Capítulo 2

Explorando a Explicabilidade da Inteligência Artificial - Técnicas para Compreender e Interpretar Modelos de Aprendizado de Máquina

Júlio V. M. Marques, Clésio A. Gonçalves, Pablo de Abreu Vieira, Armando L. Borges, Viviane B. Leal Dias, Willians S. Santos e Romuere R. V. Silva

Abstract

Artificial Intelligence (AI) has become powerful, but explainability is crucial for its reliability in sensitive areas. Interpretable models offer transparent insights, while "black boxes"lack explanation. Visualization techniques, such as heatmaps, aid in understanding. Pursuing explainability is ethical and essential to ensure trust and accountability in applications, especially in medical diagnostics. In this context, this book chapter explores its significance, techniques, evaluation, applications, challenges, limitations, and key conclusions in this field.

Resumo

A Inteligência Artificial (IA) tornou-se poderosa, mas a explicabilidade é crucial para sua confiabilidade em áreas sensíveis. Modelos interpretáveis oferecem insights transparentes, enquanto as "caixas pretas"carecem de explicação. Técnicas de visualização, como mapas de calor, auxiliam na compreensão. Buscar explicabilidade é ético e essencial para garantir a confiança e a responsabilidade em aplicações, especialmente em diagnósticos médicos. Nesse contexto, este capítulo de livro explora sua importância, técnicas, avaliação, aplicações, desafios, limitações e as principais conclusões dessa área.

2.1. Introdução

Nos últimos anos, a Inteligência Artificial (IA) consolidou-se como uma poderosa ferramenta com aplicações que abrangem desde diagnósticos médicos [Grif and Avush 2018] até à condução autônoma de veículos [Nivas et al. 2016]. No entanto, à medida que a complexidade dos modelos de IA aumenta, a demanda por compreender e explicar suas decisões torna-se crítica [Arrieta et al. 2020]. A explicabilidade na IA desempenha um

papel fundamental na construção de sistemas responsáveis e éticos, garantindo a confiança dos usuários e a conformidade com regulamentações, especialmente em setores sensíveis.

Existem diferentes tipos de IA, algumas são interpretáveis e outras são consideradas caixas pretas. Os modelos interpretáveis são projetados para serem transparentes, proporcionando uma visão direta de como tomam decisões e fazem previsões, assim, revelando o processo de tomada de decisão e oferecendo *insights* compreensíveis sobre como chegam a suas conclusões [Agarwal and Das 2020], exemplos desse modelo são as árvores de decisão [Jou 1986], regressões lineares [Yan and Su 2009] e máquinas de vetores de suporte [Cortes and Vapnik 1995]. Os modelos consideradas "caixas pretas"produzem previsões ou decisões sem oferecer uma explicação clara do raciocínio subjacente [Ribeiro et al. 2016a]. Isso pode ser especialmente problemático em cenários críticos, como cuidados de saúde e justiça criminal, onde a compreensão das razões por trás das decisões é crucial. Além disso, a transparência é essencial para evitar que os modelos de IA perpetuem informações indesejadas. Portanto, a explicabilidade desempenha um papel essencial na construção de sistemas de IA confiáveis e morais. Isso levou à adoção de modelos de aprendizado de máquina intrinsecamente interpretáveis.

Adicionalmente, as técnicas de visualização, como os mapas de calor, desempenham um papel fundamental na compreensão de modelos complexos, como redes neurais. Os mapas de calor destacam as partes mais influentes dos dados de entrada, especialmente em tarefas de visão computacional, onde revelam quais regiões de uma imagem são mais relevantes para uma determinada classificação [Selvaraju et al. 2019].

A busca pela explicabilidade na IA é uma resposta essencial à crescente complexidade dos modelos atuais e às necessidades de setores críticos. A transparência não é apenas uma questão ética, mas também uma exigência crucial para construir uma aplicação de confiança. À medida que a tecnologia avança, a comunidade científica e os profissionais da área continuarão aprimorando as técnicas de explicabilidade, com o objetivo de impulsionar a confiança, a responsabilidade e a aplicação ética em uma ampla gama de campos, que incluem resolução de desafios complexos como diagnósticos médicos.

Este trabalho está estruturado da seguinte forma: a Seção 2.2 explora a Importância da Explicabilidade na Inteligência Artificial; a Seção 2.3 discute as Técnicas de Explicabilidade, incluindo Modelos Interpretáveis, Mapas de Calor (*Heatmaps*), *Class Activation Mapping* (Grad-CAM) e *Local Surrogate* (LIME); a Seção 2.4 aborda a Avaliação da Explicabilidade; a Seção 2.5 destaca Aplicações Práticas; a Seção 2.6 explora os Desafios e Limitações; por fim, a Seção 2.7 oferece principais *insights* e conclusões.

2.2. Importância da Explicabilidade na Inteligência Artificial

Modelos de IA interpretáveis tornam possível que suas previsões sejam examinadas e explicadas, sendo possível identificar quaisquer desvios indesejados e até mesmo antiéticos em seu comportamento e forma de analisar os dados. Neste contexto, esta seção aborda alguns fatores que destacam a importância da explicabilidade de modelos de IA e a análise de suas previsões para identificar comportamentos inadequados do ponto de vista ético e técnico.

Segundo [Pedreshi et al. 2008], modelos de classificação treinados com dados históricos podem ser discriminatórios no sentido social e negativo, visto que estes dados podem conter padrões preconceituosos fortemente enraizados na sociedade atual. Neste contexto, partindo da ideia de que esses modelos de IA podem, e são utilizados como forma de apoio a decisão em vários contextos, como liberação de crédito bancário ou acesso a serviços públicos, é evidente que eles podem incluir comportamentos socialmente, racialmente e etnologicamente segregativos, prejudicando negativamente classes menos favorecidas, replicando assim, comportamentos antiéticos tradicionais.

Nas décadas de 1970 e 1980, a Faculdade de Medicina do Hospital St George's usou um algoritmo para triagem de candidatos que concorriam a oportunidades de emprego na instituição. Este programa utilizava informações contidas nos formulários do candidatos, o qual não continha qualquer referência étnica. No entanto, descobriu-se que o programa discriminava de forma injusta as minorias étnicas e pessoas do sexo feminino, inferindo esta informação através dos seus nomes e locais de nascimento, reduzindo suas probabilidades de serem selecionados pelo algoritmo e para uma futura entrevista [Lowry and Macpherson 1988]. Comportamentos como este podem se repetir em muitos outros casos em que é aplicado o uso da IA para realizar tomadas de decisão. O estudo conduzido por [Caliskan et al. 2017] revela preconceitos humanos encontrados em textos e corpus da web, onde constatou-se que nomes de pessoas negras estavam mais associados a termos negativos e desagradáveis em comparação com nomes de pessoas brancas. Com isso, é possível notar que a principal causa dos comportamentos negativamente discriminatórios dos sistemas baseados em IA advém dos dados que são usados em seu treino, visto que eles podem conter vieses antiéticos, mesmo que sutis, mas que podem levar esses sistemas a tais tipos de atitude.

A IA está continuamente sendo aplicada e aprimorada nos mais diversos campos e áreas do conhecimento, principalmente na área da saúde para auxiliar no diagnóstico de doenças por imagens [Borges et al. 2022] [Gennatas and Chen 2021]. Aplicações baseadas em sistemas como esse tem um grande potencial de auxiliar profissionais no desempenho de suas atividades. No entanto, diversos modelos de IA, com ênfase naqueles baseados em redes neurais profundas, não oferecem esclarecimentos claros sobre como chegaram em determinada conclusão, e portanto, são conhecidos como "caixas-pretas"[Rudin 2019]. Esse problema pode se tornar mais grave quando trazido a contextos delicados, em que é de suma importância a garantia de que determinado fator ou objeto tenha sido analisado e levado em conta pelo modelo antes do mesmo dar um resposta a inferência, como por exemplo, na análise de exames médicos. Neste contexto, tais modelos podem aprender padrões presentes em sua base de treino que são irrelevantes ao diagnóstico enquanto descarta os padrões que deveriam ser unicamente levados em conta, podendo levar a erros ocasionais em futuras inferências.

Estas situações podem se replicar em quaisquer outros contextos, de outros campos e áreas. Portanto, é necessário entender como estes modelos funcionam e como seus resultados podem ser explicados e justificados, com o intuito de compreender os elementos que estão sendo analisados e considerados pelo sistema para formular suas respostas, evitando assim, graves erros, principalmente em contextos delicados, nos quais eles podem ocasionar em danos e/ou prejuízos para as pessoas. Para isso, existem diversas técnicas de explicabilidade, como Mapas de Calor (*Heatmaps*), *Class Activation Map-*

ping (Grad-CAM), *Local Surrogate* (LIME) e os próprios Modelos Interpretáveis. Tais técnicas serão discutidas e analisadas ao longo do capítulo.

2.3. Técnicas de Explicabilidade

Nesta seção, é explorado uma variedade de técnicas de explicabilidade em *machine learning* e inteligência artificial. Essas técnicas são projetadas para tornar os modelos de aprendizado de máquina mais transparentes e compreensíveis, permitindo que os usuários entendam como e por que os modelos tomam decisões.

2.3.1. Modelos Interpretáveis

Os modelos interpretáveis são modelos de aprendizado de máquina que têm a capacidade de serem explicados e compreendidos de forma direta. Esses modelos são construídos com a intenção de manter uma relação clara entre as características de entrada, as saídas ou previsões, tornando-os transparentes e interpretáveis. A interpretabilidade é uma característica importante em muitas aplicações de IA, especialmente em áreas onde a transparência, a responsabilidade e a confiança são críticas, como medicina, direito, finanças e sistemas de suporte à decisão. Aqui estão alguns exemplos de modelos interpretáveis:

- **Regressão Linear** [Yan and Su 2009]: Um modelo linear estabelece uma relação linear entre as características de entrada e a variável de saída. É possível ver diretamente como as características afetam a saída por meio dos coeficientes, contribuindo para uma interpretabilidade clara.
- **Árvores de Decisão** [Jou 1986]: As árvores de decisão dividem os dados em um conjunto de regras hierárquicas baseadas nas características de entrada. A interpretabilidade é natural, pois pode-se seguir o caminho da árvore para entender as decisões.
- **Regressão Logística** [Cox 1958]: A regressão logística é usada para problemas de classificação binária e fornece uma interpretação direta dos coeficientes em relação às características.
- **Regras de Associação** [Pei 2009]: Esses modelos geram regras do tipo "se... então..."com base nas relações entre as características de entrada e a variável de saída, Gerando assim a interpretabilidade.
- **Máquinas de Vetores de Suporte (SVM)** [Cortes and Vapnik 1995]: Quando um *kernel* linear é usado em SVMs, o limite de decisão é uma linha reta, tornando o modelo interpretável em problemas de classificação.

Algumas técnicas podem ser aplicadas a modelos interpretáveis para fornecer uma compreensão mais completa de como eles funcionam e tomam decisões. A escolha das técnicas específicas depende do modelo, do problema e do público-alvo. Dentre elas, destacam-se:

- **Visualização de Coeficientes**: Em modelos lineares, como a regressão linear ou logística, pode-se visualizar diretamente os coeficientes atribuídos a cada característica, destacando a influência de cada uma.

- **Gráficos de Barras de Importância de Características**: Para árvores de decisão e modelos baseados em regras, pode-se criar gráficos de barras que mostram a importância relativa de cada característica na tomada de decisões.
- **Gráficos de Dependência Parcial**: Esses gráficos mostram como a variável de saída se relaciona com uma variável de entrada específica, mantendo outras variáveis constantes. São úteis para ilustrar as relações entre as características e as previsões em modelos como árvores de decisão.
- **Matrizes de Correlação**: Em modelos interpretáveis, como regressões lineares, exibir matrizes de correlação pode ajudar a identificar relações lineares entre características e a variável de saída.
- **Análise de Resíduos**: Para modelos de regressão, a análise de resíduos ajuda a avaliar as suposições do modelo.
- **Regras de Explicação**: Para modelos baseados em regras, pode explicar as decisões por meio de regras simples do tipo "se... então...", facilitando a compreensão das decisões do modelo.

2.3.2. Mapas de Calor (*Heatmaps*)

Mapas de calor (*heatmaps*) são uma técnica de explicabilidade amplamente utilizada em modelos de IA para visualizar e comunicar como as características de entrada afetam as previsões ou saídas do modelo. Essa técnica é especialmente útil quando se deseja entender como as características contribuem para as decisões do modelo em dados tabulares.

Os mapas de calor são normalmente aplicados a uma única instância de dados de entrada ou a um conjunto de dados de entrada. Cada instância de dados contém características (atributos) que são alimentadas ao modelo de IA. Para cada característica de entrada, o mapa de calor calcula a contribuição relativa dessa característica para a saída ou previsão do modelo.

As contribuições são representadas visualmente por meio de cores em um mapa de calor. Geralmente, as cores quentes, como vermelho ou amarelo, indicam uma contribuição positiva da característica para a saída, enquanto as cores frias, como azul ou verde, indicam uma contribuição negativa ou nula. A intensidade da cor pode representar a magnitude da contribuição.

O mapa de calor é gerado de forma que cada característica seja mapeada em um eixo (horizontal ou vertical) e a saída do modelo ou a importância da característica seja mapeada no outro eixo. Isso cria uma representação visual que destaca quais características têm maior impacto nas previsões ou saídas do modelo. A Figura 2.1 abaixo ilustra um exemplo de mapa de calor. Podemos notar a concentração populacional por meio das cores, onde, quanto mais escura a tonalidade maior é a densidade populacional. Por meio desses mapas de calor podemos extrair informações visuais das regiões com maiores significância para a tarefa estudada.

O uso dos mapas de calor na explicabilidade de modelos de IA contribui principalmente na interpretação de modelos complexos, identificação de características impor-

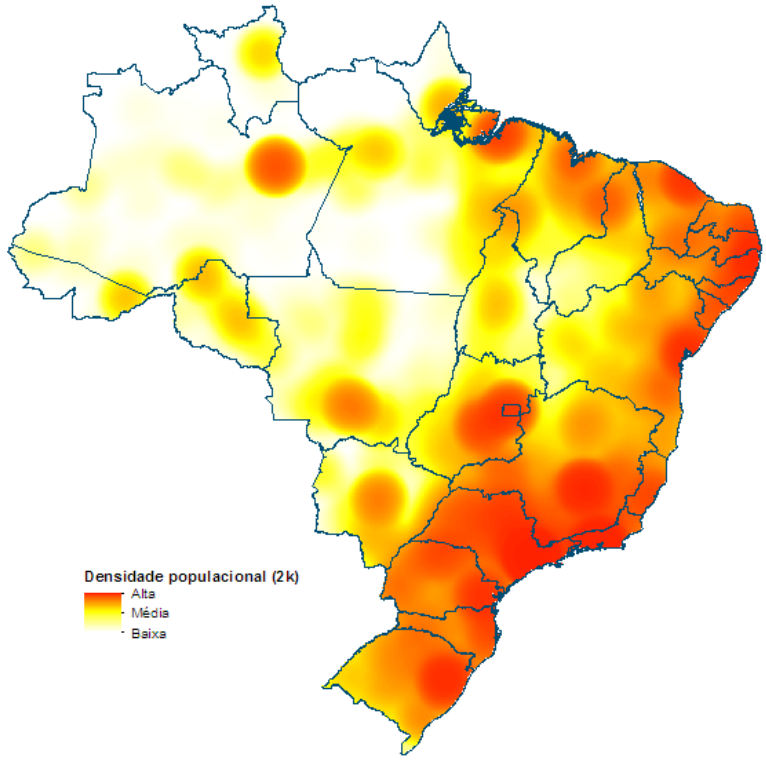

Figura 2.1: Densidade populacional do Brasil em 2010. As áreas com cores mais quentes representam uma maior densidade populacional. Fonte: [Forest-gis]

tantes, detecção de relações não lineares, diagnóstico de erros, além da validação e ajuste de modelos.

É importante notar que a interpretação de mapas de calor pode ser subjetiva, e os resultados podem depender da técnica específica de cálculo da contribuição das características. Portanto, é aconselhável usar mapas de calor em conjunto com outras técnicas de explicabilidade para obter uma compreensão completa do comportamento do modelo de IA.

2.3.3. *Class Activation Mapping* (Grad-CAM)

O Grad-CAM, ou *Class Activation Mapping*, é uma técnica de interpretação de modelos de aprendizado profundo que ajuda a visualizar quais partes de uma imagem são mais influentes na tomada de decisão do modelo durante a classificação [Selvaraju et al. 2017]. Ele se tornou uma ferramenta essencial para entender como as redes neurais convolucionais (CNNs) "olham"para as imagens e é amplamente utilizado em tarefas de visão computacional, como classificação de imagens e detecção de objetos.

O Grad-CAM é aplicado principalmente a modelos de CNN, que são amplamente usados em tarefas de visão computacional. Esses modelos são compostos por várias camadas convolucionais que extraem características relevantes das imagens. Essas camadas convolucionais produzem mapas de ativação, que destacam as regiões da imagem onde certas características foram detectadas. Quanto mais profunda a camada, mais abstratas

são as características que ela representa. Em um modelo de classificação de imagem, as camadas convolucionais são seguidas por camadas totalmente conectadas que fazem a decisão final sobre a classe à qual a imagem pertence. O Grad-CAM se concentra em entender como essas camadas totalmente conectadas ponderam as características das camadas convolucionais.

O processo pode ser explicado da seguinte forma: Primeiramente, uma imagem é alimentada ao modelo, que gera uma previsão. A classe alvo é determinada a partir dessa previsão. Em seguida, o Grad-CAM calcula os gradientes da classe alvo em relação às ativações na camada convolutiva mais profunda. Isso indica quais ativações foram mais influentes na decisão da classe. Então o Grad-CAM combina os gradientes com as ativações da camada convolutiva usando uma média ponderada [Chattopadhay et al. 2018]. Essa média ponderada é usada para obter os pesos de ativação de cada canal na camada convolutiva. Por fim, os pesos de ativação são usados para criar um mapa de ativação que destaca as regiões da imagem que mais contribuíram para a decisão da classe alvo. Na Figura 2.2, temos uma demonstração da utilização do Grad-CAM para visualizar as zonas de ativações do modelo. As zonas em azul, se concentram nas regiões que os modelos consideram mais importante para a tarefa, podemos notar que em sua maioria as regiões se concentram dentro do parênquima pulmonar, o que mostra que o modelo realmente está levando em consideração regiões que são de fato importante para a detecção da COVID-19.

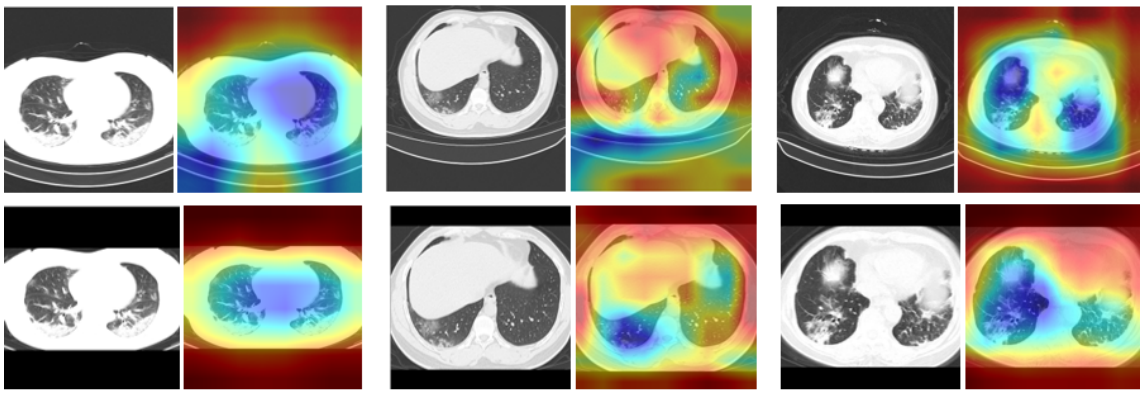

Figura 2.2: Zonas de ativações geradas pelo Grad-CAM para um modelo de detecção de COVID-19 em imagens de tomografia computadorizada. Fonte: [Marques et al. 2023]

O Grad-CAM oferece várias vantagens, ele torna as decisões dos modelos de *deep learning* mais transparentes, permitindo que os usuários compreendam quais características da imagem levaram à classificação. O Grad-CAM não apenas indica a classe alvo, mas também localiza as regiões da imagem que foram mais relevantes para essa classificação, fornecendo *insights* sobre como esses modelos tomam decisões. Sua capacidade de localizar características relevantes em imagens torna-o valioso em uma variedade de domínios, desde medicina, mostrando as regiões em destaques que revelam características relevantes para o problema, ate segurança, podendo ser utilizado para identificar regiões onde o modelo consideram importantes para manter a segurança. À medida que a pes-

quisa em interpretabilidade de modelos de aprendizado profundo avança, o Grad-CAM continua sendo uma das técnicas mais amplamente utilizadas.

2.3.4. *Local Surrogate* (LIME)

O *Local Interpretable Model-agnostic Explanations* (LIME) é uma técnica que visa tornar os modelos de aprendizado de máquina mais transparentes e interpretáveis, especialmente em nível local [Ribeiro et al. 2016b]. Ele se concentra em explicar as previsões de modelos de *machine learning* para instâncias de dados individuais, permitindo que os usuários entendam como o modelo chegou a uma decisão específica para um caso particular.

Modelos de aprendizado de máquina modernos, como redes neurais profundas, são frequentemente complexos e difíceis de interpretar. O LIME aborda essa complexidade fornecendo explicações compreensíveis para previsões de modelos de alta dimensionalidade. O LIME cria modelos locais (*surrogates*) que são muito mais simples do que o modelo de *machine learning* original, mas ainda são explicativos. Esses modelos locais são criados para imitar o comportamento do modelo original em torno da instância de dados de interesse. O LIME funciona introduzindo pequenas alterações ou modificações nos dados de entrada da instância que está sendo analisada e observando como as previsões do modelo original mudam em resposta a essas alterações. Isso ajuda a entender a sensibilidade do modelo à variação nos dados de entrada.

O processo do LIME envolve os seguintes passos: Primeiramente, uma instância de dados individual para a qual se deseja explicar a previsão é selecionada. O LIME cria uma série de instâncias com pequenas alterações a partir da instância de interesse. Essas variações e alterações podem ser feitas de várias maneiras, como adição de ruído ou remoção de recursos. As instâncias com alterações são alimentadas ao modelo de *machine learning* original, e suas previsões são registradas. Um modelo local, geralmente linear ou outro modelo, é treinado usando as instâncias com alterações e as previsões correspondentes do modelo original. Esse modelo local atua como um substituto ou *surrogates* do modelo original. O modelo local é interpretável, permitindo que os usuários entendam as relações entre os recursos de entrada e as previsões do modelo original para a instância de interesse.

O LIME oferece várias vantagens: Ele fornece explicações interpretáveis para previsões de modelos complexos em nível local, o que é útil quando se deseja entender o raciocínio do modelo para casos específicos. O LIME é "agnóstico"em relação ao tipo de modelo de *machine learning* usado, o que significa que pode ser aplicado a uma ampla variedade de modelos sem necessidade de conhecimento interno sobre eles. Pode ser aplicado em várias tarefas, incluindo classificação, regressão e até mesmo em tarefas de processamento de linguagem natural. Na Figura 2.3, podemos observar uma representação do modelo LIME para varias entradas. Nessa figura, temos as instâncias de entrada com variações e alterações, o modelo LIME interpreta essa variação para cada instância de entrada e tenta representa-la por meio de cores, mostrando a região de ativação para cada entrada.

O LIME é uma ferramenta valiosa para tornar os modelos de *machine learning* mais interpretáveis, fornecendo explicações compreensíveis para previsões em nível local. Sua capacidade de criar modelos locais simples que imitam o comportamento de

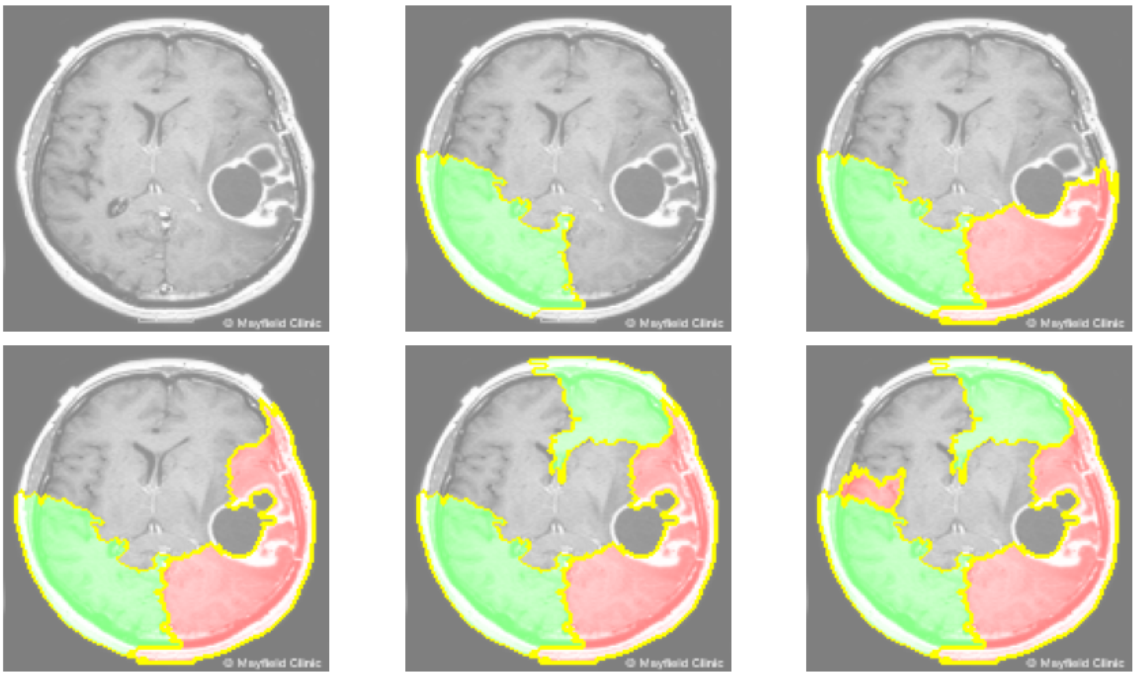

Figura 2.3: Zonas de ativações geradas pelo LIME para um modelo de detecção de tumores.

modelos complexos torna-o útil em uma variedade de domínios, como a medicina, onde é possivel obter uma explicação dos resultados de modelos para a detecção de glaucoma [Volkov and Averkin 2023], ajudando os usuários a entender como e por que os modelos tomam decisões específicas para casos individuais.

2.4. Avaliação da Explicabilidade

A interpretabilidade e explicabilidade desempenham um papel crucial na adoção e confiança em modelos de IA em uma ampla variedade de domínios, desde a medicina, a área jurídica, onde aplicações podem ser utilizadas para tratar de privacidade de dados e legislação [Thommandru et al. 2023] e até na área financeira, com modelos capazes de analisar contramedidas financeiras para o desenvolvimento da economia [Shuguang 2011]. Nesta seção, exploramos como avaliar a explicabilidade de modelos de IA e as técnicas discutidas anteriormente, como Modelos Interpretáveis, Mapas de Calor, Grad-CAM e LIME.

A avaliação da explicabilidade é essencial para determinar a eficácia das técnicas utilizadas e garantir que os resultados sejam confiáveis e úteis para os usuários finais. Várias abordagens podem ser consideradas ao avaliar a explicabilidade de um modelo: Uma maneira de avaliar a explicabilidade é por meio de métricas quantitativas que mensuram o desempenho do modelo na explicação de suas decisões. Por exemplo, para o Grad-CAM e Mapas de Calor, pode-se avaliar a precisão com que essas técnicas identificam as regiões de influência nas imagens em relação às previsões reais. Para o LIME, pode-se medir o quão bem os modelos locais criados se aproximam das previsões do modelo ori-

ginal para as instâncias de interesse. Métricas como precisão, *recall* e erro médio podem ser usadas dependendo da tarefa como classificação ou segmentação. Além das métricas quantitativas, a avaliação qualitativa desempenha um papel crucial na compreensão da explicabilidade. Isso envolve a revisão e interpretação das explicações fornecidas pelo modelo. Os usuários podem avaliar a adequação das explicações à tarefa em questão, verificar se as informações fornecidas são compreensíveis e se correspondem às expectativas. Essa avaliação geralmente envolve especialistas humanos que examinam as explicações geradas pelo modelo.

Uma abordagem útil na avaliação da explicabilidade é comparar o desempenho das técnicas utilizadas com modelos de referência ou *benchmarks*. Isso ajuda a determinar se as técnicas estão fornecendo explicações significativas e superando abordagens padrão. Por exemplo, ao usar o Grad-CAM para a detecção de objetos em imagens, pode-se comparar seu desempenho com métodos tradicionais de segmentação de objetos para verificar se ele fornece *insights* adicionais. Também podemos coletar *feedback* de usuários reais que interagem com o sistema alimentado por modelos de IA para fornecer informações valiosas sobre a eficácia das explicações. Isso pode ajudar a identificar áreas que requerem melhorias e ajustes nas técnicas de explicabilidade para atender às necessidades dos usuários.

A avaliação da explicabilidade deve levar em consideração o contexto e o domínio da aplicação. O que é considerado uma explicação eficaz pode variar dependendo da tarefa. Em algumas situações, uma explicação visual, como um mapa de calor, pode ser mais adequada, enquanto em outras, uma explicação textual ou baseada em regras pode ser preferível. É importante adaptar as técnicas de explicabilidade ao contexto específico. Em resumo, a avaliação da explicabilidade é um componente crítico na implantação de modelos de IA em aplicações do mundo real. Métricas quantitativas, avaliação qualitativa, comparação com modelos de referência, *feedback* dos usuários e consideração do contexto são todos aspectos importantes a serem considerados. A escolha das técnicas de explicabilidade também deve ser orientada pelo objetivo da interpretabilidade em um determinado cenário. Ao adotar uma abordagem abrangente de avaliação, é possível garantir que os modelos de IA sejam mais transparentes, confiáveis e éticos em sua tomada de decisão.

2.5. Aplicações Práticas

As aplicações práticas da IA são tecnologias que utilizam algoritmos e modelos de IA para solucionar uma variedade de problemas cotidianos, seja de forma autônoma ou com assistência, com a finalidade de emular a capacidade cognitiva humana. Estas aplicações operam através da aquisição e análise de dados, empregando o aprendizado automático para efetuar escolhas ou realizar ações com base nas informações disponíveis. Elas têm uma ampla variedade de finalidades e oferecem benefícios em diversas áreas na sociedade.

A IA explicável pode ser definida como aquela que produz informações ou argumentos que tornam seu funcionamento de fácil compreensão. Neste contexto, a interpretabilidade pode ser definida como a capacidade de um modelo de explicar ou fornecer o significado em termos compreensíveis para humanos, afirma [Arrieta et al. 2020]. A interpretabilidade da IA é crucial, visto que a IA está se tornando cada vez mais integrada à

vida cotidiana da sociedade. É fundamental que as pessoas possam entender e confiar nas decisões e ações executadas por esses sistemas. Nesta seção serão apresentados exemplos de aplicações com IA e sua interpretabilidade em diferentes setores.

2.5.1. Aplicação na área da saúde

Na área da medicina, a IA desempenha um papel essencial com diversas aplicações que transformam a prática médica. Uma das aplicações de destaque da IA é a detecção de doenças por meio da análise de imagens médicas, como radiografias e ressonâncias magnéticas. Essa abordagem automatizada é significativamente mais rápida do que os métodos tradicionais, uma vez que os computadores podem processar grandes volumes de dados em um intervalo de tempo muito menor em comparação com um médico humano.

No trabalho proposto por [Ahsan et al. 2020] o LIME é utilizado para identificar os recursos específicos em imagens de radiografia de tórax, as imagens tem duas classes, pacientes com COVID-19 e pacientes sem a doença. A Figura 2.4 apresenta a metodologia usada. Após as imagens de entrada, o LIME é aplicado e encontra o conjunto de *superpixels* que contém a ligação mais válida com um rótulo de previsão. Em outras palavras, o LIME está tentando entender quais partes da imagem são mais influentes para a previsão feita pelo modelo.

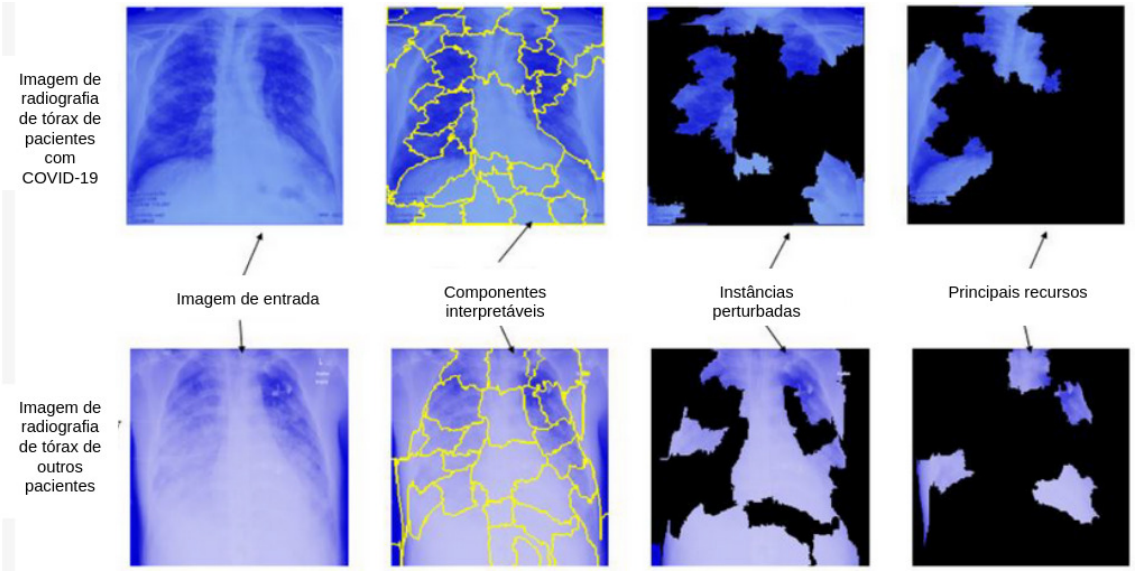

Figura 2.4: Uso do LIME em imagens de radiografia de tórax. Fonte: [Ahsan et al. 2020]

No entanto, a verdadeira eficácia dessa colaboração entre IA e especialistas em saúde reside na interpretabilidade das conclusões. A interpretabilidade reflete a capacidade de fornecer clareza sobre a localização das características indicativas da doença na imagem, permitindo aos médicos uma compreensão precisa e rápida dos resultados. Além da detecção de doenças, a IA na medicina abrange áreas como o desenvolvimento de tratamentos personalizados com base em genômica, a análise de dados de pacientes em larga escala para identificar tendências e aprimorar a gestão de hospitais e clínicas. A interpre-

tabilidade desempenha um papel vital nessas áreas, garantindo que as recomendações da IA sejam compreensíveis e confiáveis para os profissionais de saúde, promovendo, assim, uma assistência médica mais precisa e eficaz.

2.5.2. Aplicação na área da comunicação e marketing

A IA tem desempenhado um papel revolucionário no setor de comunicação e marketing, transformando a maneira como as empresas se relacionam com seus consumidores. Uma profunda análise de grandes conjuntos de informações pode fornecer às empresas valiosas informações sobre o comportamento, preferências e necessidades de seus públicos-alvo, resultando em estratégias de marketing mais eficazes.

A personalização de campanhas publicitárias é outra área em que a IA se destaca. Ao explorar minuciosamente os dados dos consumidores, a IA identifica padrões e características individuais, permitindo que as empresas adaptem suas mensagens de marketing com precisão para atender segmentos específicos do mercado. O resultado são campanhas mais direcionadas e envolventes, estabelecendo conexões mais sólidas entre os consumidores e os produtos ou serviços oferecidos.

A criação automatizada de conteúdo publicitário também desempenha um papel significativo na IA aplicada à comunicação e marketing. A IA tem a capacidade de gerar textos, imagens e vídeos com base em critérios definidos e dados disponíveis, acelerando consideravelmente o processo de produção de conteúdo, ao mesmo tempo em que garante consistência e eficiência na criação de mensagens de marketing. Na figura 2.5, podemos observar uma nuvem de palavras [Jin 2017]. Os modelos de processamento de linguagem natural (PLN) [Iqbal and Qureshi 2022], são bastantes utilizados para a geração de texto. Para entender o que esses modelos estão aprendendo ou levando em consideração é possível utilizar as nuvens de palavras para tentar identificar os padrões utilizados pelos modelos.

Figura 2.5: Nuvens de palavras. Fonte: [Jin 2017]

Em todas essas aplicações, a interpretabilidade da IA desempenha um papel crucial. É fundamental que as empresas compreendam de forma clara como a IA toma suas

decisões, possibilitando ajustes nas estratégias para que estejam alinhadas com seus objetivos e valores. A habilidade de explicar o processo de tomada de decisão da IA promove a transparência e confiança, elementos essenciais para a construção de relacionamentos sólidos com os consumidores e para garantir o sucesso a longo prazo das estratégias de marketing baseadas em IA. Em outras palavras, a interpretabilidade não só aprimora a eficácia da IA, mas também fortalece a conexão entre as empresas e seus públicos, moldando de forma positiva o futuro da comunicação e marketing.

2.5.3. Aplicação na área da agricultura

A agricultura moderna se beneficia da aplicação da IA, uma vez que desempenha um papel essencial na otimização dos recursos agrícolas, tais como irrigação, fertilizantes e pesticidas. A IA opera por meio da coleta de dados provenientes de uma variedade de fontes, incluindo sensores e satélites, permitindo um monitoramento detalhado das condições do solo, do clima, do crescimento das plantas e das pragas. Com base na interpretabilidade desses dados, os agricultores podem adquirir uma visão precisa das necessidades de suas culturas, o que lhes permite tomar decisões com o objetivo de maximizar a produtividade e reduzir o desperdício.

O trabalho de [Zhang et al. 2022] mostra a aplicação do Grad-CAM em imagens da *Spodoptera frugiperda* em uma plantação de milho. A Figura 2.6 apresenta o Grad-CAM da imagem de entrada, a aplicação da menor área retangular convexa, a localização do objeto e a remoção da área que não é de interesse. Podemos notar que o Grad-CAM nesse exemplo trás a cor vermelha para destacar a região de maior ativação do modelo, onde ele identifica que aquela região é a região mais importante para sua previsão, por fim, após o processamento notamos que realmente aquela região é onde apresenta a *Spodoptera frugiperda* na imagem.

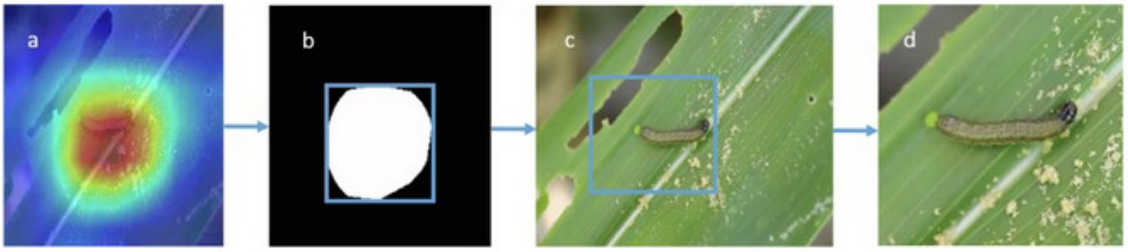

Figura 2.6: IA na agricultura. Ilustração do uso de Grad-CAM na detecção de *Spodoptera frugiperda*. Fonte: [Zhang et al. 2022]. Em a), temos a imagem gerada pelo Grad-CAM, destacando em vermelho a região mais importante para a previsão do modelo. Em b), a imagem segmentada dessa região. Em c), apresenta a *Spodoptera frugiperda* em destaque dentro da região de ativação feita pelo Grad-CAM e por fim, em d), temos a *Spodoptera frugiperda* segmentada e aproximada, mostrando que o modelo realmente é capaz de identificar as *Spodoptera frugiperda* em imagens.

A interpretabilidade da IA na agricultura é essencial, com ela é possivel que os agricultores compreendam de maneira clara e transparente como as recomendações da IA

são geradas. Essa compreensão é de importância vital, pois possibilita que os agricultores ajam de forma confiante em relação aos recursos agrícolas, resultando em uma gestão mais eficaz das operações. Em suma, a interpretabilidade da IA é uma peça fundamental que promove uma agricultura mais eficiente e sustentável, ao capacitar os agricultores a tirar o máximo proveito das vantagens oferecidas por essa tecnologia inovadora.

2.6. Desafios e Limitações

Embora existam diversas técnicas para a explicabilidade e modelos naturalmente interpretáveis, ainda existem alguns problemas relacionados ao entendimento de seus resultados, e que precisam ser levados em consideração antes de sua utilização. Neste sentido, esta seção descreve alguns pontos importantes a respeito desse contexto, tendo em vista a importância de estar ciente dessas limitações antes de utilizar os métodos abordados, pois elas podem afetar diretamente a qualidade e confiabilidade dos resultados.

No estudo conduzido por[Arrieta et al. 2020], os autores apresentam um quadro geral acerca do nível de explicabilidade de alguns modelos naturalmente interpretáveis. Neste sentido, com relação aos modelos de regressão linear/logística, é explicado que, apesar de serem legíveis por humanos e poderem ter sua complexidade reduzida para torná-los mais interpretáveis, suas variáveis e interações podem ser muito complexas para serem analisadas sem ferramentas matemáticas. Além disso, o número de interações entre variáveis independentes pode ser tão grande que seria necessário uma decomposição do modelo em partes menores para ser entendido.

Os mesmos autores ainda explanam acerca das árvores de decisão, onde eles as descrevem como modelos que podem ser facilmente simulados por humanos e oferecem uma explicação clara de como os dados estão sendo interpretados, permitindo um entendimento direto do processo de predição. No entanto, é importante enfatizar que a estrutura desse modelo pode conter muitos nós e regras, dificultando o processo de rastreio de cada decisão, afetando assim interpretação do modelo, até mesmo para especialistas. Por outro lado, as Máquinas de Vetores de Suporte (SVMs) geralmente necessitam que o modelo seja simplificado ou que técnicas de explicação locais sejam aplicadas [Arrieta et al. 2020]. Ainda assim, SVMs com kernel linear, não fornecem uma explicação direta e clara acerca de como os dados são relacionados às variáveis dependentes. Sendo assim, elas somente usam uma reta como limite para classificar os dados com base em um hiperplano de decisão, dado a linearidade de sua natureza.

Os *Heatmaps*, por sua vez, se baseiam em uma estimativa da importância que cada característica, que alimenta o modelo, tem para a composição de sua resposta. Neste sentido, por serem estimativas, tais cálculos podem ser imprecisos devido a fatores relativos a base de dados e ao modelo em si, como por exemplo a complexidade do modelo ou a presença de ruído nos dados de entrada. Outro ponto importante a se considerar é a potencial subjetividade presente na interpretação dos mapas de calor, ou seja, do ponto de vista técnico, a interpretação pode variar dependendo da ótica de quem esteja interpretando os dados exibidos nele.

O *Grad-CAM* consiste em realizar uma análise dos gradientes das ativações das camadas convolucionais da rede neural, com o objetivo de identificar as áreas de maior influência na decisão do modelo. No entanto, é importante reconhecer que esses gradien-

tes, embora relevantes, também estão sujeitos a restrições, como, resolução da imagem, arquitetura da rede neural, camadas convolucionais disponíveis, limitações de interpretação, dependência da qualidade do modelo e dependência da qualidade do modelo. Essas falhas podem surgir devido a características ambíguas nas ativações das camadas convolucionais, a sobreposição de características relevantes e irrelevantes nas imagens e a complexidade do próprio modelo.

O *LIME* de forma semelhante, fornece interpretações compreensíveis para as decisões de modelos de aprendizado de máquina em um nível local. Entretanto, é fundamental destacar os desafios do uso do modelo. Por exemplo, a eficácia do LIME pode depender da escolha adequada de hiperparâmetros e da seleção de instâncias de dados representativas para a explicação. Além disso, as interpretações geradas pelo LIME são aproximações e podem não capturar totalmente o comportamento do modelo original em todas as situações, essa característica pode gerar interpretações ambíguas em alguns cenários, como em modelos altamente não lineares, regiões de decisão complexas, conjuntos de dados desbalanceados, instabilidades nos modelos, modelos de alta dimensionalidade e ruído nos dados de entrada.

Levando em consideração os pontos ressaltados anteriormente, ao utilizar qualquer técnica de interpretabilidade ou modelo naturalmente interpretável, é importante estar ciente das limitações e desafios dessas abordagens. Uma compreensão aprofundada desses elementos possibilita uma avaliação mais precisa da confiabilidade das interpretações geradas. Portanto, a consideração cuidadosa sobre essas limitações desempenha um papel essencial na garantia de que as interpretações resultantes sejam não apenas relevantes, mas também confiáveis para contribuir na compreensão e na tomada de decisões fundamentadas em modelos de aprendizado de máquina.

2.7. Conclusões

Em conclusão, a interpretabilidade e a explicabilidade desempenham um papel fundamental no campo da Inteligência Artificial (IA). À medida que a IA se torna cada vez mais integrada em diversas áreas, desde a saúde até o marketing e a agricultura, compreender como os modelos de IA tomam decisões se torna crucial. Este entendimento não apenas promove a confiança dos usuários, mas também ajuda a evitar resultados indesejados e antiéticos.

No entanto, a busca pela interpretabilidade não é isenta de desafios e limitações. Modelos complexos, como redes neurais profundas, podem ser difíceis de interpretar, mesmo com técnicas de explicabilidade. Além disso, as interpretações geradas por essas técnicas podem ser aproximadas e sujeitas a imprecisões, especialmente em situações de alta complexidade ou com dados ruidosos.

Apesar dessas limitações, é essencial adotar uma abordagem abrangente para avaliar a interpretabilidade da IA em contextos específicos. Métricas quantitativas, avaliação qualitativa, comparação com modelos de referência e o *feedback* dos usuários desempenham um papel importante na avaliação da eficácia das técnicas de explicabilidade.

A interpretabilidade da IA continuará a ser uma área de pesquisa e desenvolvimento importante, à medida que os modelos de IA se tornam mais complexos e integrados

em nossa sociedade. A transparência e a compreensão das decisões de IA são essenciais para construir sistemas responsáveis, éticos e confiáveis, garantindo que a IA beneficie a humanidade em diversos setores e aplicações.

À medida que a tecnologia avança, a interpretabilidade da IA continuará a evoluir, ajudando a resolver desafios complexos e a promover uma tomada de decisão informada em diversas áreas, desde a medicina até a agricultura e além. Com a conscientização das limitações e a busca constante por melhorias, a interpretabilidade da IA continuará a desempenhar um papel crucial no progresso da IA e em sua integração responsável em nossa sociedade.

Portanto, o objetivo deste capítulo foi concluído ao expor de forma conceitual e intuitiva, uma compreensão relativamente ampla do domínio da interpretabilidade e a explicabilidade de IA, orientando assim o público interessado sobre os principais conceitos, técnicas e abordagens no contexto desse capítulo, de forma que não se limitem somente a isto, mas que também procurem outros contextos e formas de se trabalhar dentro deles utilizando todo o conhecimento abordado, sendo isto, a principal contribuição deste capítulo.

Como trabalhos futuros, pretende-se realizar um levantamento do estado da arte a fim de explorar novas técnicas e metodologias para a interpretabilidade e a explicabilidade de IA. Com isso, é possível realizar melhorias neste capítulo, contribuindo para o estado da arte e realizar comparações com as técnicas apresentadas.

Agradecimentos

Este capítulo foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001. Agradecemos também ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado do Piauí (FAPEPI). Todos os códigos podem ser encontrados em https://github.com/Julio-M39/SINFO2022MINICURSO.

Referências

[Jou 1986] (1986). Induction of decision trees. volume 1, pages 81–106. Kluwer Academic Publishers.

[Agarwal and Das 2020] Agarwal, N. and Das, S. (2020). Interpretable machine learning tools: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1528–1534. IEEE.

[Ahsan et al. 2020] Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M. L., and Shakhawat Hossain, M. (2020). Covid-19 symptoms detection based on nasnetmobile with explainable ai using various imaging modalities. *Machine Learning and Knowledge Extraction*, 2(4):490–504.

[Arrieta et al. 2020] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

[Borges et al. 2022] Borges, A. L., Gonçalves, C. d. A., Dias, V. B. L., Sousa, E. A., Costa, C. H. N., and Silva, R. R. V. e. (2022). Visceral leishmaniasis detection using deep learning techniques and multiple color space bands. In *International Conference on Intelligent Systems Design and Applications*, pages 492–502. Springer.

[Caliskan et al. 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[Chattopadhay et al. 2018] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847.

[Cortes and Vapnik 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[Cox 1958] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

[Forest-gis] Forest-gis. Densidade populacional no brasil – heatmaps. Acesso em 29 set. 2023.

[Gennatas and Chen 2021] Gennatas, E. D. and Chen, J. H. (2021). Chapter 1 - artificial intelligence in medicine: past, present, and future. In Xing, L., Giger, M. L., and Min, J. K., editors, *Artificial Intelligence in Medicine*, pages 3–18. Academic Press.

[Grif and Avush 2018] Grif, M. G. and Avush, Y. (2018). The development of medical diagnostic system based on integration of traditional and eastern medicines. In *2018 XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, pages 511–515.

[Iqbal and Qureshi 2022] Iqbal, T. and Qureshi, S. (2022). The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*, 34(6, Part A):2515–2528.

[Jin 2017] Jin, Y. (2017). Development of word cloud generator software based on python. *Procedia Engineering*, 174:788–792. 13th Global Congress on Manufacturing and Management Zhengzhou, China 28-30 November, 2016.

[Lowry and Macpherson 1988] Lowry, S. and Macpherson, G. (1988). A blot on the profession. *British medical journal (Clinical research ed.)*, 296(6623):657.

[Marques et al. 2023] Marques, J. V. M., de Araújo Gonçalves, C., de Carvalho Ferreira, J. F., de Melo Souza Veras, R., de Andrade Lira Rabelo, R., and Veloso e Silva, R. R. (2023). Detection of covid-19 in computed tomography images using deep learning. In Abraham, A., Pllana, S., Casalino, G., Ma, K., and Bajaj, A., editors, *Intelligent Systems Design and Applications*, pages 143–152, Cham. Springer Nature Switzerland.

[Nivas et al. 2016] Nivas, V. M., Krishnan, P. G., and Fredrhic, A. C. (2016). Automated guided car (agc) for industrial automation. In *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, pages 1–6.

[Pedreshi et al. 2008] Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 560–568, New York, NY, USA. Association for Computing Machinery.

[Pei 2009] Pei, J. (2009). *Association Rules*, pages 140–142. Springer US, Boston, MA.

[Ribeiro et al. 2016a] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). "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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

[Ribeiro et al. 2016b] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

[Rudin 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

[Selvaraju et al. 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

[Selvaraju et al. 2019] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

[Shuguang 2011] Shuguang, W. (2011). Analysis on finance countermeasures of regional economy development in china. In *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, pages 6142–6145.

[Thommandru et al. 2023] Thommandru, A., Mone, V., Mitharwal, S., and Tilwani, R. (2023). Exploring the intersection of machine learning, money laundering, data privacy, and law. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pages 149–155.

[Volkov and Averkin 2023] Volkov, E. N. and Averkin, A. N. (2023). Possibilities of explainable artificial intelligence for glaucoma detection using the lime method as an

example. In *2023 XXVI International Conference on Soft Computing and Measurements (SCM)*, pages 130–133.

[Yan and Su 2009] Yan, X. and Su, X. G. (2009). *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Co., Inc., USA.

[Zhang et al. 2022] Zhang, H., Zhao, S., Song, Y., Ge, S., Liu, D., Yang, X., and Wu, K. (2022). A deep learning and grad-cam-based approach for accurate identification of the fall armyworm (spodoptera frugiperda) in maize fields. *Computers and Electronics in Agriculture*, 202:107440.