 2022b)

Sobre o que pode ser o futuro de XAI, de acordo com Turek (2022), XAI futuramente deverá resultar na alteração do processo de desenvolvimento de modelos, de forma que a capacidade de explicação e compreensão estaria integrada dentro dos próprios modelos, ou seja, os sistemas de IA trariam consigo explicações intrínsecas. Isso pode ser observado na Figura 4.

Figura 4 – Implementação das camadas de explicação junto aos modelos, possibilitando o entendimento dos resultados



Fonte: Adaptado de Turek (2022)

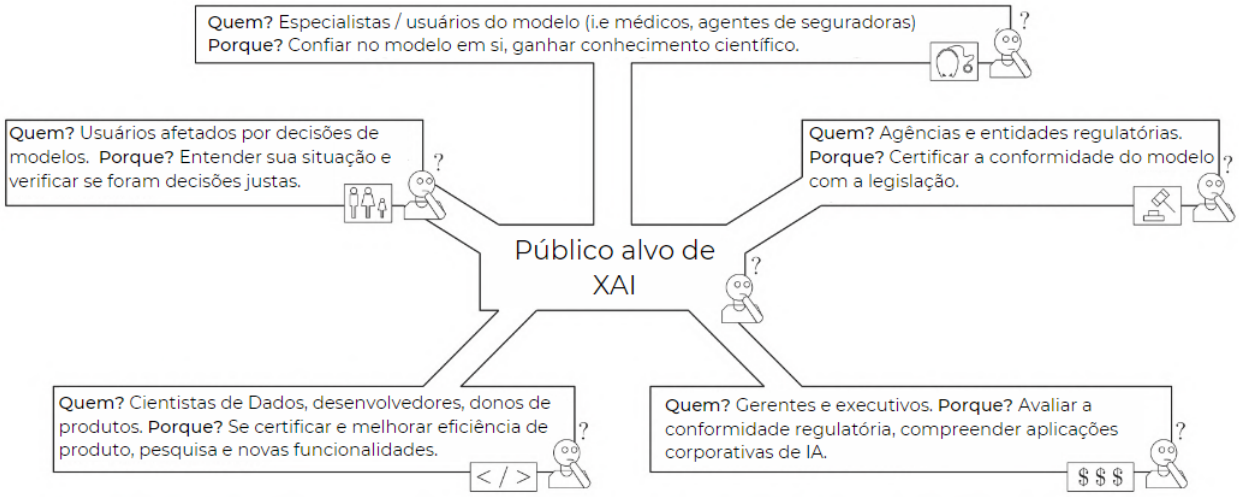
> "Diferentes métodos de IA são afetados por preocupações sobre a explicabilidade de diferentes maneiras, e diferentes métodos ou ferramentas podem fornecer diferentes tipos de explicação. Existem exemplos de sistemas de IA que não são facilmente explicáveis, mas podem ser implantados sem preocupação; também há casos em que o uso de métodos de IA explicáveis é necessário e também precisa ser apoiado por sistemas mais amplos para garantir a responsabilidade em todo o pipeline analítico – desde a coleta de dados até a decisão. Aqueles que desenvolvem e implantam IA precisam levar em conta as necessidades dos diferentes grupos que interagem com o sistema, considerando quais tipos de explicação podem ser úteis e para qual finalidade."[2] (ROYALSOCIETY.ORG, 2019)

A Figura 5 apresenta quais poderiam ser as partes interessadas em explicações geradas por XAI e a razão do interesse.

Existe um fenômeno que é comumente aceito na comunidade de ciência de dados, que é a afirmação de que existe uma troca entre acurácia do modelo e interpretabilidade. Ao trabalhar com grandes bases de dados e utilizar modelos complexos, como os caixa-preta, geralmente é possível alcançar boa performance, entretanto baixa interpretabilidade.

[2] Tradução realizada pelo autor. Texto original: "Different AI methods are affected by concerns about explainability in different ways, and different methods or tools can provide different types of explanation. There are examples of AI systems that are not easily explainable, but can be deployed without concern; there are also cases where the use of explainable AI methods is necessary and also needs to be supported by wider systems to ensure accountability across the full analytics pipeline – from data collection to decision. Those developing and deploying AI need to take into account the needs of different groups interacting with the system, considering what types of explanation might be useful and for what purpose.".

Figura 5 – Demonstração das diferentes partes interessadas na explicação de modelos



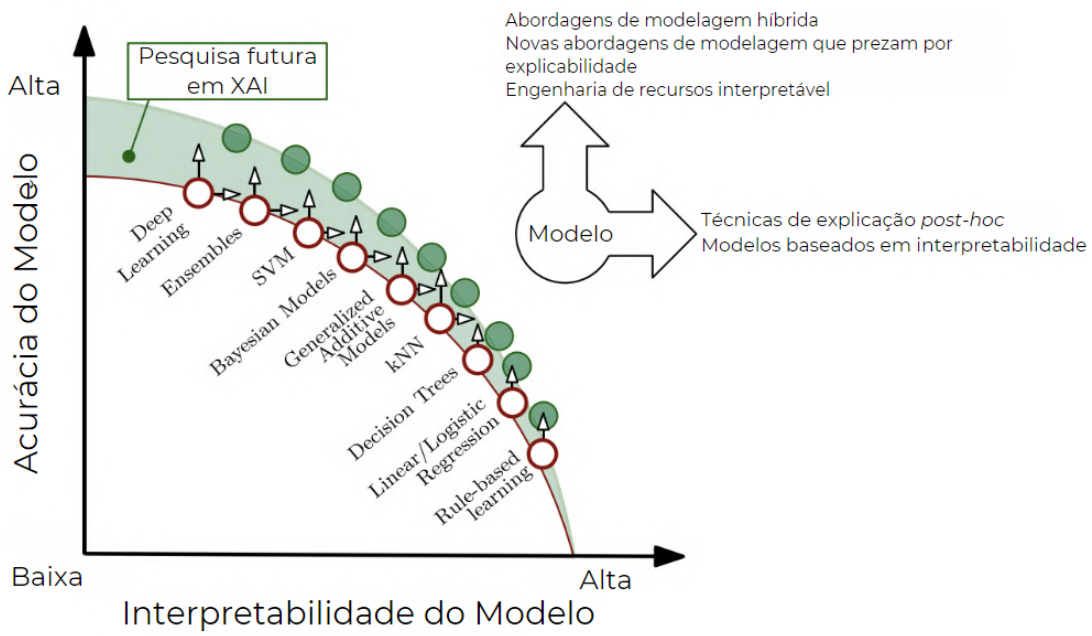
Fonte: Adaptado de Arrieta (2019)

Pode-se dizer também o contrário, em modelos simples, como modelos lineares, costuma ser possível interpretar a intuitivamente compreender os cálculos realizados pelo modelo, mas geralmente apresentando resultados piores.

Neste contexto, técnicas de explicação *post-hoc* surgem, possibilitando a interpretação de modelos complexos e garantindo tanto bons resultados quanto interpretabilidade de seu funcionamento. (CHU, 2020)

É possível observar na Figura 6 a relação entre acurácia e interpretabilidade de modelos citada acima.

Figura 6 – Interpretabilidade versus acurácia em diferentes modelos



Fonte: Adaptado de Arrieta (2019)

2.3.1 Ferramentas de XAI

Como apresentado por Neto (2021), existem diversas abordagens para prover explicações dos modelos de ML. Eles diferem normalmente em função do tipo da técnica e objetivos. A seguir são apresentadas algumas das principais categorias de classificação destas.

- Global se refere uma explicação geral sobre o modelo, buscando explicar, por exemplo, o comportamento geral, os atributos de maior influência e a lógica da IA.
- Local é a explicação de instâncias individualmente, por exemplo, demonstrar os fatores que foram mais impactantes para chegar a uma determinada predição.

É possível ainda separar as explicações em "específico a modelo"e "agnóstico a modelo".

- Específico a modelo, ou *model-specific*, são aquelas que são capazes de explicar apenas o tipo de modelo para o qual foram especificamente criadas.
- Agnóstico a modelo, ou *model-agnostic*, se refere a técnicas de XAI que conseguem gerar explicações para qualquer tipo de modelo em que sejam aplicadas

Por fim, existem outras duas classificações que se dão através do momento que a explicação é gerada.

- *Transparent*, *Intrinsic*, ou *White-box* fornecem interpretações em suas próprias estruturas, ou seja, no próprio modelo. Exemplos são árvores de decisão e modelos lineares.
- *Post-hoc* se refere a técnicas que são aplicadas no modelo, após seu treinamento, para criar explicações.

Atualmente existem muitas ferramentas para trabalhar com XAI, sendo bibliotecas ou *frameworks* disponíveis na linguagem Python, cada uma com suas particularidades e forma de funcionamento. Pode-se pontuar dentre algumas das mais populares: SHAP, LIME, Shapash, ExplainerDashboard, Dalex, Explainable Boosting Machines (EBM) e ELI5.

Tendo em vista a relevância e popularidade do assunto de XAI, existem ainda muitas outras ferramentas para Python que podem ser citadas, como Alibi, Skater, EthicalML, Aix360 da IBM, DiCE, ExplainX.Ai, entre outras. (BHATNAGAR, 2021)

Neste trabalho são utilizados as duas ferramentas mais populares para Explicabilidade de IA: SHAP e LIME, duas ferramentas agnósticas a modelo que geram explicações locais, *post-hoc* (CHU, 2022) Desta forma, somente estas duas terão suas subseções na fundamentação teórica.

2.3.2 LIME

> "Lime[3] é a abreviação para Local Interpretable Model-Agnostic Explanations. Cada parte do nome reflete em algo que desejamos em explicações. Local se refere à fidelidade local - i.e., queremos que a explicação realmente reflita o comportamento do classificador "em volta"da instância sendo prevista. Essa explicação é inútil a não ser que seja interpretável - ou seja, a não ser que um ser humano consiga compreendê-la. Lime é capaz de explicar qualquer modelo sem precisar "espiar"dentro dele, então é agnóstica a modelo."[4](RIBEIRO, 2016)

As explicações do LIME são geradas através da aproximação de um modelo caixa-preta localmente utilizando um modelo mais simples, que pode ser obtido causando perturbações na instância a ser explicada. Desta maneira, LIME tenta obter os atributos que influenciam a predição do modelo, através da atribuição de pesos baseados na similaridade entre instâncias que sofreram perturbações e a instância de interesse, em suma, são realizadas perturbações na imagem para ver como a saída do modelo se comporta, aprendendo assim quais fatores têm maior importância. Como consequência podem ser geradas explicações que não representam o comportamento do modelo para todas as instâncias, somente aquela sendo explicada.

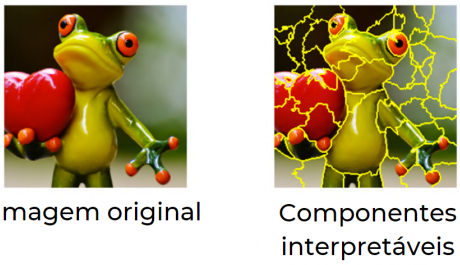
[3] (RIBEIRO; SINGH; GUESTRIN, 2016)

[4] Tradução realizada pelo autor. Texto original: "Lime is short for Local Interpretable Model-Agnostic Explanations. Each part of the name reflects something that we desire in explanations. Local refers to local fidelity - i.e., we want the explanation to really reflect the behaviour of the classifier "around"the instance being predicted. This explanation is useless unless it is interpretable - that is, unless a human can make sense of it. Lime is able to explain any model without needing to 'peak' into it, so it is model-agnostic.

Entretanto, o funcionamento descrito é um ponto positivo em termos de interpretabilidade das explicações obtidas, já que as perturbações são realizadas em componentes que podem ser compreendidos por humanos, como pedaços de imagem ou palavras em uma frase, configurando uma abordagem intuitiva.

Para exemplificar como LIME funciona em modelos de classificação de imagens, primeiramente a imagem é dividida em diversos componentes interpretáveis, ilustrado na Figura 7.

Figura 7 – Componentes interpretáveis do LIME



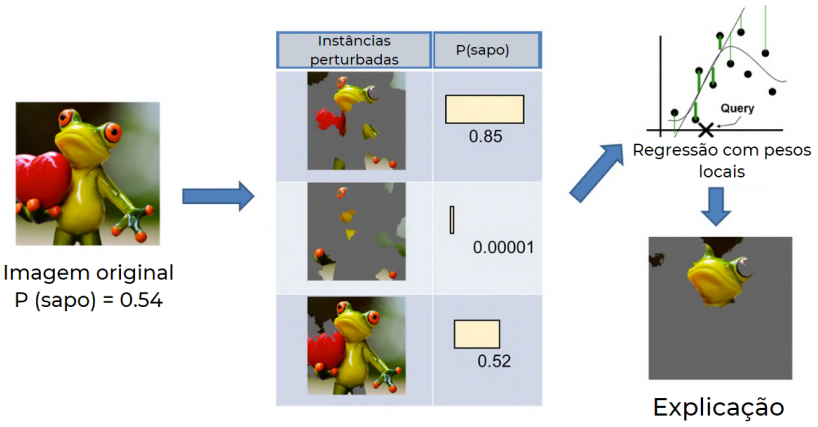
Fonte: Adaptado de Ribeiro Sameer Singh (2016)

Em seguida, é gerado um conjunto de imagens obtidas através de perturbações na imagem original. Para cada instância de perturbação, é obtida a predição do modelo indicando a probabilidade da classe desejada estar na imagem.

A partir daí, é obtido um modelo simples, geralmente linear, que represente o comportamento do modelo caixa-preta desejado a um nível local e que possa retornar os componentes da imagem de maior peso, ou seja, com maior importância local. (RIBEIRO SAMEER SINGH, 2016)

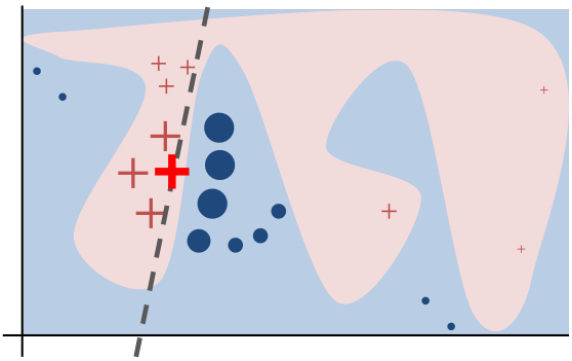
A etapa de obtenção de um modelo linear pode ser representada na Figura 8

Figura 8 – Perturbação de instâncias e obtenção do modelo linear



Fonte: Adaptado de Ribeiro Sameer Singh (2016)

Figura 9 – Aproximação do modelo



Fonte: Ribeiro (2016)

A Figura 9 contém uma aproximação para visualizar o modelo, onde o fundo rosa e azul representa as fronteiras de decisão do modelo caixa-preta e o símbolo de cruz vermelho é a instância sendo explicada. Através das perturbações, as outras instâncias são geradas. Ao final disso, o modelo linear utilizado para explicar a instância é obtido. (RIBEIRO, 2016)

Por fim espera-se que LIME seja capaz de trazer boas explicações locais, mas tendo em vista que o desempenho de LIME se baseia na quantidade de instâncias geradas ao redor da instância a ser explicada, portanto pode ser necessário atenção ao gerar explicações, já que, assim como todo sistema, é passível de falhas. (CHU, 2020)

2.3.3 SHAP

SHAP[5] é a abreviação de Shapley Additive exPlanations. SHAP é um *framework* com ferramentas computacionalmente eficientes para calcular os valores de Shapley, um conceito dentro de teoria dos jogos cooperativos, introduzido por Lloyd Shapley.

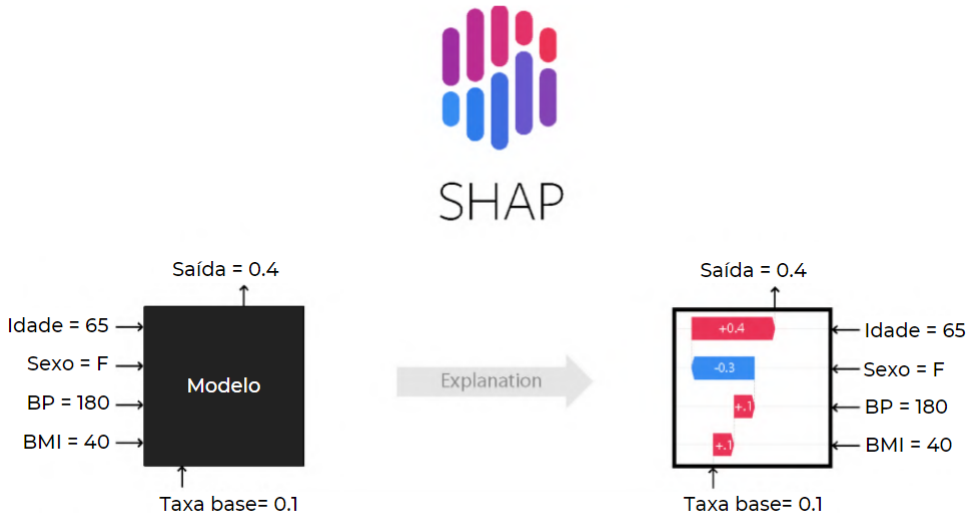
A Figura 10 apresenta a ideia básica de funcionamento de SHAP.

É importante pontuar em um primeiro momento que, como um *framework*, SHAP contém uma gama de possíveis aplicações, com diferentes módulos focados em várias abordagens de modelos. Dentre eles estão Deep SHAP, Linear SHAP e Kernel SHAP, por exemplo.

Valores de Shapley, a base do SHAP é um método para calcular a contribuição de cada membro dentro de uma coalizão para determinado resultado. A intuição mais simples para este conceito é a de um grupo de pessoas que trabalham em conjunto para realizar uma tarefa. Considerando que esta tarefa tenha uma recompensa diretamente proporcional ao esforço, busca-se responder como seria possível calcular com precisão a parte merecida de cada uma das pessoas do grupo, levando em consideração o quanto cada uma contribuiu para alcançar o objetivo. De forma a exemplificar, a tarefa pode ser considerada um jogo e a recompensa corresponde à pontuação.

[5] (LUNDBERG; LEE, 2017)

Figura 10 – SHAP



Fonte: Adaptado de SHAP (2022)

Para o cálculo, é necessário descobrir qual seria o resultado do jogo para todas as combinações possíveis de jogadores. O que é chamado de coalizão é uma única combinação dentre os jogadores. Considerando uma situação de quatro jogadores, existiriam $2^4 = 16$ combinações possíveis, sabendo que cada jogador poderia estar ou não naquela coalizão.

Ao obter os resultados para cada configuração possível de jogadores, calcula-se o Shapley value para cada jogador individualmente. O cálculo se dá através de uma soma ponderada da diferença entre a pontuação dos jogos nos quais o jogador estava presente e os que ele não estava. Ao resultado dessa soma ponderada é aplicado ainda um peso respectivo e por fim, chega-se ao resultado da contribuição daquele jogador.

Voltando ao contexto de ML, pode-se considerar os jogadores como atributos de um modelo e a recompensa de cada um como seu impacto na predição, ou seja, valores de Shapley podem ser utilizados para descobrir como a predição se distribui entre cada um dos atributos.

Analisando a situação descrita, pode parecer simples realizar o cálculo, entretanto o problema surge ao aplicar o mesmo processo em modelos de aprendizado de máquina, em especial, de aprendizado profundo, que possuam quantidade maiores de atributos. Sabendo que a quantidade de combinações possíveis é de 2^N, onde N é a quantidade de atributos, pode-se alcançar valores muito altos a depender do modelo e dos dados. Por isso, a ferramenta SHAP tem esse processo implementado eficientemente, podendo ser aplicado a todo tipo de modelos. (CHU, 2022)

3 Metodologia

Neste capítulo é abordada a metodologia utilizada para a obtenção dos resultados no experimento de XAI, apresentando a base de dados e as ferramentas utilizadas para o desenvolvimento. Em um primeiro momento foi realizada a pesquisa teórica para compreender sobre o tema e encontrar os principais métodos e ferramentas para trabalhar com explicabilidade de modelos.

Isto estabelecido, foi implementado o exemplo prático, iniciado pela busca da base de dados a ser utilizada, implementação do modelo e por fim, aplicação das ferramentas de explicabilidade.

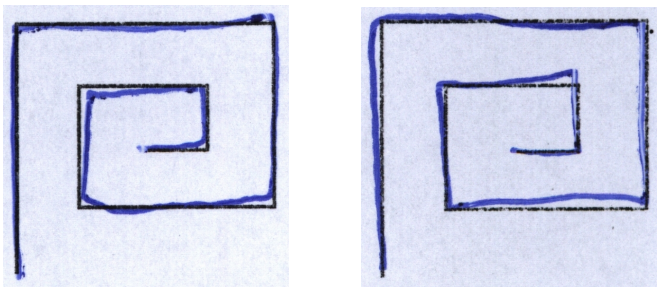
3.1 Base de Dados

A base de dados utilizada *HandPD*[1] *Dataset* é populada com exames de escrita de dois grupos de indivíduos: Grupo saudável e o grupo de pacientes, contendo indivíduos que sofrem com Parkinson.

Os exames de escrita foram coletados na Faculdade de Medicina de Botucatu, estado de São Paulo, Brasil. Para os indivíduos, o exame consistia em seguir com caneta os formatos de quatro espirais e quatro meandros, que foram posteriormente recortados e salvos no formato de imagem "jpg".

Ao todo, a base de dados contém os exames de 92 indivíduos, sendo 18 do grupo saudável e 74 do grupo de pacientes, somando o total de 736 imagens, divididas entre os dois grupos. As Figuras 11 a 14 apresentam algumas imagens que estão presentes na base de dados.

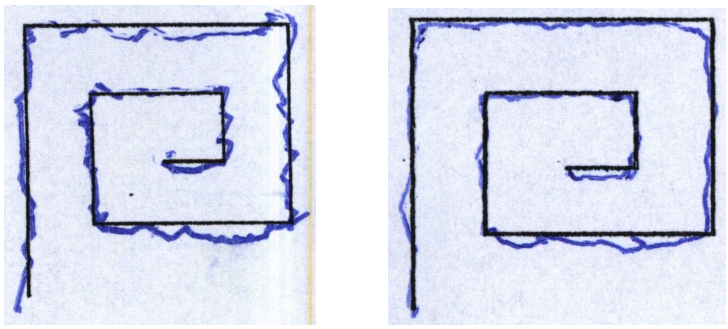
Figura 11 – Exames com meandro do grupo de pessoas saudáveis



Fonte: Pereira et al. (2016).

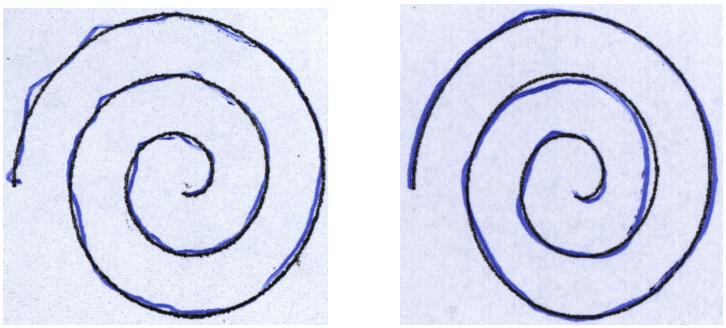
[1] (PEREIRA et al., 2016)

Figura 12 – Exames com meandro do grupo de pessoas com Parkinson



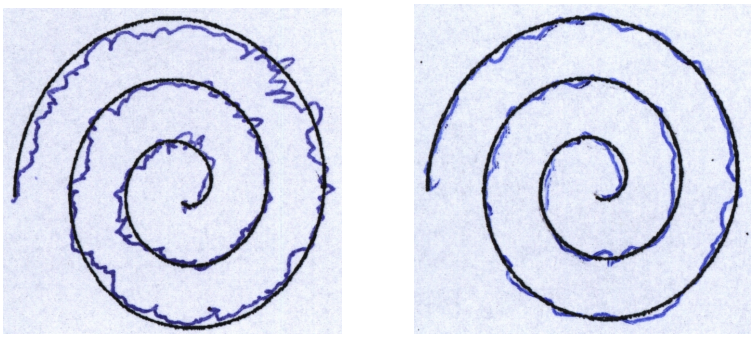
Fonte: Pereira et al. (2016).

Figura 13 – Exames com espiral do grupo de pessoas saudáveis



Fonte: Pereira et al. (2016).

Figura 14 – Exames com espiral do grupo de pessoas com Parkinson



Fonte: Pereira et al. (2016).

3.2 Ferramentas

Para o desenvolvimento do projeto foram utilizadas as seguintes ferramentas:

- Google Colab, ou Colaboratory é um produto gratuito do Google Research, área de pesquisas científicas do Google. O Colab permite que qualquer pessoa escreva e execute código Python pelo navegador e é especialmente adequado para aprendizado de máquina, análise de dados e educação. Mais tecnicamente, o Colab é um serviço de *notebooks* hospedados do Jupyter que não requer nenhuma configuração para usar e oferece acesso sem custo financeiro a recursos de computação como GPUs.
- Python é uma linguagem de programação orientada a objetos e de alto nível amplamente usada em aplicações da *Web*, desenvolvimento de *software*, ciência de dados e aprendizado de máquina.
- SHAP é um *framework* utilizado em Python que contém métodos matemáticos para explicar localmente as predições de modelos de aprendizado de máquina. É baseado em conceitos de teoria dos jogos introduzidos por Lloyd Shapley em 1951, que servem como base para calcular a contribuição de cada atributo para determinada predição de modelo.
- LIME é um *framework* para Python utilizado para obter explicações de modelos localmente, baseado em realizar perturbações na instância a ser explicada e desta forma calcular a contribuição de cada atributo ou módulo da instância para o resultado do modelo.
- Keras é uma API (*Application Programming Interface*) efetiva e de alto nível utilizada para trabalhar com redes neurais, escrita em Python.
- NumPy é uma biblioteca de Python que providencia objetos de vetores multidimensionais, diversos objetos derivados, rotinas para operações em vetores, incluindo manipulação matemática e lógica de formas, transformações discretas de Fourier, álgebra linear básica, operações estatísticas básicas e muito mais.
- Matplotlib é uma biblioteca compreensível para criar visualizações estáticas, animadas e interativas em Python.
- Skimage, ou Scikit-image é uma biblioteca *open-source* para processamento de imagens.

4 Desenvolvimento

Neste capítulo será abordado o processo de desenvolvimento do projeto, do experimento e mais ao fim, a discussão sobre os resultados obtidos.

Para o desenvolvimento do projeto, primeiramente foi realizado o levantamento bibliográfico e estudo do tema de explicabilidade de inteligência artificial, ao mesmo tempo, foi realizada a busca pelas ferramentas de XAI que seriam utilizadas no projeto. Desta forma, foram escolhidos LIME e SHAP, já que são as duas ferramentas mais populares no tema devido a suas características e capacidades apresentadas na fundamentação teórica.

Ao final do trabalho com a bibliografia, foi realizado outro levantamento, neste segundo momento, o modelo CNN a ser utilizado, bem como a base de dados HandPD para o treinamento do modelo e para os testes foram escolhidos.

Por fim, a última etapa do projeto foi o desenvolvimento, que se deu pela implementação do modelo, seguida pela aplicação das técnicas de XAI e avaliação dos resultados obtidos.

4.1 Experimento

O experimento conduzido foi a aplicação de técnicas de inteligência artificial explicável sobre o problema de classificação de imagens de exames médicos de pacientes com Parkinson. A intenção é a de compreender quais fatores têm maior influência na decisão do modelo para dada instância.

Para o desenvolvimento, se faz necessário um conjunto de imagens sobre as quais será aplicada a explicação, estas que podem ser chamadas também de imagens de exemplo, foram escolhidas de maneira arbitrária de dentro do conjunto de testes do modelo. O critério utilizado foi a escolha de imagens que supostamente teriam maiores chances de gerar boas análises, ou seja, ao mesmo tempo que foram escolhidas imagens que poderiam ser consideradas fáceis de classificar e explicar tal classificação, tanto por humanos quanto pela máquina, também existem imagens que geram dúvidas quanto à sua classificação correta para humanos e predições incorretas para o modelo.

Visando obter resultados uniformes sobre suas capacidades, ambas ferramentas geraram explicações para cada uma das instâncias do conjunto de imagens de exemplo.

4.1.1 Modelagem

O problema foi abordado como classificação binária, buscando predizer se o indivíduo pertencia ao grupo de indivíduos saudáveis (0) ou ao grupo de pacientes de Parkinson (1). Na implementação do modelo foi utilizada a biblioteca Keras para desenvolver uma Rede Neural Convolucional (CNN).

A Figura 15 apresenta o código da modelagem da CNN. Logo após inicializar o modelo (*classifier*), é aplicado o primeiro passo de convolução, com 32 filtros de dimensão 3x3 nos dados de entrada que têm o formato (512, 512, 3) e utilizando a função de ativação ReLu. Em seguida é aplicado o primeiro agrupamento do tipo *max pooling*, usando janelas de dimensão 2x2. Após isso, repete-se o primeiro passo de convolução e outro de agrupamento. Como próximos passos, aplica-se o achatamento e é realizada a conexão total das camadas utilizando a função de ativação *sigmoid*. Ao final da definição das camadas, a CNN é compilada com o otimizador *Adam* e está pronta para ser treinada com os dados.

Figura 15 – Modelagem da CNN.

```
#Inicializando a CNN
classifier = Sequential()
# Convolucao
classifier.add(Convolution2D(32,3,3, input_shape = (512,512,3),
    activation = 'relu'))

# Max pooling
classifier.add(MaxPooling2D(pool_size = (2,2)))
# Segunda convolucao e max pooling
classifier.add(Convolution2D(32,3,3, activation = 'relu'))
classifier.add(MaxPooling2D(pool_size = (2,2)))
# Flatten
classifier.add(Flatten())
# Full Connection hidden layer
classifier.add(Dense(128, activation = 'relu'))
# Output layer
classifier.add(Dense(1, activation = 'sigmoid'))
#Compilando a CNN
classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy',
    metrics = ['accuracy'])
```

Fonte: Elaborada pelo autor.

Após o modelo de classificação ter sido definido, foram feitos os processamentos necessários com a base de dados para obter os conjuntos de treino e teste. A dimensão utilizada para as imagens de entrada foi de 512 por 512 por 3 (RGB).

Com o processamento dos dados pronto, o treinamento do modelo foi realizado com 10 épocas e atingiu a acurácia de validação de 85,62%. Importante ressaltar que a escolha de apenas 10 épocas para treinamento da rede foi feita visando obter uma boa acurácia, mas com

certa margem de erros para que fosse possível analisar o comportamentos das ferramentas de XAI nessas instâncias. O treinamento é apresentado na Figura 16

Figura 16 – Treino do modelo.

```
# Fit
classifier.fit_generator(
    train_set,
    steps_per_epoch = 588,
    epochs = 10,
    validation_data = test_set,
    validation_steps = 146
)
```

Fonte: Elaborada pelo autor.

Ao final da modelagem, as imagens de exemplo já citadas acima e que serão utilizadas para obter a explicabilidade com as ferramentas de XAI, passam pelo mesmo processo de normalização que as imagens de entrada do modelo, portanto, também possuem as dimensões de (512, 512, 3).

4.1.2 LIME

Concluída a etapa de modelagem, a atenção se volta para a aplicação das ferramentas de explicabilidade, primeiramente utilizando LIME. Em um primeiro momento, é necessário definir o *explainer* do LIME, neste caso, foi utilizado o LimeImageExplainer(), tendo em vista que busca-se obter explicações de um modelo baseado em imagens. Apresentado na Figura 17

Figura 17 – LIME Image Explainer.

```
explainer = lime_image.LimeImageExplainer()
```

Fonte: Elaborada pelo autor.

Em razão do LIME gerar apenas explicações locais, ou seja, de ser capaz de explicar apenas uma instância por vez, é necessário gerar o objeto de explicação de cada uma das imagens de exemplo individualmente. Para isso, utiliza-se o método explain_instance() do *explainer* definido anteriormente. Os parâmetros para este método são: a imagem a ser usada no exemplo, a predição do modelo para esta instância, os principais *labels* atribuídos pelo modelo a esta instância e o número de amostras que serão utilizadas pelo LIME para gerar esta explicação. Para este exemplo, em todas as instâncias foi definido o valor de rótulos como 5 e a quantidade de amostras como 1000, apresentado na Figura 18.

Figura 18 – Explain instance.

```
exp = explainer.explain_instance(normalized_img_mc3[0].astype('double'
    ),
                                 classifier.predict,
                                 top_labels=10,
                                 hide_color=0,
                                 num_samples=1000)
```

Fonte: Elaborada pelo autor.

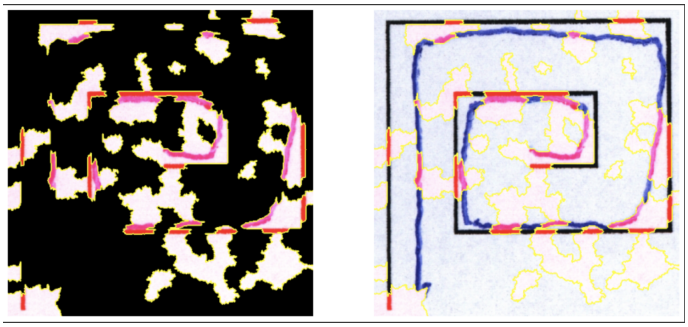
Ao final do processamento, foi obtido o objeto de explicação para a instância determinada. A partir daí é realizada a plotagem da explicação. Um parâmetro para a plotagem é a quantidade de atributos a serem incluídos para aquela explicação. Todas as instâncias seguem com o número de 50 atributos.

Repete-se o mesmo processo para todas as instâncias do conjunto de imagens de exemplo e desta forma, foram obtidas as suas explicações respectivas.

As legendas das imagens contém a classificação correta daquela instância (paciente ou grupo controle), seguido pelo resultado numérico apresentado pelo modelo para aquela instância. Valores que se aproximam de 0 indicam um indivíduo que pertenceria ao grupo de pessoas saudáveis e valores mais próximos de 1, ao grupo com Parkinson. Além disso, as explicações obtidas foram separadas e apresentadas em quatro grupos para serem interpretadas, como pode ser visto a seguir.

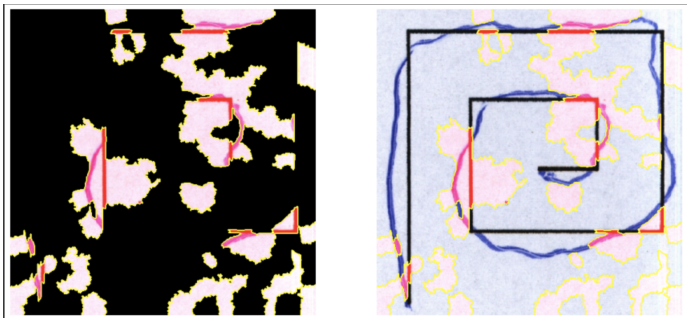
Grupo 1 (Figuras 19 a 21): Explicações de instâncias do grupo de pacientes, que indicam relevância em áreas da imagem correspondentes aos pontos da imagem com traços de caneta fora do delineado base ou irregulares. Sendo assim, podem ser entendidas como explicações condizentes com a interpretação humana.

Figura 19 – Paciente - Resultado modelo: 0.99951226.



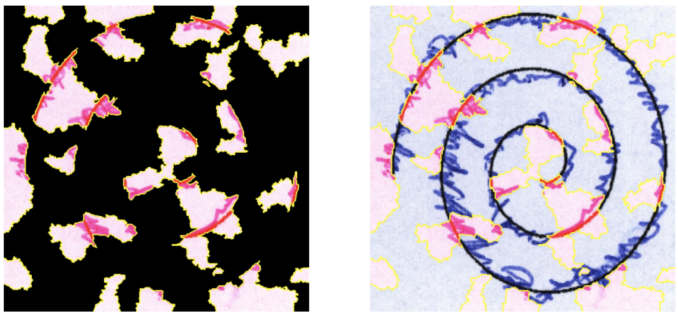
Fonte: Elaborada pelo autor.

Figura 20 – Paciente - Resultado modelo: 0.9995695.



Fonte: Elaborada pelo autor.

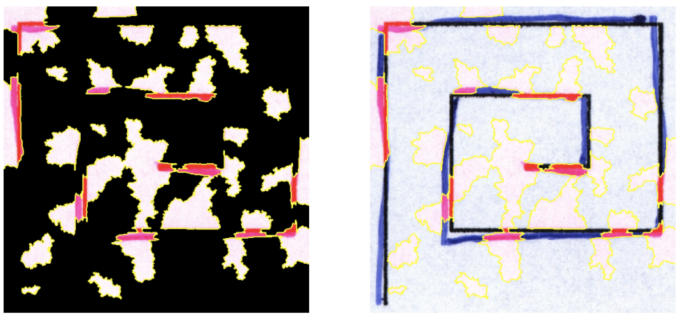
Figura 21 – Paciente - Resultado modelo: 0.9993524.



Fonte: Elaborada pelo autor.

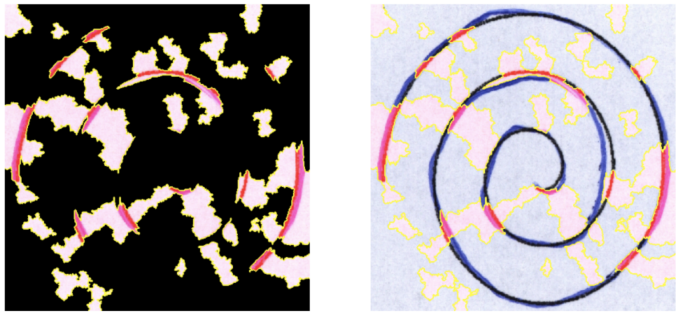
Grupo 2 (Figuras 22 e 23): Explicações de instâncias do grupo de indivíduos saudáveis, que indicam relevância em áreas com traços de caneta que acompanham ou se mantêm próximos ao delineado base. Novamente poderiam ser entendidas como explicações condizentes com a interpretação humana.

Figura 22 – Controle - Resultado modelo: 0.23667398.



Fonte: Elaborada pelo autor.

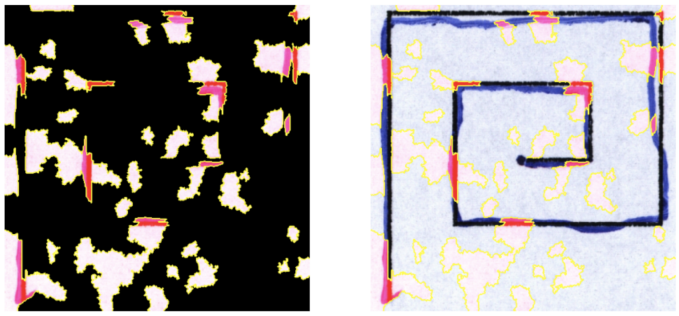
Figura 23 – Controle - Resultado modelo: 0.0436979.



Fonte: Elaborada pelo autor.

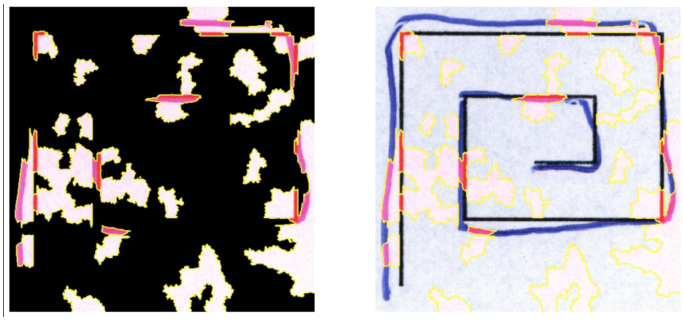
Grupo 3 (Figuras 24 e 25): Explicações para predições incorretas do modelo. Na Figura 24 pode ser visto que LIME apontou relevância em áreas próximas ao delineado base, explicando fielmente o resultado do modelo de uma instância que pertenceria ao grupo controle. Observa-se ainda nas duas figuras, principalmente na Figura 25, que até mesmo um ser humano teria dificuldades em apontar corretamente a qual grupo este exame pertence. Novamente a explicação trabalha para embasar o resultado do modelo, seja este correto ou não.

Figura 24 – Paciente - Resultado modelo: 0.0155741.



Fonte: Elaborada pelo autor.

Figura 25 – Controle - Resultado modelo: 0.9517495.

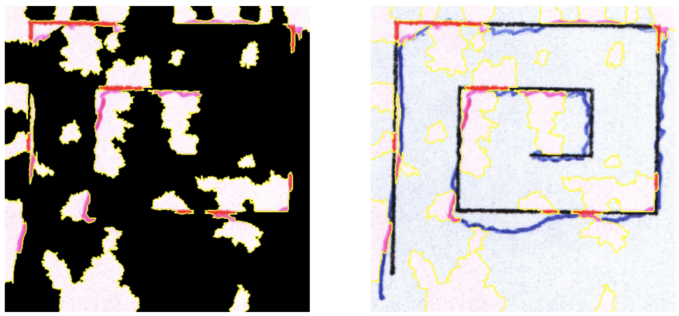


Fonte: Elaborada pelo autor.

Grupo 4 (Figura 26): Observa-se que LIME indicou relevância em vários pontos da imagem de linha reta, mesmo com o resultado do modelo indicar fortemente que a predição

tende a ser Parkinson, além das visíveis áreas onde o traçado de caneta não acompanha a base e que, em grande parte, não foram indicadas como relevantes. Esta instância poderia ser entendida como uma explicação que não necessáriamente condiz com a interpretação humana.

Figura 26 – Paciente - Resultado modelo: 0.97905695.



Fonte: Elaborada pelo autor.

4.1.3 SHAP

Para explicar utilizando o SHAP, de maneira similar ao LIME, deve ser definido o *explainer* no início. Tendo em vista o modelo de aprendizado profundo CNN utilizado, a explicação será feita com o módulo DeepExplainer.

Como parâmetros, é passado o conjunto de imagens que será utilizado juntamente ao modelo, no primeiro parâmetro, para computar as explicações. Este conjunto de imagens é chamado de *background*, neste caso representado por x, como pode ser visto na Figura 27.

Figura 27 – Shap Deep Explainer.

```
# shap DeepExplainer
shap_explainer = shap.DeepExplainer(classifier, x)
```

Fonte: Elaborada pelo autor.

Obtido o *explainer*, não é necessário calculá-lo novamente para cada instância a ser explicada, portanto o mesmo continuará sendo utilizado para todas as imagens. Entretanto, a cada instância foram gerados seus respectivos shap_values utilizando o *explainer* e a imagem a ser explicada. A obtenção dos shap_values está na Figura 28

Figura 28 – Computando SHAP Values.

```
# computar shap values
shap_values = shap_explainer.shap_values(normalized_img2_mc1)
```

Fonte: Elaborada pelo autor.

Feito isso, pode ser realizada a plotagem de cada instância utilizando seus respectivos shap_values, como visto na Figura 29.

Figura 29 – Plotagem dos resultados

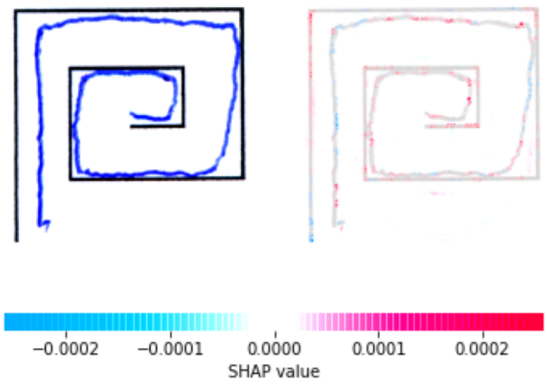
```
# plotagem
shap.image_plot(shap_values, normalized_img2_mc1)
```

Fonte: Elaborada pelo autor.

Com o objetivo de possibilitar a comparação entre os resultados de SHAP e LIME, serão apresentadas explicações das mesmas instâncias vistas anteriormente com LIME e no mesmo agrupamento. Vale ainda dizer que para melhorar a visualização, as explicações do SHAP tiverem seu contraste e saturação aumentados.

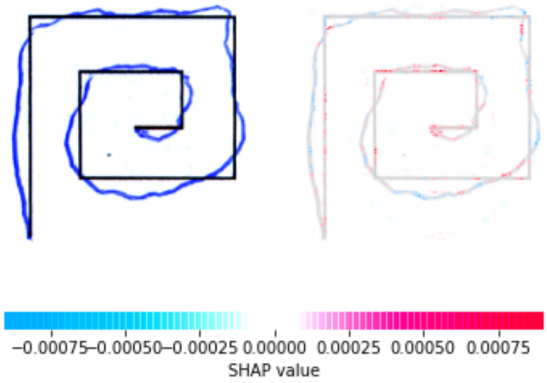
Grupo 1 (Figuras 30 a 32): Explicações de instâncias do grupo de pacientes, que indicam relevância principalmente em áreas da imagem correspondentes aos pontos da imagem com traços de caneta fora do delineado base ou irregulares. Sendo assim, podem ser entendidas como explicações condizentes com a interpretação humana. Observa-se que, em maior parte, os pontos azuis estão presentes onde o delineado está mais distante da base e os vermelhos, de certa forma, onde os traços são mais próximos ao delineado

Figura 30 – Paciente - Resultado modelo: 0.99951226.



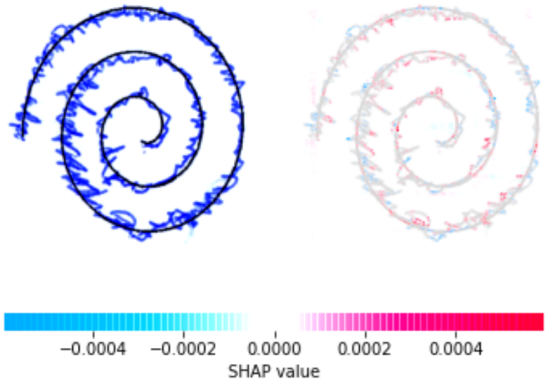
Fonte: Elaborada pelo autor.

Figura 31 – Paciente - Resultado modelo: 0.9995695.



Fonte: Elaborada pelo autor.

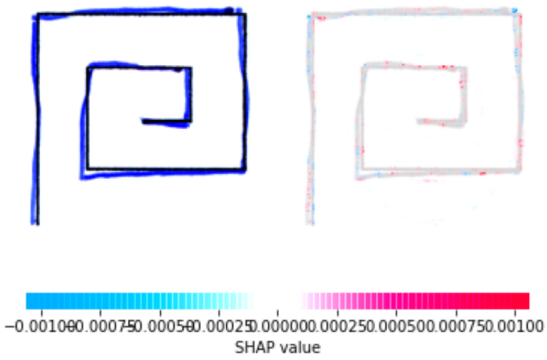
Figura 32 – Paciente - Resultado modelo: 0.9993524.



Fonte: Elaborada pelo autor.

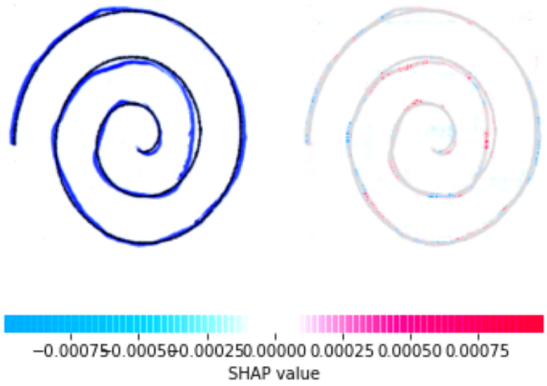
Grupo 2 (Figuras 33 e 34): Explicações de instâncias do grupo de indivíduos saudáveis, que indicam relevância em pontos onde os traços de caneta acompanham ou se mantêm próximos ao delineado base. Novamente poderiam ser entendidas como explicações condizentes com a interpretação humana. A distribuição de pontos azuis e vermelhos se mantém bem próximas tanto entre si quanto em relação ao delineado base, entretanto na Figura 34 observa-se que em dois pontos centrais específicamente, onde pode ser visto que o delineado do indivíduo se distânciou um pouco da base, existem pontos vermelhos, diferente das outras instâncias apresentadas acima desta figura.

Figura 33 – Controle - Resultado modelo: 0.23667398.



Fonte: Elaborada pelo autor.

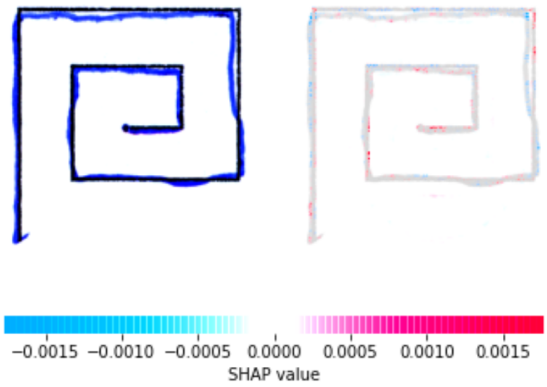
Figura 34 – Controle - Resultado modelo: 0.0436979.



Fonte: Elaborada pelo autor.

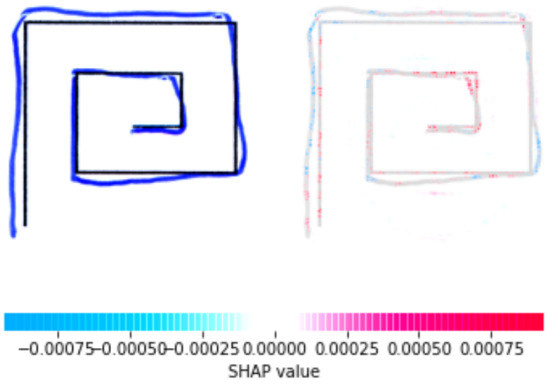
Grupo 3 (Figuras 35 e 36): Explicações para predições incorretas do modelo. Na Figura 35 a distribuição de cores dos pontos se mantém como a maioria das outras instâncias apresentadas, com pontos azuis em locais que apresentam traços mais distantes da base e irregulares, além de ter seus pontos de relevância distruibuídos próximos ao delineado base, representando o resultado do modelo de uma instância que pertenceria ao grupo controle. A Figura 36 tem seus suas cores distribuídas com a mesma lógica, mas os pontos estão mais distribuídos nos locais com traçado que se distanciam da base, podendo ser entendido novamente, que a explicação trabalha para embasar o resultado do modelo.

Figura 35 – Paciente - Resultado modelo: 0.0155741.



Fonte: Elaborada pelo autor.

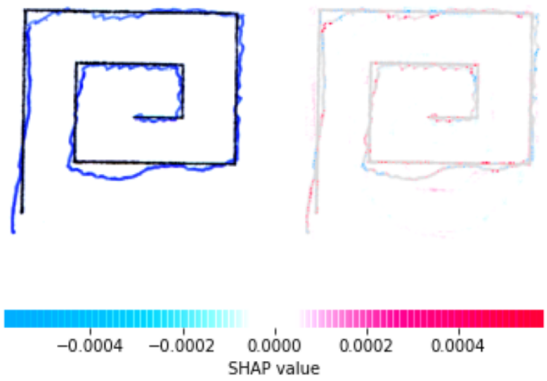
Figura 36 – Controle - Resultado modelo: 0.9517495.



Fonte: Elaborada pelo autor.

Grupo 4 (Figura 37): Os pontos estão presentes principalmente em áreas de traçado distantes da base, como se esperaria de uma predição de paciente do modelo. Entretanto a lógica de distribuição das cores azul e vermelha parece não seguir uma lógica interpretável, tornando difícil sua compreensão, representando uma explicação que, novamente, parece não condizer com a interpretação humana.

Figura 37 – Paciente - Resultado modelo: 0.97905695.



Fonte: Elaborada pelo autor.

É possível observar que as explicações obtidas com SHAP tendem a ser mais regulares no sentido de que os pontos de relevância indicados costumam frequentemente acompanhar os traços de caneta. As explicações com SHAP costumam apontar para *pixels* que de fato se esperaria que impactassem na decisão do modelo.

Entretanto, existe certa confusão ao interpretar as cores que deveriam representar influência positiva e negativa para determinada predição. De instância para instância, mesmo quando fazem parte do mesmo grupo de indivíduos, pode parecer que as cores dos pontos não seguem a mesma lógica.

4.2 Resultados e Discussão

Ao final do desenvolvimento, foi possível obter explicações satisfatórias para o modelo proposto utilizando as duas ferramentas: LIME e SHAP. Ao final da análise, podem ser citados alguns detalhes sobre as implementações das técnicas de explicabilidade com cada ferramenta.

As explicações obtidas através de LIME possuem estrutura que torna mais interessante a sua interpretação; a separação da imagem em áreas e os elementos de explicação baseados nestas áreas contribuem para uma legibilidade intuitiva, já que seres humanos não compreendem a imagem por meio da análise individual de cada *pixel*, mas analisam recortes das áreas de interesse como uma totalidade para alcançar suas conclusões.

Vale ainda dizer que LIME apresenta a quantidade de atributos mais relevantes desejada, contribuindo para se obter com prioridade apenas os pontos que possuem maior grau de influência no resultado final, elevando o nível de interpretação.

Um aparente ponto negativo do LIME nesta implementação seriam algumas áreas que foram obtidas como importantes para a decisão apesar de intuitivamente não conterem informações relevantes, como áreas da imagem pontuadas mesmo vazias, sem nenhum tipo de

traço. Devido a isso, surgem dúvidas quanto ao que teria levado LIME a indicar aquela área.

Em razão do SHAP analisar cada *pixel*, é possível encontrar em suas explicações todos os pontos individuais que teriam influenciado para a decisão. Isso pode ser um fator positivo quando se busca ter um panorama geral de compreensão.

A dinâmica utilizada para interpretar as explicações do SHAP pode ser mais complexa, já que são indicados pontos que influenciam tanto positivamente para uma decisão quanto negativamente. Por conta disso, se faz necessário voltar e repensar a forma que aquela explicação deve ser interpretada. Somado a isso, outro fator que torna menos interpretável é a visibilidade dos *pixels* em meio à imagem, tornando difícil por vezes enxergar os pontos, bem como suas respectivas cores.

Vale pontuar que, devido ao critério de escolha das imagens de exemplo utilizado, foi possível observar explicações que intuitivamente podem parecer corretas para seres humanos e que poderiam ser consideradas como boas explicações, bem como as que podem ir contra esta intuição, cumprindo a intenção inicial da escolha arbitrária das imagens de exemplo, que seria a de gerar discussões sobre o desempenho e as capacidades dos modelos de explicabilidade.

Para o contexto deste experimento, visando avaliar as explicações, podem ser feitas algumas perguntas: Foi possível interpretar a explicação obtida? Faz sentido o que foi apresentado e pode ser considerado correto, levando em conta o raciocínio humano para resolver o mesmo problema? Ao imaginar a situação hipotética: um profissional da saúde apresenta ao paciente a decisão do modelo de classificá-lo como saudável ou apresentando sinais da doença de Parkinson; juntamente à explicação obtida, existem informações suficientes para embasar uma argumentação de forma a convencer o paciente de que o modelo poderia ser confiado?

A depender das respostas para as perguntas acima, seria possível avaliar o desempenho da aplicação das técnicas de XAI. O contexto tem grande relevância para o valor de uma explicação. Neste caso, discutivelmente o fator mais importante seria a capacidade de auxiliar os pacientes na compreensão do resultado de seu exame.

Como uma consideração final quanto aos resultados, tanto LIME quanto SHAP podem ser classificadas como geradoras de boas explicações. Respondendo às perguntas colocadas acima, quanto à interpretabilidade, LIME se sobressaiu ao explicar este modelo de imagem, porém ambas têm explicações interpretáveis. Quanto à confiança que pode ser colocada nas explicações, é sempre necessário atenção quanto a possíveis erros, mas, no panorama geral, os dois podem ser considerados confiáveis. Por fim, respondendo à pergunta hipotética, através dos resultados obtidos das duas ferramentas, pode-se deduzir que as explicações serviriam como base de argumentação para explicar as decisões do modelo, de forma a passar confiança nos resultados, facilitando o caminho de entendimento e auxiliando o profissional da saúde.

5 Conclusão

Neste projeto de conclusão de curso foram apresentadas implementações de técnicas de inteligência artificial explicável utilizando duas das ferramentas mais populares neste contexto: LIME e SHAP, ambas aplicadas a um modelo de rede neural convolucional utilizado para classificar imagens de exames médicos de escrita, pertencentes a um grupo de indivíduos saudáveis e outro grupo de pacientes de Parkinson. O objetivo da implementação foi o de auxiliar na explicação dos resultados obtidos por esta IA.

Ao mesmo tempo que o objetivo de implementação foi alcançado, foi possível apresentar maiores detalhes sobre XAI, bem como analisar as capacidades e desempenho dos métodos utilizados.

Os resultados obtidos apontam que as duas ferramentas utilizadas para obter explicações apresentam bons resultados, ressaltando suas individualidades e seus pontos positivos e negativos.

5.1 Trabalhos Futuros

Considerando uma continuidade deste trabalho, ou mesmo trabalhos distintos, mas ainda no contexto de inteligência artificial explicável, é possível citar as seguintes possibilidades:

- Continuar explorando as capacidades de SHAP em modelos de imagem, possivelmente tentando o mesmo problema com outros módulos, como Kernel Explainer ou Gradient Explainer, que também podem ser aplicados em modelos de aprendizado profundo e têm potencial para gerar outros tipos de explicações que podem ser interessantes.
- Explorar com maior profundidade os benefícios que pode-se alcançar com aplicação de XAI em empresas ou contextos médicos reais.
- Buscar relacionar o processo de pensamento humano à maneira que XAI funciona, de forma a encontrar possíveis semelhanças.
- Abordar outros problemas com XAI, em diferentes tipos de modelos e dados com o objetivo de continuar explorando este tema.

Referências

ALI, A. *Convolutional Neural Network(CNN) with Practical Implementation*. 2019. Disponível em: <https://medium.com/machine-learning-researcher/convlutional-neural-network-cnn-2fc4faa7bb63>. Acesso em: 19 dezembro 2022.

ANKARSTAD, N. *What is Explainable AI (XAI)?*: Logistic regression. 2020. Disponível em: <https://towardsdatascience.com/what-is-explainable-ai-xai-afc56938d513>. Acesso em: 12 novembro 2022.

ARRIETA, e. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward responsible AI*. 2019.

BHATNAGAR, P. *Explainable AI (XAI) - A guide to 7 packages in your python to explain your models*. 2021. Disponível em: <https://towardsdatascience.com/explainable-ai-xai-a-guide-to-7-packages-in-python-to-explain-your-models-932967f0634b>. Acesso em: 23 dezembro 2022.

BRASIL, M. da S. *Doença de Parkinson*. 2022a. Disponível em: <https://bvsms.saude.gov.br/doenca-de-parkinson/>. Acesso em: 14 dezembro 2022.

BRASIL, M. da S. *11/4 – Dia Mundial de Conscientização da Doença de Parkinson: avançar, melhorar, educar, colaborar!* 2022b. Disponível em: <https://bvsms.saude.gov.br/11-4-dia-mundial-de-conscientizacao-da-doenca-de-parkinson-avancar-melhorar-educar-colaborar/>. Acesso em: 21 janeiro 2022.

BROWN, S. *Machine learning, explained*. 2021. Disponível em: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>. Acesso em: 2 dezembro 2022.

CHU, L. *Opening black-box models with LIME- Beauty and the Beast*. 2020. Disponível em: <https://towardsdatascience.com/opening-black-box-models-with-lime-beauty-and-the-beast-9daaf02f584a>. Acesso em: 21 dezembro 2022.

CHU, L. *Model Explainability*. 2022. Disponível em: <https://pub.towardsai.net/model-explainability-shap-vs-lime-vs-permutation-feature-importance-98484efba066>. Acesso em: 14 dezembro 2022.

DWARAKANATH, A.; AHUJA, M.; SIKAND, S.; RAO, R. M.; BOSE, R. J. C.; DUBASH, N.; PODDER, S. Identifying implementation bugs in machine learning based image classifiers using metamorphic testing. In: *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*. [S.l.: s.n.], 2018. p. 118–128.

EINSTEIN, H. I. A. *Parkinson*. 2022. Disponível em: <https://www.einstein.br/doencas-sintomas/parkinson>. Acesso em: 14 dezembro 2022.

GONZALEZ-USIGLI, H. A. *Doença de Parkinson (DP)*. 2022. Disponível em: <https://www.msdmanuals.com/pt-br/casa/distÃžrbios-cerebrais,-da-medula-espinal-e-dos-nervos/doenÃğas-do-movimento/doenÃğa-de-parkinson-dp>. Acesso em: 21 janeiro 2022.

HULSTAERT, L. *Understanding Model Predictions with LIME*. 2018. Disponível em: <https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b>. Acesso em: 20 novembro 2022.

IBM. *Convolutional Neural Networks*. 2022a. Disponível em: <https://www.ibm.com/topics/convolutional-neural-networks>. Acesso em: 5 dezembro 2022.

IBM. *Explainable AI (XAI)*. 2022b. Disponível em: <https://www.ibm.com/watson/explainable-ai>. Acesso em: 13 novembro 2022.

IBM. *What is Deep Learning?* 2022c. Disponível em: <https://www.ibm.com/topics/deep-learning>. Acesso em: 5 dezembro 2022.

IBM. *What is Machine Learning?* 2022d. Disponível em: <https://www.ibm.com/topics/machine-learning>. Acesso em: 14 dezembro 2022.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, v. 30, 2017.

MICROSOFT. *O que é um modelo de machine learning?* 2022. Disponível em: <https://learn.microsoft.com/pt-br/windows/ai/windows-ml/what-is-a-machine-learning-model>. Acesso em: 16 setembro 2022.

NETO, M. G. *Explainable AI: Extraindo explicações e aumentando a confiança dos modelos de ML*. 2021.

PARSONS, C. *What Is a Machine Learning Model?* 2021. Disponível em: <https://blogs.nvidia.com/blog/2021/08/16/what-is-a-machine-learning-model/>. Acesso em: 03 janeiro 2023.

PEREIRA, C. R.; PEREIRA, D. R.; SILVA, F. A.; MASIEIRO, J. P.; WEBER, S. A. T.; HOOK, C.; PAPA, J. P. A new computer vision-based approach to aid the diagnosis of parkinson's disease. *Computer Methods and Programs in Biomedicine*, Elsevier North-Holland, Inc., New York, NY, USA, v. 136, p. 79–88, 2016.

RIBEIRO, M. *LIME - Local Interpretable Model-Agnostic Explanations*. 2016. Disponível em: <https://homes.cs.washington.edu/~marcotcr/blog/lime/>. Acesso em: 16 novembro 2022.

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.

RIBEIRO SAMEER SINGH, C. G. M. *Local Interpretable Model-Agnostic Explanations (LIME): An introduction*. 2016. Disponível em: <https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/>. Acesso em: 22 dezembro 2022.

ROYALSOCIETY.ORG. *Explainable AI*. 2019. Disponível em: <https://royalsociety.org/topics-policy/projects/explainable-ai/>. Acesso em: 20 dezembro 2022.

SHAP. *Documentation*. 2022. Disponível em: <https://shap.readthedocs.io/en/latest/>. Acesso em: 15 outubro 2022.

THORN, J. *Explainable Artificial Intelligence*. 2020. Disponível em: <https://towardsdatascience.com/explainable-artificial-intelligence-14944563cc79>. Acesso em: 8 dezembro 2022.

THUNG, F.; WANG, S.; LO, D.; JIANG, L. An empirical study of bugs in machine learning systems. In: IEEE. *2012 IEEE 23rd International Symposium on Software Reliability Engineering*. [S.l.], 2012. p. 271–280.

TUREK, M. *Explainable Artificial Intelligence (XAI)*. 2022. Disponível em: <https://www.darpa.mil/program/explainable-artificial-intelligence>. Acesso em: 22 dezembro 2022.

ZITNIK, M.; NGUYEN, F.; WANG, B.; LESKOVEC, J.; GOLDENBERG, A.; HOFFMAN, M. M. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, Elsevier, v. 50, p. 71–91, 2019.