

Luiz Felipe de Camargo

Estratégia de Uso de Inteligência Artificial Explicável no Apoio à Reabilitação Neuromotora

Bauru, São Paulo, Brasil

26 de agosto de 2024

Luiz Felipe de Camargo

Estratégia de Uso de Inteligência Artificial Explicável no Apoio à Reabilitação Neuromotora

Tese para defesa de Doutorado sobre o Uso
de Inteligência Artificial Explicável (XAI) no
Apoio à Reabilitação Neuromotora

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Programa de Pós-Graduação em Ciência da Computação

Orientador: Dr. José Remo Ferreira Brega

Bauru, São Paulo, Brasil

26 de agosto de 2024

Este trabalho é dedicado à minha família e meus amigos, os meus maiores pontos de apoio, desde sempre.

C172e

Camargo, Luiz Felipe de

Estratégia de Uso de Inteligência Artificial Explicável no Apoio à
Reabilitação Neuromotora / Luiz Felipe de Camargo. -- Bauru, 2024
133 p. : il., tabs., fotos

Tese (doutorado) - Universidade Estadual Paulista (UNESP),
Faculdade de Ciências, Bauru

Orientador: José Remo Ferreira Brega

1. Inteligência Artificial. 2. Inteligência Artificial Explicável. 3.
Uso. 4. Classificação de Movimentos. 5. Reabilitação Neuromotora. I.
Título.

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

Essa ficha não pode ser modificada.

ATA DA DEFESA PÚBLICA DA TESE DE DOUTORADO DE LUIZ FELIPE DE CAMARGO, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, DA FACULDADE DE CIÊNCIAS - CÂMPUS DE BAURU.

Aos 09 dias do mês de agosto do ano de 2024, às 14:00 horas, por meio de Videoconferência, realizou-se a defesa de TESE DE DOUTORADO de LUIZ FELIPE DE CAMARGO, intitulada **Estratégia de Uso de Inteligência Artificial Explicável no Apoio a Reabilitação Neuromotora**. A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. JOSE REMO FERREIRA BREGA (Orientador(a) - Participação Virtual) do(a) Departamento de Computação / UNESPCâmpus de Bauru, Prof. Dr. DANIEL CARLOS GUIMARÃES PEDRONETTE (Participação Virtual) do(a) IGCE / UNESP / Rio Claro (SP), Prof. Dr. KELTON AUGUSTO PONTARA DA COSTA (Participação Virtual) do(a) Departamento de Computação / UNESPCâmpus de Bauru, Prof. Dr. EDGARD AFONSO LAMOUNIER JUNIOR (Participação Virtual) do(a) Universidade Federal de Uberlândia, Prof. Dr. MARCELO DE PAIVA GUIMARÃES (Participação Virtual) do(a) Departamento de Realidade Virtual / Universidade Federal de São Paulo. Após a exposição pelo doutorando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: **Aprovado**. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Prof. Dr. JOSE REMO FERREIRA BREGA

Agradecimentos

Expresso aqui meus agradecimentos ao meu orientador Prof. José Remo Ferreira Brega, pelas produtivas conversas, sempre repletas de conselhos, ideias e de amizade, agradeço ao Luis Guilherme Silva Rodrigues, pela dedicação, pelas sugestões sempre certeiras e por todo auxílio, agradeço à equipe da DTI-FMB onde presto serviço, pela paciência, pelos conselhos e sugestões que agregaram bastante qualidade a este trabalho, agradeço aos colegas de doutorado, pelo compartilhamento de conhecimento, pela colaboração e pelo crescimento conjunto, agradeço aos professores do programa e do Departamento de Computação da Faculdade de Ciências em Bauru por toda a colaboração na minha formação. Agradeço a todos os colegas do grupo de pesquisa em classificação de movimentos, professores e alunos, que durante todo o processo sempre compartilharam boas opiniões, agradeço a todos os colegas do Laboratório de Interface e Visualização (LIV), orientados do Prof. Remo, que me ajudaram a entender cada vez mais o processo de desenvolvimento de um trabalho científico na área de Ciência da Computação. Agradeço a minha família e demais amigos por serem sempre meu suporte. E, por fim, acima de tudo, agradeço a Deus, em especial pelo dom da minha vida e por nunca me abandonar.

*"Era preciso que o senhor naufragasse
nos próprios fracassos para aceitar o colete
salva-vidas que ele lhe lançou.- Eugen Herrigel,
A Arte Cavalheiresca do Arqueiro Zen*

Resumo

Cada vez mais soluções baseadas em Inteligência Artificial são utilizadas para resolver problemas em diversos domínios. Entretanto, muitas pessoas demonstram não se sentirem a vontade com esse tipo de solução por não compreender seu funcionamento. Diante disso surge a chamada Inteligência Artificial Explicável (XAI), buscando além de fornecer as respostas produzidas pela Inteligência Artificial, oferecer também aspectos de explicabilidade, detalhando o processo de decisão e gerando confiança. A área de saúde é promissora para aplicação de Inteligência Artificial Explicável, apresentando a possibilidade de uso na reabilitação de pacientes que sofre de acidente vascular cerebral e outras doenças que afetam as capacidades neuromotoras. Nesse contexto é apresentada uma revisão de literatura sobre o uso de Inteligência Artificial Explicável na Reabilitação Neuromotora, um breve estudo comparativo entre as bibliotecas mais populares para esse uso e um aprofundamento na temática de avaliação de explicabilidade e do processo de compreensão. A partir dos resultados obtidos, foi construída uma estratégia de uso de XAI para apoio à reabilitação neuromotora, com foco na classificação de movimentos. Como parte da estratégia de uso foi desenvolvida a ferramenta MoveXAI, que integra a visualização do movimento executado em um avatar com as informações de explicabilidade. A ferramenta se encontra em testes e tem demonstrado resultados relevantes segundo *feedback* de colaborador pesquisador da área de reabilitação.

Palavras-chave: Inteligência Artificial. Inteligência Artificial Explicável. Uso. Classificação de Movimentos. Reabilitação Neuromotora.

Abstract

More and more solutions based on Artificial Intelligence are being used to solve problems in various fields. However, many people feel uncomfortable with this type of solution because they do not understand how it works. As a result, the so-called Explainable Artificial Intelligence (XAI) has emerged, seeking not only to provide the answers produced by Artificial Intelligence, but also to offer aspects of explainability, detailing the decision-making process and generating trust. The area of health is promising for the application of Explainable Artificial Intelligence, with the possibility of using it in the rehabilitation of patients suffering from strokes and other diseases that affect neuromotor abilities. In this context, a literature review is presented on the use of Explainable Artificial Intelligence in Neuromotor Rehabilitation, a brief comparative study between the most popular libraries for this use and an in-depth look at the topic of evaluating explainability and the comprehension process. Based on the results obtained, a strategy was built for using XAI to support neuromotor rehabilitation, with a focus on classifying movements. As part of the use strategy, the MoveXAI tool was developed, which integrates the visualization of the movement performed on an avatar with the explainability information. The tool is currently being tested and has shown relevant results, according to feedback from a researcher in the rehabilitation area.

Keywords: Artificial Intelligence. Explainable Artificial Intelligence. Usage. Movement Classification. Neuromotor Rehabilitation.

Lista de ilustrações

Figura 1 – Os cenários com a IA tradicional e XAI conforme definição da DARPA.	21
Figura 2 – Quadrinho tratando a temática de IA de forma bem humorada, com foco na falta de compreensão dos mecanismos de IA.	23
Figura 3 – Diagrama mostrando as vertentes de interpretabilidade de acordo com diferentes públicos.	32
Figura 4 – Diagrama mostrando as principais abordagens de XAI.	35
Figura 5 – Diagrama mostrando a criação do conceito de IA Responsável através de diversos aspectos.	36
Figura 6 – Diagrama mostrando as etapas da revisão sistemática de literatura.	45
Figura 7 – Gráfico mostrando a evolução da quantidade de estudos durante os anos.	46
Figura 8 – Diagrama demonstrando o fluxo de dados para realização de uma explicação através da biblioteca LIME.	56
Figura 9 – Biblioteca de XAI LIME sendo utilizada para explicar classificação de texto nas classes "ateísmo" ou "cristianismo". Palavras destacadas em azul colaboram com a classificação como "ateísmo", palavras destacadas em laranja colaboram com a classificação como "cristianismo".	56
Figura 10 – Biblioteca de XAI LIME sendo utilizada para explicar classificação de imagem de gato, indicando áreas que colaboraram positiva (área verde) e negativamente (área vermelha) para classificação.	57
Figura 11 – Biblioteca de XAI SHAP sendo utilizada para explicar classificação da opinião sobre um filme como positiva ou negativa, indicando trechos que colaboraram positiva (cor rosa) e negativamente (cor azul) para classificação.	58
Figura 12 – Diagrama demonstrando o fluxo de dados para realização de uma explicação através da biblioteca SHAP.	59
Figura 13 – Exemplo de uso da biblioteca ELI5, indicando a área mais relevante para a classificação através de um mapa de calor sobreposto a imagem original.	60
Figura 14 – Exemplo de uso da biblioteca AI Explainability 360, gerando uma explicação textual, obtido executando a demonstração da ferramenta disponível na página do projeto.	61
Figura 15 – Exemplo de uso da biblioteca Skater, indicando as áreas da imagem que levaram a classificação da imagem como "panda".	62
Figura 16 – Exemplo de uso da biblioteca Captum, com 2 exemplos, indicando a classificação das imagens como "gato" e "navio", após o devido tratamento das imagens.	63

Figura 17 – Exemplo de uso da biblioteca TruLens, aplicando uma máscara sobre a imagem original, indicando as áreas mais relevantes para a classificação através de janelas nessa máscara.	64
Figura 18 – Exemplo de uso da biblioteca VisualDL, demonstrando um <i>dashboard</i> , com informações sobre modelo e seus parâmetros.	64
Figura 19 – Exemplo de uso da ferramenta Netron, exibindo a estrutura da rede neural.	65
Figura 20 – Exemplo de uso da ferramenta Zetane Viewer, exibindo a estrutura da rede neural através de uma visualização 3D, permitindo a análise de cada camada.	66
Figura 21 – Resultados dos experimentos realizados no Exemplo 1, utilizando a biblioteca LIME em um modelo de identificação de caracteres.	69
Figura 22 – Resultados dos experimentos realizados no Exemplo 2, utilizando a biblioteca LIME em um modelo de identificação de imagens.	70
Figura 23 – Resultados dos experimentos realizados no Exemplo 3, utilizando a biblioteca SHAP em um modelo de identificação de caracteres.	71
Figura 24 – Resultados dos experimentos realizados no Exemplo 4, utilizando a biblioteca SHAP em um modelo de previsão que utiliza dados tabulares.	72
Figura 25 – Diagrama mostrando um modelo conceitual do processo de explicação, no contexto do XAI.	75
Figura 26 – Exemplo de teste do modelo de estimativa de pose em 3D, com a exibição de um avatar executando os movimentos detectados.	80
Figura 27 – Exemplo de teste do modelo de estimativa de pose em 3D, com a exibição de um avatar executando os movimentos detectados.	80
Figura 28 – Exemplo de teste do modelo de avaliação de ângulo, com a exibição de um avatar sobreposto ao vídeo original e a estimativa do ângulo de flexão do tronco.	81
Figura 29 – Exemplo de teste do modelo de avaliação de ângulo, com a exibição de um avatar sobreposto ao vídeo original e a estimativa do ângulo de flexão do tronco.	82
Figura 30 – Diagrama mostrando a arquitetura de um dos modelo de rede neural utilizados nos estudos de (RODRIGUES et al., 2021a).	85
Figura 31 – Diagramas mostrando o fluxo de dados para classificação de movimentos com a ferramenta existente e proposta inicial de utilização de XAI.	86
Figura 32 – Captura de tela mostrando a aplicação do rastreamento de pose em novo vídeo produzido para o estudo atual.	87

Figura 33 – Captura de tela mostrando os resultados dos testes executados com a biblioteca LIME, com o impacto de cada articulação em cada eixo e em cada quadro do vídeo para a classificação correta do movimento em uma amostra.	87
Figura 34 – Captura de tela mostrando a manipulação dos dados obtidos com a biblioteca LIME, agregando os diferentes valores relacionados a mesma articulação.	88
Figura 35 – Captura de tela mostrando os resultados dos testes executados com a biblioteca SHAP, com o impacto de cada articulação em cada eixo no tempo para a classificação correta do movimento de forma global. . . .	88
Figura 36 – Captura de tela mostrando que ao aplicar a biblioteca LIME no modelo de IA, os dados eram retornados com indicação de <i>frame</i> e não forneciam uma explicação consistente.	91
Figura 37 – Captura de tela mostrando que ao aplicar a biblioteca SHAP no modelo de IA, cada <i>frame</i> era tratado como um elemento a ser explicado, gerando uma visualização sobreposta e confusa.	91
Figura 38 – Diagrama mostrando a análise de uma amostra considerando recurso, evento e célula no qual se baseia a biblioteca TimeSHAP.	92
Figura 39 – Gráficos diversos fornecidos pelo TimeSHAP obtidos executando um exemplo disponível no repositório.	93
Figura 40 – Imagem mostrando um estudo sobre a correlação dos valores das articulações, os eixos mostram as articulações, cada célula indica a correlação de dois valores de articulações, quanto mais escura a cor, mais correlacionados estão os valores.	94
Figura 41 – Captura de tela mostrando o protótipo de <i>player</i> compatível com dados no formato CSV que já existia.	95
Figura 42 – Imagem mostrando a captura de tela durante execução de <i>script</i> para atualização da aplicação no contêiner em execução no servidor.	96
Figura 43 – Escala de cores escolhida para exibição dos dados de explicabilidade no <i>player</i> , variando as cores nas articulações, onde vermelho indica contribuição positiva e azul negativa.	96
Figura 44 – Imagem mostrando a execução dos exercícios para produção do conjunto de dados UI-PRMD.	97
Figura 45 – Captura de tela mostrando a interface inicial do MoveXAI com o campo de seleção de conjunto de dados.	99
Figura 46 – Diagrama mostrando o fluxo de dados da ferramenta MoveXAI.	100
Figura 47 – Captura de tela mostrando a parte superior da interface após seleção de uma amostra.	101

Figura 48 – Captura de tela mostrando a parte superior da interface após seleção do modo de <i>upload</i>	102
Figura 49 – Captura de tela mostrando a interface do MoveXAI na aba " <i>Player</i> ", com a articulação do ombro destacada em vermelho, indicando o impacto positivo dela na classificação e a barra indicando a precisão da classificação no <i>frame</i> indicado.	103
Figura 50 – Captura de tela mostrando a interface do MoveXAI na aba "Linha do Tempo", onde o gráfico de <i>pruning</i> da Biblioteca XAI é utilizado para demonstrar o crescimento da precisão do modelo no decorrer dos <i>frames</i>	104
Figura 51 – Captura de tela mostrando a interface do MoveXAI na aba "Frames", onde o gráfico de eventos da Biblioteca XAI é utilizado para demonstrar o impacto de cada <i>frame</i> na classificação. Os <i>frames</i> são indicado na lateral esquerda, com os valores de Shapley calculados, onde quanto mais intensa a cor, maior o impacto do <i>frame</i>	105
Figura 52 – Captura de tela mostrando a interface do MoveXAI na aba "Articulações", onde o gráfico de características da Biblioteca XAI é utilizado para demonstrar o impacto de cada articulação na classificação. As articulações são mostradas na lateral esquerda e os valores de impacto são exibidos através de um gráfico de barras.	106
Figura 53 – Captura de tela mostrando a interface do MoveXAI na aba "Células", onde o gráfico de células da Biblioteca XAI é utilizado para demonstrar o impacto das principais articulações e <i>frames</i> na classificação. É apresentada uma matriz com o eixo horizontal mostrando as articulações e o eixo vertical os <i>frames</i> , com a cor variando de acordo com o impacto da célula (articulação/ <i>frame</i>) na classificação.	107
Figura 54 – Captura de tela mostrando a interface do MoveXAI na aba "Correlação", onde um gráfico demonstra a correlação da articulações. Os eixos mostram as articulações, cada célula indica a correlação de dois valores de articulações, quanto mais escura a cor, mais correlacionados estão os valores.	108

Lista de tabelas

Tabela 1 – Público-alvo em XAI	31
Tabela 2 – Questões que compuseram o formulário de extração.	43
Tabela 3 – Categorias de objetivos dos estudos.	46
Tabela 4 – Presença da expressão XAI.	47
Tabela 5 – Uso das bibliotecas.	47
Tabela 6 – Bibliotecas citadas.	48
Tabela 7 – Estudos separados por momento de geração da explicabilidade.	48
Tabela 8 – Métodos de visualização citados.	49
Tabela 9 – Informações obtidas sobre as ferramentas de XAI analisadas.	67
Tabela 10 – Estudos incluídos na revisão.	128

Lista de abreviaturas e siglas

AAAI	Associação para Avanços da Inteligência Artificial (do inglês, <i>Association for the Advancement of Artificial Intelligence</i>)
ACM	<i>Association of Computing Machinery</i>
ALE	<i>Accumulated local effects</i>
AVC	Acidente Vascular Cerebral
CEPID BRAINN	Instituto Brasileiro de Neurociências e Neurotecnologia
DARPA	Agência de Projetos de Pesquisa Avançada de Defesa (do inglês, <i>Defense Advanced Research Projects Agency</i>)
EUA	Estados Unidos da América
GDPR	Regulamento Geral de Proteção de Dados (do inglês, <i>General Data Protection Regulation</i>)
IA	Inteligência Artificial
IBM	<i>International Business Machines Corporation</i>
IEEE	<i>Institute of Electrical and Eletronics</i>
ISI	<i>Institute for Scientific Information</i>
LightGBM	<i>Light Gradient Boosting Machine</i>
LIME	<i>Local Interpretable Model-agnostic Explanations</i>
MIT	<i>Massachusetts Institute of Technology</i>
ML	Aprendizado de Máquina (do inglês, <i>Machine Learning</i>)
ONU	Organização das Nações Unidas
PFI	<i>Permutation Feature Importance</i>
ReLU	Unidade Linear Retificada (do inglês, <i>Rectified Linear Unit</i>)
RFI	<i>Random forest feature importance</i>
RSL	Revisão Sistemática da Literatura

SGD	Gradiente Descendente Estocástico (do inglês, <i>Stochastic Gradient Descent</i>)
SHAP	<i>SHapley Additive exPlanations</i>
UI-PRMD	<i>University of Idaho – Physical Rehabilitation Movement Dataset</i>
XAI	Inteligência Artificial Explicável (do inglês, <i>Explainable Artificial Intelligence</i>)

Sumário

1	INTRODUÇÃO	20
1.1	Inteligência Artificial e explicabilidade	20
1.2	Justificativas	23
1.3	Hipótese	24
1.4	Objetivos	25
1.5	Considerações finais do capítulo	26
1.6	Estrutura da tese	26
2	FUNDAMENTAÇÃO TEÓRICA	27
2.1	Considerações iniciais do capítulo	27
2.2	Inteligência Artificial	27
2.3	Inteligência Artificial Explicável	28
2.3.1	Abordagens conceituais	29
2.3.2	Termos e definições	30
2.3.3	Público alvo	31
2.3.4	Métricas	33
2.3.5	Necessidade de explicações	33
2.3.6	Abordagens de Inteligência Artificial Explicável	34
2.4	IA e XAI em saúde	36
2.5	Acidente Vascular Cerebral e Reabilitação	37
2.6	Considerações finais do capítulo	38
3	REVISÃO SISTEMÁTICA DE LITERATURA	39
3.1	Considerações iniciais do capítulo	39
3.2	Metodologia	39
3.2.1	Classificação dos estudos	39
3.2.2	Objetivos da revisão	40
3.2.3	Questões de pesquisa	40
3.3	Desenvolvimento do protocolo	40
3.3.1	Estratégia de busca e bases de dados	40
3.3.2	Critérios de inclusão e exclusão	41
3.3.3	Extração de dados	42
3.3.4	Ferramenta Parsifal	43
3.4	Resultados	44
3.4.1	Condução	44
3.4.2	Estudos excluídos e incluídos	45

3.4.3	Anos dos estudos	45
3.4.4	Objetivos dos estudos	46
3.4.5	Uso da expressão XAI	47
3.4.6	Bibliotecas de apoio	47
3.4.7	Momento de geração da explicação	48
3.4.8	Fontes de dados	48
3.4.9	Visualização da informação	49
3.4.10	Avaliação ou validação com usuários	49
3.4.11	Estudos secundários	49
3.5	Discussão	50
3.5.1	Tendências	50
3.5.2	Descrição dos estudos	51
3.5.3	Relevância do estudo	52
3.5.4	Limitações do estudo	52
3.6	Considerações finais do capítulo	53
4	FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL	54
4.1	Considerações iniciais do capítulo	54
4.2	<i>Frameworks</i>, bibliotecas e métodos	54
4.3	LIME	54
4.4	SHAP	57
4.5	Outras ferramentas	59
4.5.1	ELI5	59
4.5.2	AI Explainability 360	60
4.5.3	Skater	61
4.5.4	Captum	62
4.5.5	TruLens	63
4.5.6	VisualDL	64
4.5.7	Netron	65
4.5.8	Zetane Viewer	65
4.6	Comparativo	66
4.7	Exemplos de uso	68
4.7.1	Exemplo 1 - LIME	68
4.7.2	Exemplo 2 - LIME	69
4.7.3	Exemplo 3 - SHAP	70
4.7.4	Exemplo 4 - SHAP	71
4.7.5	Considerações finais do capítulo	73
5	AVALIAÇÃO DE MÉTODOS DE INTELIGÊNCIA ARTIFICIAL EX- PLICÁVEL	74

5.1	Considerações iniciais do capítulo	74
5.2	O processo de entendimento	74
5.3	A curiosidade	76
5.4	Medidas de avaliação de Inteligência Artificial Explicável	76
5.4.1	Modelo mental	76
5.4.2	Explicação de utilidade e satisfação	76
5.4.3	Confiança e confiabilidade do usuário	77
5.4.4	Desempenho de tarefa Humano-Inteligência Artificial	77
5.4.5	Medidas computacionais	77
5.5	Como avaliar	77
5.6	Considerações finais do capítulo	78
6	ESTRATÉGIA DE USO	79
6.1	Considerações iniciais do capítulo	79
6.2	Estudos preliminares	79
6.3	Cenário	82
6.4	Classificação de movimentos com IA	83
6.5	Estratégia	85
6.6	Considerações finais do capítulo	89
7	DESENVOLVIMENTO	90
7.1	Considerações iniciais do capítulo	90
7.2	Soluções inadequadas	90
7.3	TimeSHAP	92
7.4	Interface	94
7.5	Conjunto de dados UI-PRMD	97
7.6	Pesquisador da área de reabilitação	98
7.7	Considerações finais do capítulo	99
8	RESULTADOS	100
8.1	Considerações iniciais do capítulo	100
8.2	A ferramenta	101
8.3	Infraestrutura computacional	109
8.4	A área de explicabilidade	109
8.5	Considerações finais do capítulo	110
9	CONCLUSÃO	111
9.1	Considerações iniciais do capítulo	111
9.2	Relevância	111
9.3	Limitações	112
9.4	Trabalhos futuros	112

9.5	Considerações finais	113
9.6	Publicações no período do doutorado	113
	REFERÊNCIAS	115
	APÊNDICES	126
	APÊNDICE A – DADOS DOS ESTUDOS INCLUÍDOS	127
	APÊNDICE B – ESTUDOS EXCLUÍDOS	132

1 Introdução

1.1 Inteligência Artificial e explicabilidade

Cada vez mais pode-se observar o uso massivo da Inteligência Artificial (IA) nas mais diversas áreas da tecnologia. Todo celular moderno conta com um sistema de assistente pessoal para auxiliar na rotina diária, assistente esse que toma suas decisões com auxílio da IA. Um avião atualmente consegue passar grande parte de seu tempo de voo sendo controlado pelo piloto automático, não havendo necessidade da interferência humana. A IA é capaz de realizar os mais diversos tipos de atividades, quase sempre de forma superior e mais confiável que os seres humanos ([ABUSELIDZE; MAMALADZE, 2021](#)) ([GILLATH et al., 2021](#)).

Nos últimos anos pode ser observado ainda o advento da Inteligência Artificial Generativa, com capacidade de produção de textos e imagens com qualidades impressionantes, emulando a inteligência humana e servindo de assistentes na produção de conteúdos antes não possíveis para não especialistas ([HESSEL; LEMES, 2024](#)).

Com a adoção de IA existe a expectativa, por exemplo, que haverá uma redução no número de vagas de emprego. Porém, isso não necessariamente deve acontecer, mas sim uma mudança na natureza do trabalho a ser executado, passando muitas vezes de trabalhos penosos, com grande desgaste físico, para trabalhos de controle de equipamentos, baseados em esforço intelectual e demandando capacitação prévia maior ([ABUSELIDZE; MAMALADZE, 2021](#)). Entretanto, a disseminação de informações incorretas sobre a utilização de IA pode prejudicar a adoção deste tipo de tecnologia ([CARVALHO, 2021](#)).

O uso comercial da IA em sistemas de informação permite oferecer aos usuários uma experiência superior, customizada a suas necessidades, otimizando negócios e acelerando processos. O processo de utilização da IA quase sempre consiste na introdução de dados em um modelo computacional de forma a treiná-lo para solucionar um determinado problema. Este modelo aprende os padrões destes dados de treinamentos e se torna capaz de identificá-los e prevê-los em outras massas de dados, identificando situações e fazendo recomendações. Porém, a estrutura interna do modelo, que leva ao aprendizado e toma as decisões, quase sempre é complexa e de difícil compreensão, principalmente para o usuário médio ([SARKAR, 2019](#)).

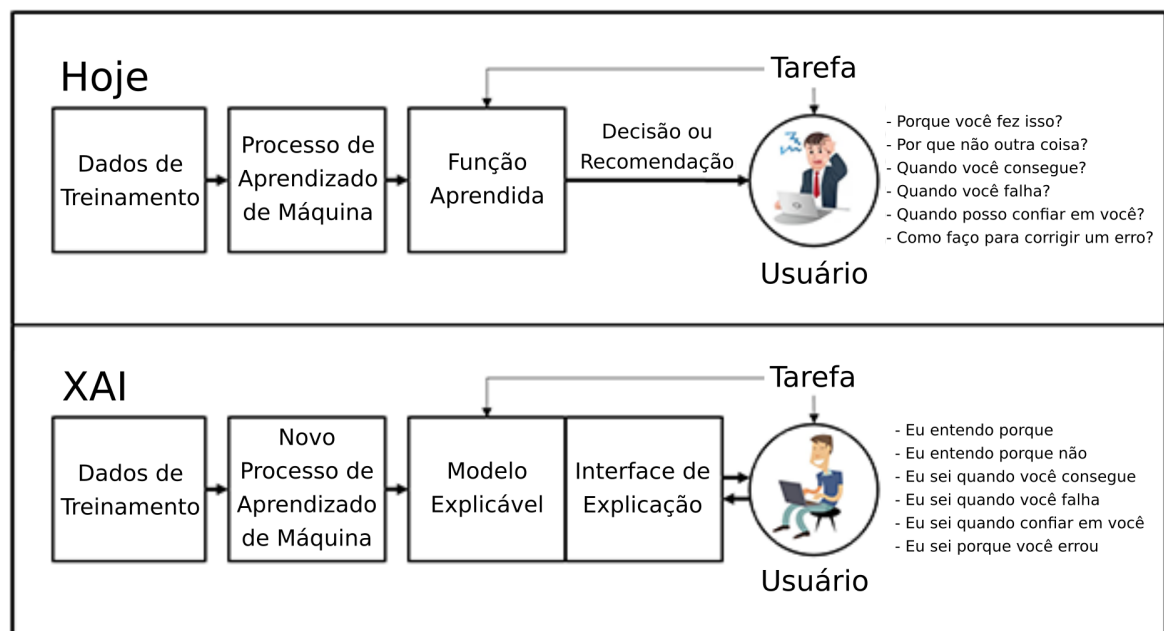
Essa complexidade, somada da falta de transparência, gera grande receio e desconforto para os usuários, ainda que especialistas em sua área do conhecimento, que tem a possibilidade de utilizar ferramentas com IA, criando barreiras para a adoção deste tipo de tecnologia ([HAIBE-KAINS et al., 2020](#)). Situações com risco de morte, como o uso de

IA para diagnóstico médico, tendem a gerar ainda mais receio. A IA pode tomar decisões superiores muitas vezes ao ser humano, mas normalmente não pode explicar e justificar sua decisão (DECAMP; LINDVALL, 2020).

Neste cenário, surge o conceito de Inteligência Artificial Explicável (do inglês *eXplainable Artificial Intelligence* (XAI), buscando tornar a IA mais acessível, produzindo, além do resultado esperado, mecanismos que auxiliam a compreensão dos resultados e do caminho percorrido até eles. A primeira iniciativa para desenvolvimento de XAI partiu da Agência de Projetos de Pesquisa Avançada de Defesa, em inglês *Defense Advanced Research Projects Agency* (DARPA), agência governamental de pesquisa americana focada em projetos de defesa, onde foram desenvolvidas também as bases da Internet. Entretanto, diversas outras iniciativas explorando XAI vêm surgindo nos últimos anos, inclusive, provenientes de gigantes como o Google (GUNNING et al., 2021) (SARKAR, 2019).

O XAI busca oferecer uma interface de explicação para o funcionamento da IA, desmistificando seu funcionamento e aumentando o grau de confiança do usuário, de forma a aumentar assim a aceitabilidade deste tipo de ferramenta. Segundo os autores do livro *Creative Confidence*, com cada inovação vem a necessidade de torná-la factível, viável e desejável (KELLEY; KELLEY, 2013). Na Figura 1 pode-se visualizar o cenário atual da IA e o proposto por meio da XAI.

Figura 1 – Os cenários com a IA tradicional e XAI conforme definição da DARPA.



Fonte - Turek (2018) - Traduzido pelo autor.

Cada vez mais veem sendo utilizados modelos chamados profundos, que possuem uma grande quantidade de parâmetros, divididos muitas vezes em diversas camadas. Esses modelos apresentam bons resultados, mas devido a sua complexidade, quase sempre

são de difícil compreensão e, portanto, opacos. Buscar compreender como os diferentes componentes dos dados de treinamento impactam nos resultados é um ponto crucial da XAI, que auxilia no desenvolvimento de uma IA mais compreensível e assim mais justa e ética. Ainda que se baseiem puramente em técnicas matemáticas, as ferramentas de IA utilizam dados do passado em seu processo de aprendizado, podendo assim perdurar vieses e preconceitos presentes nos dados de treinamento. A IA pode ser considerada o futuro da tecnologia, mas é um futuro que aprende com o passado, que pode não ser totalmente desejável (MEHRABI et al., 2021).

Quando se trata de explicabilidade, um fator fundamental é a própria audiência, o público-alvo tem papel de protagonista, visto que pode ser afetado pelas decisões dos modelos, dessa forma é fundamental que entendam a situação, buscando garantir que as decisões sejam tomadas de forma justa. Entretanto, é complexo quantificar o entendimento, a melhora na compreensão, sem adentrar campos relacionados a Psicologia e ao comportamento humano.

Por utilizar métodos complexos, fazendo uso de técnicas matemáticas e de computação avançadas, muitas vezes a população leiga e até mesmo especialistas não tem total conhecimento sobre o funcionamento dos métodos de IA, esse assunto é abordado com humor no quadrinho de [xkcd](#) (2017) que pode ser visto traduzido na Figura 2.

Figura 2 – Quadrinho tratando a temática de IA de forma bem humorada, com foco na falta de compreensão dos mecanismos de IA.



Fonte - [xkcd](#) (2017)

No documentário *Coded Bias*, a pesquisadora do *Massachusetts Institute of Technology* (MIT) Joy Buolamwini trata das falhas no treinamento de algoritmos de reconhecimento facial baseados em IA que possuem baixa precisão e acabam perpetuando preconceitos e gerando discriminação, principalmente por conta do treinamento com conjuntos dados com vieses. ([KANTAYYA, 2020](#)). Situações como essa levam cada vez mais cidades, principalmente nos EUA, a banir tecnologias de reconhecimento facial pelo poder público ([POLLO, 2020](#)).

A preocupação com o impacto das tecnologias de IA na sociedade é tão grande que a Alta Comissária para Direitos Humanos da Organização das Nações Unidas (ONU) Michelle Bachelet, juntamente com seu escritório, produziu um relatório demandando ações urgentes para evitar riscos a privacidade provenientes da IA ([RIGHTS, 2021](#)).

1.2 Justificativas

Com a evolução cada vez maior da IA, a necessidade de confiança nos dados vêm crescendo. A lei europeia *General Data Protection Regulation* (GDPR) possui um trecho onde torna obrigatória a explicação sobre o funcionamento dos mecanismos de

tratamento de dados (VALENTE, 2019). De acordo com a empresa de consultoria Gartner, estimou-se em 2017 que até 2020, cerca de 85% dos gerentes de informação teriam testado programas de IA, buscando comprar, criar ou contratar uma ferramenta deste tipo. Com esse crescimento, cabe ao ser humano encontrar mecanismos para garantir a confiabilidade (GOASDUFF, 2017).

A presença da participação humana nas análises de dados, com auxílio de mecanismos como XAI, permite que os usuários entendam, confiem e gerenciem adequadamente os resultados, obtendo *feedback*, mesmo com pouco conhecimento de domínio, alcançando assim modelos mais precisos (CHEN et al., 2016).

Um fato que mostra a relevância da temática foi a premiação da professora Cynthia Rudin, da Universidade de Duke, com o prêmio Squirrel AI de Inteligência Artificial para o Benefício da Humanidade da Associação para Desenvolvimento da Inteligência Artificial (AAAI), após 15 anos de pesquisas sobre algoritmos de aprendizado de máquina interpretáveis (KINGERY, 2021).

Em 2023 o Reino Unido convocou diversos países, incluindo os Estados Unidos e a China, juntamente com a União Europeia, para assinar o primeiro acordo mundial estabelecendo um entendimento compartilhado das oportunidades e riscos apresentados pela IA, ressaltando a necessidade dos países trabalharem juntos para enfrentar os desafios que tendem a surgir (SUNAK; DONELAN; CLEVERLY, 2023). Além disso, o próprio governo dos EUA vem trabalhando em diversas frentes para utilização da IA com a devida segurança e confiança (HOUSE, 2024).

O próprio Papa Francisco, chefe do Estado do Vaticano e líder mundial da Igreja Católica Apostólica Romana, escolheu como tema de sua carta para o Dia Mundial da Paz de 2024 "Inteligência Artificial e Paz", reforçando o cuidado necessário com o avanço tecnológico, em especial a IA, segundo ele: "Os avanços tecnológicos que não conduzem a uma melhoria da qualidade de vida da humanidade inteira, nunca poderão ser considerados um verdadeiro progresso" (FRANCISCO, 2024).

1.3 Hipótese

Diversas áreas da Saúde, entre elas a reabilitação neuromotora, se beneficiam de aplicações computacionais. Uma necessidade no processo de reabilitação é a execução dos movimentos adequados propostos pelo profissional de saúde, a identificação do movimento geralmente demanda uma análise do profissional, utilizando toda sua experiência prévia e conhecimento técnico. Uma ferramenta computacional que auxilie na classificação do movimento, utilizando IA, pode auxiliar na análise do profissional, acelerando o processo de identificação do movimento (RODRIGUES et al., 2022). Através da XAI, como uma expansão da IA, espera-se que seja possível aumentar a quantidade de dados fornecidas

ao profissional de saúde, permitindo, além da identificação do movimento, uma avaliação mais detalhada.

A aplicação de XAI em dados temporais, que tem em sua estrutura uma dimensão representando a passagem do tempo, é ainda uma área pouco explorada, representando um desafio e uma vez que dados de movimentos e reabilitação tendem a ter características temporais, a aplicação de XAI nesse contexto encara esse desafio (ŠIMIĆ; SABOL; VEAS, 2021). Além disso, os processos de avaliação de XAI, buscando compreender o nível de entendimento do público alvo, ainda são poucos definidos, conforme será demonstrado através da Revisão Sistemática de Literatura (RSL) (LOPES et al., 2022) (SISK et al., 2022).

Dessa forma, cria-se a seguinte hipótese: uma estratégia utilizando XAI pode colaborar de forma positiva com o processo de análise dos dados de reabilitação neuromotora?

1.4 Objetivos

O objetivo principal desta tese é apresentar uma estratégia de uso de XAI no apoio à reabilitação neuromotora, com foco na classificação de movimentos.

Os objetivos secundários abordados foram:

- Revisão Sistemática de Literatura: de modo a alcançar esse objetivo principal, foi realizada uma RSL que busca explorar o campo de pesquisa do uso de XAI na reabilitação neuromotora e com isso obter informações sobre a aplicação de técnicas de XAI de forma a melhorar o entendimento do usuário acerca do funcionamento de métodos de IA;
- Comparação de Ferramentas de XAI: por meio das informações obtidas na revisão, foi realizada uma comparação sobre algumas das ferramentas para uso de XAI existentes;
- Estudo sobre Avaliação de XAI: também por meio das informações obtidas na revisão, foi realizado um aprofundamento na questão de avaliação e métricas para métodos de XAI; e
- Desenvolvimento de ferramenta de XAI: com base em todo conteúdo reunido, foi elaborada a estratégia de uso, que contemplou o desenvolvimento de uma ferramenta que se encontra em testes por um colaborador pesquisador da área de reabilitação.

1.5 Considerações finais do capítulo

No presente capítulo foram apresentadas informações sobre as motivações e sobre o contexto desta tese, de forma a criar uma base para continuidade do mesmo, expondo a tendência de uso da IA e as preocupações que esse uso pode trazer.

1.6 Estrutura da tese

Esta tese está organizado em mais 8 capítulos e 2 apêndices:

- Capítulo 1: a presente introdução;
- Capítulo 2: são expostos conceitos para fundamentar o restante do trabalho;
- Capítulo 3: é apresentada a RSL realizada;
- Capítulo 4: são descritas ferramentas e bibliotecas para uso de XAI;
- Capítulo 5: é detalhado o desafio da avaliação em XAI;
- Capítulo 6: a estratégia de uso e o contexto onde ela está inserida são apresentados;
- Capítulo 7: apresenta o processo de desenvolvimento da estratégia de uso, em especial da ferramenta desenvolvida;
- Capítulo 8: contém a apresentação e discussão dos resultados obtidos;
- Capítulo 9: apresenta a conclusão desta;
- Apêndice A: apresenta os dados dos estudos incluídos na RSL; e
- Apêndice B: apresenta uma lista dos estudos excluídos na RSL.

2 Fundamentação Teórica

2.1 Considerações iniciais do capítulo

Neste capítulo serão apresentados os principais conceitos que norteiam esta tese, sendo eles: Inteligência Artificial e Inteligência Artificial Explicável.

2.2 Inteligência Artificial

O termo foi criado na década de 1950 pelo professor e pesquisador John McCarthy, ele a definiu como: “ciência e engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes” ([MCCARTHY, 2007](#)).

Diversos pesquisadores da área discutem a definição de IA, e com isso, uma definição considerada bastante esclarecedora é a dos autores Rich e Knight: "IA é o estudo de como fazer os computadores realizarem tarefas as quais, até o momento, os homens fazem melhor"([RICH; KNIGHT, 1994](#)). Já o autor Laurière complementa: "todo problema para o qual nenhuma solução algorítmica é conhecida é um problema de IA"([LAURIERE, 1990](#)).

Diante destas definições, [Russel e Norvig \(2012\)](#) buscaram categorizar elas, definindo 4 objetivos que a IA pode buscar:

- Pensar como um humano;
- Pensar racionalmente;
- Agir como um ser humano; e
- Agir racionalmente.

A IA pode incluir um grande número de vertentes: processamento de linguagem natural, representação de conhecimento, raciocínio automatizado, aprendizado de máquina, visão computacional, robótica e etc.

Desenvolvidos entre as décadas de 1980 e 1990, os chamados sistemas especialistas foram criados para simular o raciocínio de um especialista em uma área do conhecimento, respondendo a perguntas com informações sobre a área escolhida. São utilizados para auxílio na tomada de decisões complexas em diversas áreas como medicina, engenharia, finanças e direito. Sendo considerados uma das tecnologias presentes no "guarda-chuva" do termo Inteligência Artificial. Sempre existiu a preocupação com a explicabilidade e confiança

em ferramentas baseadas em sistemas especialistas, a explicabilidade para essa situação pode ser alcançada de diversas formas, entre elas: uso de regras explícitas, desenvolvimento de visualização ilustrando o funcionamento do modelo e métodos de decomposição (CARBONERA; GONÇALVES; SOUZA, 2023).

Nos últimos anos a inteligência artificial chamada generativa se encontra em um momento de rápido avanço e grande potencial, abrindo um leque de oportunidades para diversos setores, com ela é possível gerar conteúdo inédito de forma automatizada. Modelos de linguagem como o GPT-4 e o LLaMA demonstram capacidades impressionantes na geração de textos realistas e criativos, desde poemas e roteiros até código de programação e artigos científicos (BROWN et al., 2020) (OPENAI et al., 2024) (TOUVRON et al., 2023). Já na área de imagens, ferramentas como o DALL-E 2 e o Imagen 2 permitem criar imagens realistas a partir de simples descrições textuais, com aplicações promissoras em design, marketing e entretenimento (RAMESH et al., 2022) (SAHARIA et al., 2022).

Surgem cada vez mais pesquisas discutindo a questão da explicabilidade e transparência em modelos de linguagem, ressaltando a necessidade de que toda evolução que vem acontecendo aceleradamente seja acompanhada de forma equilibrada pela valorização da interpretabilidade (CAMBRIA et al., 2024). Entretanto, modelos de linguagem são sistemas extremamente complexos, com bilhões de parâmetros, dessa forma compreender o raciocínio por trás de cada resposta se torna um desafio.

2.3 Inteligência Artificial Explicável

A Inteligência Artificial Explicável (XAI) é uma vertente da IA que tem o objetivo de explicar seu próprio propósito, além de fornecer resultados. Essa explicação do propósito deve incluir sua finalidade e seu processo de decisão, que é amplamente reconhecido como um recurso crucial para a implantação prática de modelos de IA (ARRIETA et al., 2020). As discussões sobre XAI se tornam cada vez mais presentes no mundo da tecnologia.

Uma XAI busca demonstrar sua finalidade de forma transparente por meio dos seguintes aspectos (KIE-TEC, 2019):

- As forças e fraquezas da aplicação;
- Os critérios específicos utilizados para chegar ao resultado;
- A razão pela qual uma decisão foi tomada, em detrimento de outras opções, justificando assim a decisão;
- O nível de confiança apropriado para diversos tipos de decisão;
- Erros aos quais a aplicação em questão está vulnerável; e

- Formas de corrigir estes erros.

Organizações como a OpenAI ¹, uma organização sem fins lucrativos com foco no desenvolvimento de uma IA amigável, tendem a surgir e unir grandes empresas em torno desta filosofia, deixando para trás uma IA que existe como uma caixa preta e tornando ela algo transparente.

Alguns importantes pontos relacionados aos benefícios do uso da XAI podem ser vistos a seguir (NETO, 2021):

- Detectar vieses, preconceitos, falhas e vulnerabilidades nos modelos;
- Garantir a atuação de variáveis úteis, otimizando processos;
- Obter conhecimento dos negócios por meio da interpretação dos modelos;
- Atender os domínios que necessitam de explicações das decisões, que tratam de informações sensíveis, como saúde, segurança, entre outros;
- Atender questões regulatórias e legais; e
- Aumentar robustez das soluções e a confiança proporcionada.

Um dos principais pontos a se considerar diz respeito às decisões obtidas com auxílio de um modelo que afetam de maneira direta a vida dos seres humanos, impactando em campos como a medicina, direito e segurança pública, nessas situações, existe a necessidade emergente de se buscar compreender como os modelos de IA chegaram às suas decisões (GOODMAN; FLAXMAN, 2017). Um exemplo concreto pode ser citado na chamada medicina de precisão, onde os especialistas necessitam de mais informações provenientes do modelo além de uma resposta binária obtida a partir de uma classificação (TJOA; GUAN, 2019).

Um ponto positivo dos modelos considerados transparentes é que quando se ajusta um modelo transparente com os mesmos dados, sempre se obtém os mesmos resultados. Entretanto, os modelos não transparentes ou opacos não são igualmente reproduzíveis porque usam números aleatórios para inicializar seus pesos (MASÍS, 2021).

2.3.1 Abordagens conceituais

As técnicas de XAI podem ser classificadas de diversas formas, de acordo com as informações utilizadas, a forma de integração com a IA e compatibilidade com diferentes modelos, algumas das classificações mais comuns são:

¹ <https://openai.com/>

Global ou local: diz respeito ao nível de interpretação fornecido pela técnica XAI, sendo global referente ao modelo como um todo, explicando o comportamento de forma geral e local referente a explicação em um ponto individual, expondo o comportamento da decisão para cada predição (ALI et al., 2023).

Transparente ou posterior (*post-hoc*): trata do momento em que a explicação é gerada, onde as técnicas transparentes buscam a explicabilidade durante o processo de predição do modelo, tornando o processo transparente, ou *white-box*, já as técnicas posteriores fornecem a explicação depois do treinamento e da predição, por meio dos dados de entrada e saída (ALI et al., 2023).

Específico ou *model-agnostic*: essa classificação está relacionada ao suporte a diferentes modelos pela técnica de XAI. As técnicas específicas servem para gerar explicações sobre uma técnica ou uma família de técnicas em especial, já as técnicas *model-agnostic*, como o próprio nome diz, são agnósticas de modelo, servindo em qualquer modelo preditivo. A explicação com técnicas específicas geralmente demanda um conhecimento interno do modelo, já as técnicas agnósticas utilizam somente as predições e os dados de entrada (ALI et al., 2023).

2.3.2 Termos e definições

A área de pesquisa de XAI está intimamente ligada aos termos interpretabilidade e explicabilidade, com ambos sendo usados indevidamente de forma intercambiável. Entretanto, existem diferenças entre os conceitos, interpretabilidade é definida como uma característica passiva do modelo, de modo que se refere a quanto um determinado modelo faz sentido para um observador, esse termo pode ser usado como sinônimo para transparência. Já o termo explicabilidade se refere a uma característica ativa do modelo, está relacionado às ações ou procedimentos efetuados para detalhar e esclarecer o funcionamento e as funções internas de um modelo (ARRIETA et al., 2020). Outro termo utilizado no contexto de XAI é a compreensibilidade, definido como o grau que um humano pode entender uma decisão tomada por um modelo de IA. A compreensibilidade também está relacionada à medida da capacidade do público para entender o conhecimento contido no modelo. Ou seja, as características buscadas pela XAI estão em dois lados: no modelo e também no público que irá fazer uso da explicação (SCHNEEBERGER et al., 2023).

Sintetizando esses termos, Gunning et al. (2019) define XAI da seguinte maneira: “A XAI criará um conjunto de técnicas de aprendizado de máquina que permite que os usuários humanos entendam, confiem adequadamente e gerenciem efetivamente a geração emergente de parceiros artificialmente inteligentes”. Ainda existe uma falta de concordância no vocabulário de conceitos e nas diferentes definições que cercam a XAI, entretanto a ideia central sempre permanece na questão do entendimento por parte do usuário.

2.3.3 Público alvo

Um ponto a ser abordados em XAI é a dificuldade na definição do público alvo da explicação. Este ponto apresenta grande importância, de modo que a definição do público alvo correto impacta no objetivo e na nível da explicação. Na Tabela 1 são mostradas algumas opções de público alvo e o objetivo da explicação para cada opção.

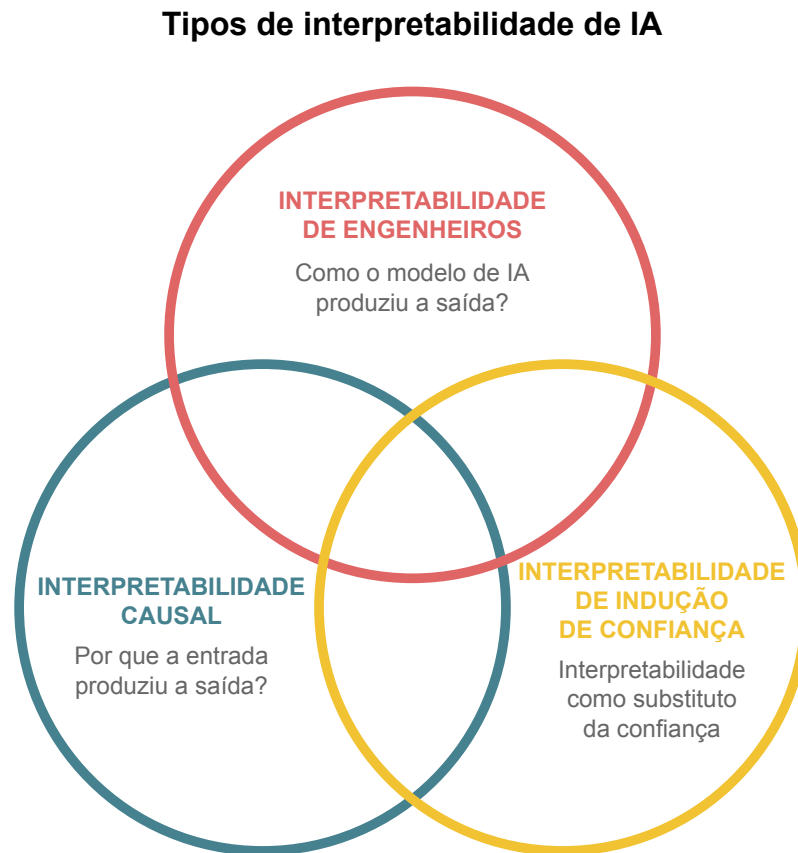
Tabela 1 – Público-alvo em XAI

Interessados	Motivação
Especialistas de domínio/usuários do modelo (por exemplo, médicos, agentes de seguros)	Confiar no próprio modelo, ganhar conhecimento científico
Entidades/agências reguladoras	Certificar modelo de cumprimento da legislação em vigor, auditorias
Gerentes e membros do conselho executivo	Avaliar a conformidade regulamentar, entender os aplicativos corporativos de IA
Cientistas de dados, desenvolvedores, proprietários de produtos	Garantir ou melhorar a eficiência do produto, pesquisa, novas funcionalidades
Usuários afetados pelas decisões do modelo	Entender a situação deles, verificar as decisões justas

Fonte: Traduzido e adaptado de [Arrieta et al. \(2020\)](#)

Nigam Shah, professor de medicina (informática biomédica) e de ciência de dados biomédicos na Universidade de Stanford, identificou três principais vertentes de interpretabilidade de acordo com diferentes públicos: explicabilidade para engenheiros, explicabilidade causal e explicabilidade que induz à confiança ([MILLER, 2021](#)). Na Figura 3 podem ser observadas essas vertentes e o modo como elas se sobrepõe, sendo assim algumas explicações podem estar na interseção das três vertentes.

Figura 3 – Diagrama mostrando as vertentes de interpretabilidade de acordo com diferentes públicos.



Fonte — Adaptado de (MILLER, 2021).

Para os engenheiros é interessante saber como é o funcionamento do modelo para auxiliar na depuração, já para médicos e no campo da saúde, a explicação causal pode ser útil na questão de ganho de confiança. “Mostrar a alguém uma explicação de uma zona diferente daquela de que ela precisa pode ter consequências inesperadas, como a perda de confiança em um modelo preciso e útil”, diz o professor Shah (MILLER, 2021).

Um grande desafio dos métodos de XAI, em especial para métodos de aprendizado profundo, é fornecer explicações que sejam acessíveis para a sociedade como um todo, inclusive para os responsáveis por formular políticas e leis. Essas explicações não devem exigir conhecimentos técnicos, de forma a tratar bem ambiguidades e auxiliar no desenvolvimento do direito à explicação para toda a sociedade, ponto que ainda não está presente em diversas legislações relacionadas a proteção de dados, como na GDPR na Europa (WACHTER; MITTELSTADT; FLORIDI, 2017).

Os especialistas em determinada área, chamados especialistas de domínio, não sendo conhecedores de IA e Aprendizado de Máquina, do inglês *Machine Learning* (ML), acabam por buscar entender a decisão da IA através de raciocínio lógico e relação de causa

e efeito, enquanto do cientistas de IA e ML buscam entender o funcionamento do modelo internamente, levando em conta seus componentes e o impacto que geram nas previsões.

2.3.4 Métricas

De acordo com [Arrieta et al. \(2020\)](#), existe uma grande necessidade em pesquisar conceitos e métricas para avaliação da explicabilidade alcançada, em especial para modelos de aprendizado profundo. Uma métrica ou grupo de métricas deve ser capaz de permitir uma comparação relevante sobre o quanto o modelo se adéqua na definição de explicável.

[Hoffman et al. \(2018\)](#) se aprofunda nesta questão da falta de métricas, questionando em especial os desafios e prospectos dessa temática. A questão que norteia esse estudo é: "Se apresentarmos a um usuário um sistema de IA que explica como funciona, como saberemos se a explicação funciona e se o usuário alcançou uma compreensão pragmática da IA?". Essa questão principal ainda se expande nas seguintes questões:

- A qualidade das explicações;
- Se os usuários estão satisfeitos com as explicações;
- Quão bem os usuários entendem os sistemas de IA;
- Como a curiosidade motiva a busca por explicações;
- Se a confiança do usuário e a confiança na IA são adequadas; e
- Como o sistema de trabalho humano-XAI funciona.

2.3.5 Necessidade de explicações

Outro ponto apresentado pela área de pesquisa de XAI é se realmente as explicações apresentam resultados positivos e são necessárias. Explicações podem levar o público alvo a confiar em um sistema que pode estar errado, essa situação é semelhante ao uso de estatística muitas vezes distorcidas para embasar conclusões erradas e tendenciosas. Em uma pesquisa realizada pela Microsoft, foram utilizados dois modelos de IA, sendo um explicável e outro não, avaliadores avaliaram positivamente o modelo explicável, ainda que as decisões estivessem claramente erradas e ficaram reticentes com resultados obtidos por meio do modelo não explicável ([POURSABZI-SANGDEH; GOLDSTEIN; HOFMAN, 2021](#)).

Um dos pontos onde a explicabilidade pode não ser desejada ou causar problemas é em situações onde existe a questão do segredo comercial, onde o funcionamento de sistemas e algoritmos faz parte da atividade-fim da empresa, sendo parte do que difere o serviço ou produto oferecido dos concorrentes. Nessas situações, a tecnologia utilizada

pela organização é protegida como um segredo comercial e demandas por explicabilidade, expondo mecanismos internos de funcionamento dessa tecnologia podem colocar em risco a exclusividade do serviço prestado, causando prejuízos a organização (ADADI; BERRADA, 2018).

Informações de explicabilidade podem também ser utilizadas para melhorar a efetividade de ataques maliciosos que buscam confundir o modelo por meio das chamadas amostras adversárias. Entretanto, nesse mesmo contexto, é possível utilizar essas informações para evitar esse tipo de ataque, executando experimentos e fortalecendo o modelo (MARINO; WICKRAMASINGHE; MANIC, 2018).

Muitas vezes a explicabilidade pode não ser considerada útil por demandar um esforço considerável sem oferecer um retorno equivalente, por exemplo, em uma situação onde um mecanismo de IA realiza a leitura de códigos de postais em encomendas, esse mecanismo pode errar, porém, seu erro causa pouco impacto na sociedade, como discriminações e injustiças e a frequência do erro pode nem mesmo gerar um problema financeiro relevante (MASÍS, 2021).

Diante desses pontos negativos relacionados à aplicação de explicabilidade, surge a diretriz de se preferir, quando possível, o uso de técnicas transparentes no desenvolvimento de um sistema baseado em IA. Levando em conta que diferentes necessidades e contextos demandam diferentes soluções, recomenda-se primeiro considerar a utilização de modelos interpretáveis como padrão em vez de métodos de modelagem sofisticados, porém opacos (RUDIN, 2019). Entretanto, geralmente modelos considerados caixa preta superam os modelos considerados caixa branca na maioria das métricas (MASÍS, 2021).

2.3.6 Abordagens de Inteligência Artificial Explicável

Diferentes iniciativas e grupos de pesquisas em todo mundo vem desenvolvendo a XAI por meio das mais diversas abordagens, Arrieta et al. (2020) elencaram as principais abordagens utilizadas, que podem ser observadas na Figura 4.

Figura 4 – Diagrama mostrando as principais abordagens de XAI.

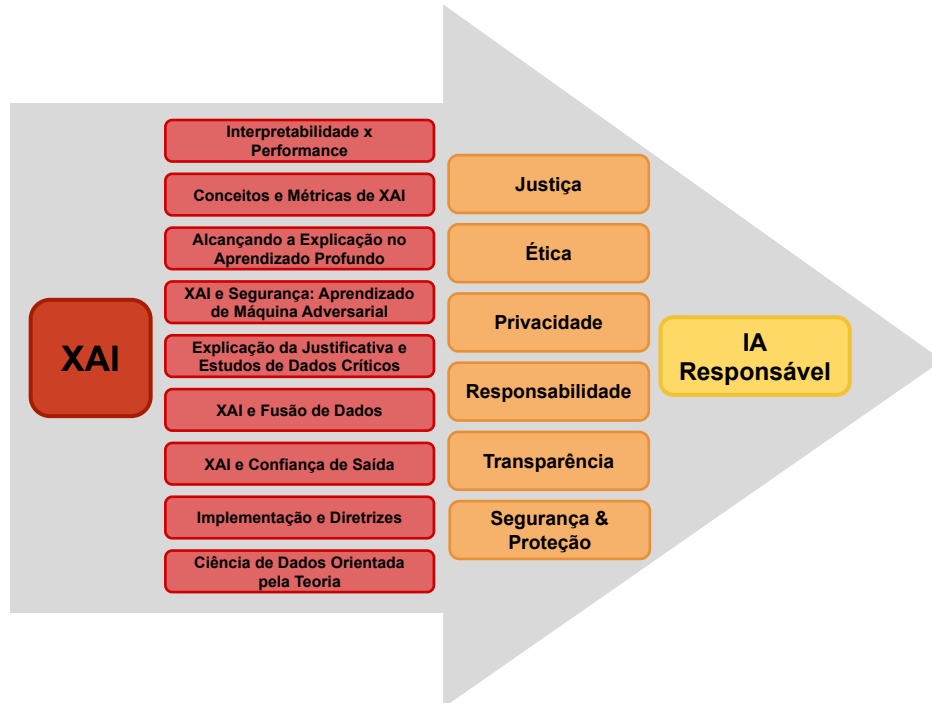
Fonte — Adaptado de (ARRIETA et al., 2020).

Realizando uma vasta pesquisa, [Arrieta et al. \(2020\)](#) buscaram relacionar os diferentes modelos de IA utilizados às diferentes abordagens de explicabilidade, gerando assim uma taxonomia que elenca os principais trabalhos envolvendo cada modelo e sua abordagem de explicação. Durante a realização de sua pesquisa, foi detectada a necessidade de criação de uma taxonomia adicional específica para modelos de aprendizagem profunda.

O trabalho de [Fjeld et al. \(2020\)](#) buscou mapear todos os documentos que propuseram, desde 2016, princípios para nortear o desenvolvimento de IA. Baseado nesse mapeamento e nos principais desafios de XAI, ([ARRIETA et al., 2020](#)) chegaram ao conceito de IA Responsável, um tipo de IA que busca por meio de diversos mecanismos, entre eles a XAI, oferecer ferramentas de tratamento de dados de forma automática e inteligente que sigam princípios éticos e de transparência. A Figura 5 demonstra a criação do conceito de IA Responsável.

Figura 5 – Diagrama mostrando a criação do conceito de IA Responsável através de diversos aspectos.

Resumo dos desafios do XAI e seu impacto nos princípios da IA responsável.



Fonte — Adaptado de [Arrieta et al. \(2020\)](#).

Quando se considera o modelo completo, sem a necessidade de melhorá-lo por meio da explicabilidade, não sendo necessário questionar o seu funcionamento, pode-se considerar o entendimento do problema como completo. Caso essa seja a situação, a recomendação seria não utilizar o aprendizado de máquina, uma vez que a aplicação de aprendizado de máquina deve ser usada quando não é possível definir um algoritmo com regras específicas para tratar o problema ([MASÍS, 2021](#)).

2.4 IA e XAI em saúde

O uso de IA já é uma realidade no setor da saúde, abrindo uma gama de possibilidades promissoras para diagnósticos, pesquisas e uma evolução na relação entre profissionais da área e pacientes. Diversas aplicações inovadoras já estão em uso, mas ainda há um caminho a ser percorrido para garantir que esse avanço tecnológico traga benefícios concretos e mensuráveis para todos os envolvidos, de forma confiável e transparente ([JUNIOR; NAKAYA; RIZZO, 2024](#)).

A IA na área de saúde tem um enorme potencial para revolucionar o setor, mas é fundamental que seu desenvolvimento e aplicação sejam guiados por princípios éticos, científicos e responsáveis. Somente assim pode-se garantir que essa tecnologia traga

benefícios reais para pacientes, profissionais e para toda a sociedade (BRUNO; PEREIRA; FALTAY, 2023). Com a preocupação sobre os aspectos de confiança e transparência no uso de IA em saúde, surgem possibilidades de aplicação de XAI na área, buscando entregar mais informações além da decisão do modelo (BORYS et al., 2023).

Ainda que a área de XAI venha se desenvolvendo muito nos últimos anos, a aplicação em dados temporais, como séries temporais e em especial, vídeos é um grande desafio (ŠIMIĆ; SABOL; VEAS, 2021). Existem muitos exemplos disponíveis de XAI sendo aplicada em imagens, dados tabulares, etc., porém ao se adicionar o tempo como uma nova dimensão, com uma relação de sequência entre as amostras, a dificuldade em se obter explicações aumenta. O trabalho de Kolarik et al. (2023) propõe uma revisão sobre a explicabilidade em modelos profundos aplicados a vídeos médicos, encarando o desafio de tratar dados temporais em uma área importante que é a saúde.

Sobre a detecção de movimentos em vídeo com IA na área de saúde, o trabalho de Leo et al. (2022) propõe uma análise da área de pesquisa e um levantamento das possibilidades de evolução para o futuro, com o foco na movimentação de bebês e em doenças neurológicas. Explorando essa área de pesquisa, com a aplicação de XAI, pode-se citar o trabalho de Sakkos et al. (2021) que propõe uma análise dos movimentos de bebês utilizando IA e XAI para detectar sinais de paralisia cerebral (ZHU et al., 2021) (MCCAY et al., 2021).

2.5 Acidente Vascular Cerebral e Reabilitação

O chamado acidente vascular cerebral (AVC) é uma situação médica onde se danifica tecido cerebral próximo de sua ocorrência, pode ser isquêmico, com bloqueio de vasos sanguíneos, ou hemorrágico, com ruptura de um vaso. Em parte das situações não é fatal, porém causa sequelas, que variam conforme a área do cérebro afetada. Dentre as sequelas, as principais são problemas visuais, na fala, perda de coordenação motora, etc., diversos tratamentos podem ser utilizados buscando amenizar essas sequelas, podendo ser citada a fisioterapia como uma das principais. É uma doença de impacto mundial, causando incapacitação a longo prazo, custos médicos para o sistema de saúde e redução na qualidade de vida dos afetados (HEALTH, 2020).

A fisioterapia se mostra uma abordagem com bons resultados devido à capacidade das vias neurais de se recuperarem parcialmente, mesmo após o AVC. No tratamento neuromotor oferecido pela fisioterapia, existe o estímulo do rearranjo das vias e circuitos motores do cérebro, por conta da capacidade de plasticidade cerebral ou neuroplasticidade, que permite que os neurônios se regenerem tanto anatomicamente quanto funcionalmente, formando novas conexões sinápticas. Quanto mais precoce for o início das abordagens para terapia, melhores serão os resultados, onde é importante também a colaboração do

paciente, buscando não abandonar o tratamento. Soluções como abordagens lúdicas e interfaces diferenciadas, como realidade aumentada e virtual, podem colaborar com o aumento do interesse por parte do paciente (REGO; MOREIRA; REIS, 2010).

No campo da Ciência da Computação, o desenvolvimento de ferramentas computacionais para auxiliar na reabilitação pós-AVC abre um leque de oportunidades promissoras para aprimorar o processo de recuperação de pacientes que sofreram um acidente vascular cerebral. Diversas ferramentas computacionais podem ser utilizadas para complementar o trabalho tradicional de fisioterapia e fonoaudiologia, oferecendo estímulos personalizados, monitoramento contínuo e feedback imediato aos pacientes. Aplicações de realidade virtual e realidade aumentada permitem a criação de ambientes imersivos e interativos que simulam situações da vida real, motivando os pacientes a praticar tarefas motoras e cognitivas de forma divertida e engajadora (RODRIGUES et al., 2022) (JOHN; SANGEETHA, 2024).

2.6 Considerações finais do capítulo

Neste capítulo foram apresentados os principais conceitos para nortear a presente tese, com base nestes conceitos, será descrita a seguir a metodologia utilizada para realização de uma RSL. Por meio da revisão apresentada no Capítulo 3, se busca compreender melhor o contexto de uso de XAI e como é construído esse tipo de ferramenta.

3 Revisão Sistemática de Literatura

3.1 Considerações iniciais do capítulo

Neste capítulo são apresentados os métodos e dados acerca do processo de revisão realizado neste estudo, descrevendo principalmente o método de RSL.

3.2 Metodologia

A RSL é uma técnica de pesquisa cujo objetivo é executar revisões abrangentes de literatura sobre um determinado tema, avaliando os resultados de forma não tendenciosa, detalhando e explicando todo o processo, incluindo seus critérios de seleção, de modo que o pesquisador que for utilizar a revisão avalie a qualidade desta e atualizá-la ou executá-la novamente se for necessário. Esta técnica foi desenvolvida inicialmente para utilização na área Médica, porém, vem cada vez mais sendo utilizada em outras áreas do conhecimento, incluindo a Ciência da Computação ([BRERETON et al., 2007](#)).

3.2.1 Classificação dos estudos

Durante a revisão os estudos são categorizados da seguinte forma:

- Estudos identificados: são os estudos retornados pelo sistema de busca selecionado, seja ele manual ou eletrônico. É registrada a quantidade de estudos encontrada e a fonte destes estudos;
- Estudos duplicados: são aqueles estudos que estão presentes em mais de uma base de dados selecionada, são contabilizados somente uma vez;
- Estudos não selecionados: são aqueles estudos que de maneira objetiva não atendem os critérios de inclusão. São excluídos durante a fase de seleção. É registrada apenas a quantidade destes estudos;
- Estudos selecionados: são aqueles que aparentemente atendem os critérios de inclusão, são incluídos na fase de seleção. Registram-se as referências completas destes estudos;
- Estudos excluídos: são aqueles estudos que, após avaliação do texto completo, não atendem aos critérios de inclusão, sendo excluídos na etapa de extração; e
- Estudos incluídos: são aqueles estudos que, após avaliação do texto completo, atendem aos critérios de inclusão, se mantém durante todo o processo, sendo utilizados na etapa de extração.

3.2.2 Objetivos da revisão

A revisão executada tem como objetivo compreender os processos de uso da chamada Inteligência Artificial Explicável no contexto da Reabilitação Neuromotora, analisando estudos que tratam da utilização deste conceito de forma concreta e prática, fornecendo explicabilidade para ferramentas que utilizam Inteligência Artificial e Aprendizado de Máquina no apoio à Reabilitação Neuromotora, e por meio disso, compreender profundamente o campo de pesquisa.

3.2.3 Questões de pesquisa

Após diversas discussões e pesquisas prévias, chegou-se à seguinte questão principal de pesquisa: Como se usa XAI na reabilitação neuromotora?

Esta pergunta principal pode ser expandida nas seguintes questões:

- Quais são as soluções de XAI (técnicas ou métodos) já existentes em uso na reabilitação neuromotora?
- Quais os recursos visuais utilizados por essas soluções de XAI na reabilitação neuromotora?;
- Quais são as fontes de dados utilizadas para aplicar a XAI na reabilitação neuromotora?; e
- Quão maduras são essas soluções de XAI na reabilitação neuromotora?

3.3 Desenvolvimento do protocolo

Nesta seção é descrito o desenvolvimento do protocolo definido para guiar a presente revisão. Foram detalhadas as seguintes etapas: escolha da estratégia de busca, escolha das bases de dados, definição dos critérios de inclusão e exclusão, definição do processo de extração de dados com o formulário de extração e do processo de síntese dos estudos.

3.3.1 Estratégia de busca e bases de dados

Durante as buscas preliminares, foram sendo selecionados termos para se obter uma busca mais focada, com uma quantidade de resultados que possibilitasse a execução da revisão, alguns termos inclusive foram sendo identificados como interessantes nos resultados obtidos, gerando um grupo de controle e um processo recursivo de refinamento. Ao final deste processo, definiu-se uma revisão sistemática utilizando os termos:

machine learning, deep learning, trust, explainable, artificial intelligence, interpretability, neurorehabilitation, neuromotor rehabilitation.

As bases de dados oferecem o recurso de busca avançada, onde os termos de buscas podem ser concatenados utilizando caracteres adequados que servem como operadores, formando assim a chamada *string* de busca. Para o presente estudo, a *string* de busca utilizada foi:

((("explainable OR interpretable") AND ("artificial intelligence"OR "machine learning"OR "deep learning")) OR XAI) AND ("neuromotor rehabilitation"OR neurorehabilitation)

Foram considerados estudos válidos artigos publicados na língua inglesa em periódicos e conferências, e somente os estudos localizados a partir do ano de 2018 até a execução das buscas e da fase de identificação dos estudos em 23/06/2024.

As buscas iniciais foram realizadas utilizando *strings* de busca nas bases de dados ACM Digital Library, IEEE Digital Library, ISI Web of Science, Science@Direct e Scopus. Estas bases foram selecionadas por fornecerem uma quantidade adequada de estudos e possuírem acesso liberado por meio dos convênios da universidade. A seleção destas foi realizada considerando a relevância, afinidade com as áreas de estudo e quantidade de resultados obtidos nos testes preliminares.

3.3.2 Critérios de inclusão e exclusão

Nesta etapa foram definidos os critérios para avaliar quais estudos avançam para a próxima etapa da revisão. Os critérios devem ser definidos a partir das questões de pesquisa. Outros critérios que podem ser utilizados na execução de uma revisão são: restrição por idioma, por exemplo, restringir a revisão a somente artigos em inglês, restrição por área, restrição por artigos primários, excluindo estudos secundários como outras revisões sistemáticas, etc. (NEIVA; SILVA, 2016).

Não foram encontrados trabalhos correlatos escritos em língua portuguesa, desta forma optou-se pela inclusão somente de artigos escritos na língua inglesa.

Para a presente revisão, os seguintes critérios de inclusão foram definidos:

- Afinidade do estudo com os temas desejados – buscou-se incluir na revisão os estudos que possuíam relação com o tema Uso de XAI e similares no contexto da reabilitação neuromotora, estudos sem relação ao assunto foram excluídos; e
- Apresenta uma aplicação prática – foram selecionados e identificados todos os estudos que apresentaram uma aplicação prática e concreta dos temas desejados, de modo que se pudesse entender o processo de desenvolvimento desta aplicação.

Como critérios de exclusão, foram definidos os seguintes critérios:

- Estudo completo não está disponível no idioma inglês – estudos retornados na pesquisa, com o resumo disponível na língua inglesa, porém, com o texto completo somente disponível em outro idioma foram excluídos da revisão pela impossibilidade de leitura;
- Estudo não possui acesso liberado para a instituição - estudos retornados na pesquisa, com o resumo disponível, porém, com o texto completo não disponível nas permissões fornecidas pela universidade, foram excluídos da revisão pela impossibilidade de leitura;
- Falta de afinidade do estudo com os temas desejados – estudos retornados na busca, porém, que não apresentavam afinidade com os temas desejados foram excluídos;
- Não é um artigo – resultados retornados que não poderiam ser considerados estudos foram desconsiderados, entre eles estão artigos de opinião, índices de anais de eventos, entre outros;
- Não trata de implementação, se limitando apenas a conceitos de XAI - estudos que tratam dos assuntos desejados de maneira conceitual, sem realizar implementação, foram excluídos da revisão;
- Revisão ou *survey* sobre o assunto - estudos que apresentavam uma revisão anterior sobre o assunto foram excluídos da presente revisão, para serem analisados de forma particular; e
- Trata de ferramenta já analisada – situações onde foram identificados mais de um artigo proveniente de um mesmo grupo de autores e se tratando da mesma ferramenta optou-se por avaliar o estudo mais recente, excluindo os demais por já terem a ferramenta em questão analisada.

3.3.3 Extração de dados

Na etapa de extração de dados, a questão de pesquisa deve ser respondida por meio da análise dos artigos selecionados na etapa anterior, mediante um formulário de extração previamente elaborado com questões que detalhem aspectos da pergunta de pesquisa principal. Os dados são registrados de maneira que fiquem vinculados aos estudos a que se referem, esse registro pode ser realizado utilizando ferramentas específicas ou métodos mais simplificados, como uma planilha. Ao se executar a leitura completa dos estudos, as possíveis respostas ao formulário de extração são registradas (NEIVA; SILVA, 2016).

Para o presente estudo, as seguintes questões compuseram o formulário de extração, conforme pode ser observado na Tabela 2:

Tabela 2 – Questões que compuseram o formulário de extração.

Número	Questão	Objetivo
1	Nome da técnica ou método	Esta pergunta visa identificar o nome dado a ferramenta e verificar se a ferramenta em questão já não foi analisada em outro estudo.
2	Qual o objetivo?	Por meio desta pergunta buscou-se identificar o objetivo principal que levou ao desenvolvimento da solução.
3	Cita diretamente uso de XAI?	Por meio desta pergunta buscou-se identificar se o estudo já citava XAI como uma área de pesquisa ou somente utilizava aspectos de XAI sem citar o termo diretamente.
4	Cita uso de alguma biblioteca de XAI? Qual?	Por meio desta pergunta buscou-se identificar as tendências de uso das bibliotecas de XAI já existentes.
5	Técnica de IA desenvolvida com explicabilidade ou explicação posterior?	Conforme entendimento de Rudin (2019) , as técnicas de explicabilidade aplicadas em IA podem ser utilizadas em dois momentos: durante o processo de execução do modelo, de forma integrada, gerando assim uma técnica desenvolvida com explicabilidade (os chamados modelos transparentes) ou posteriormente a execução do modelo, a partir dos dados resultantes (a chamada explicabilidade <i>post-hoc</i>).
6	Qual a fonte de dados utilizada?	Por meio desta questão se buscou identificar quais são as fontes de dados utilizadas, buscando também conhecer bases de dados que fossem públicas e pudessem ser úteis em outros estudos.
7	Faz uso de visualização da informação para obter explicabilidade? Como?	Por meio desta questão, buscou-se avaliar a utilização do fator visual na explicação realizada, analisando os aspectos e técnicas de Visualização da Informação utilizados.
8	Houve estudo de avaliação ou validação com usuários sobre o método de XAI?	Esta pergunta identificou se houve um processo de avaliação ou validação da ferramenta desenvolvida com usuários, de forma a se obter retorno sobre as experiências de uso da XAI.

Fonte: Produzida pelo autor.

3.3.4 Ferramenta Parsifal

Para apoio à realização da presente revisão, foi utilizada a ferramenta online Parsifal, desenvolvida para apoiar a realização de Revisões Sistemáticas da Literatura no contexto da Engenharia de Software. Dessa forma, pesquisadores distribuídos geografi-

camente podem trabalhar em conjunto em um espaço de trabalho web compartilhado, desenvolvendo o protocolo e conduzindo a pesquisa ([PARSIFAL, 2018](#)).

Esta ferramenta permite aos pesquisadores preencher todo o protocolo de uma revisão sistemática de forma facilitada. Ela possui informações sobre todas as fases presentes em uma revisão, desde o objetivo, questão de pesquisa, buscas, estratégias de seleção, critérios de inclusão e exclusão, formulário de extração até a sintetização dos resultados.

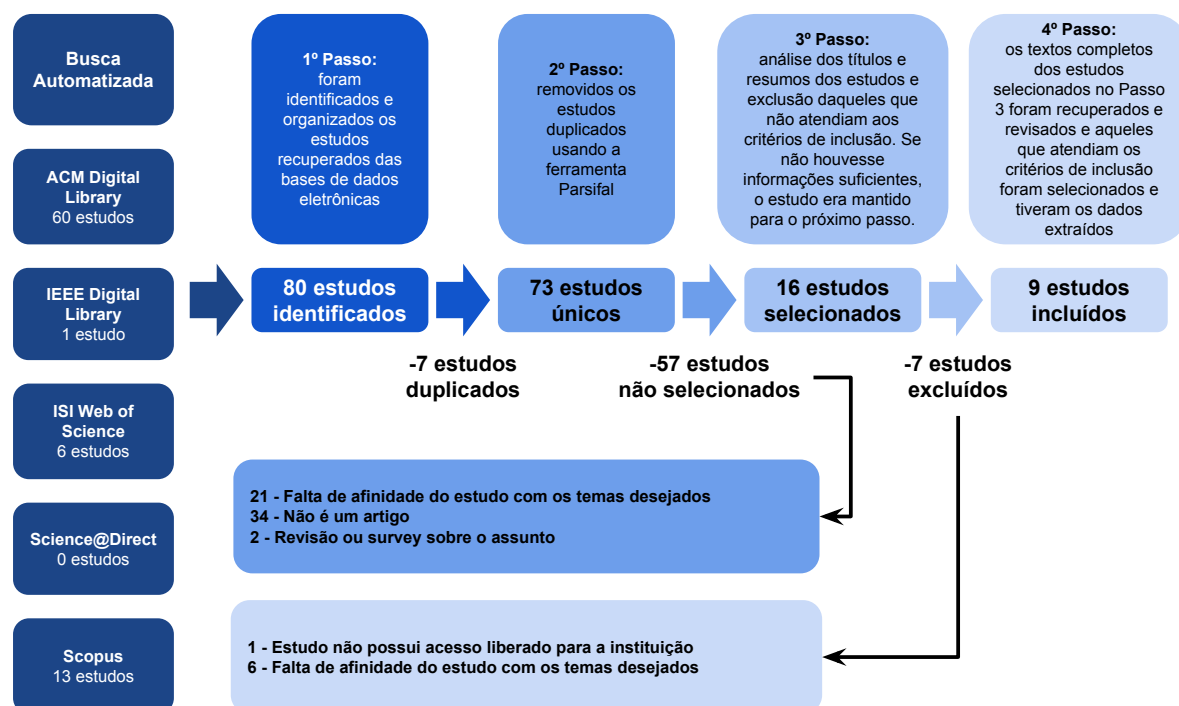
3.4 Resultados

Nesta seção são apresentados os resultados obtidos ao aplicar o protocolo de revisão nas bases científicas e extrair os dados dos estudos retornados, buscando assim obter informações para serem discutidas.

A síntese e sumarização das informações coletadas foram reunidas e expostas por meio de tabelas, de forma a mostrar a divisão dos estudos baseada nas perguntas que compõem o formulário de extração. Sempre que possível foram utilizados gráficos e recursos visuais para exposição da informação, utilizando as boas práticas de visualização, de modo a facilitar a compreensão dos dados.

3.4.1 Condução

Na Figura 6 pode-se observar em detalhes a evolução da revisão executada.

Figura 6 – Diagrama mostrando as etapas da revisão sistemática de literatura.

Fonte — Produzida pelo autor.

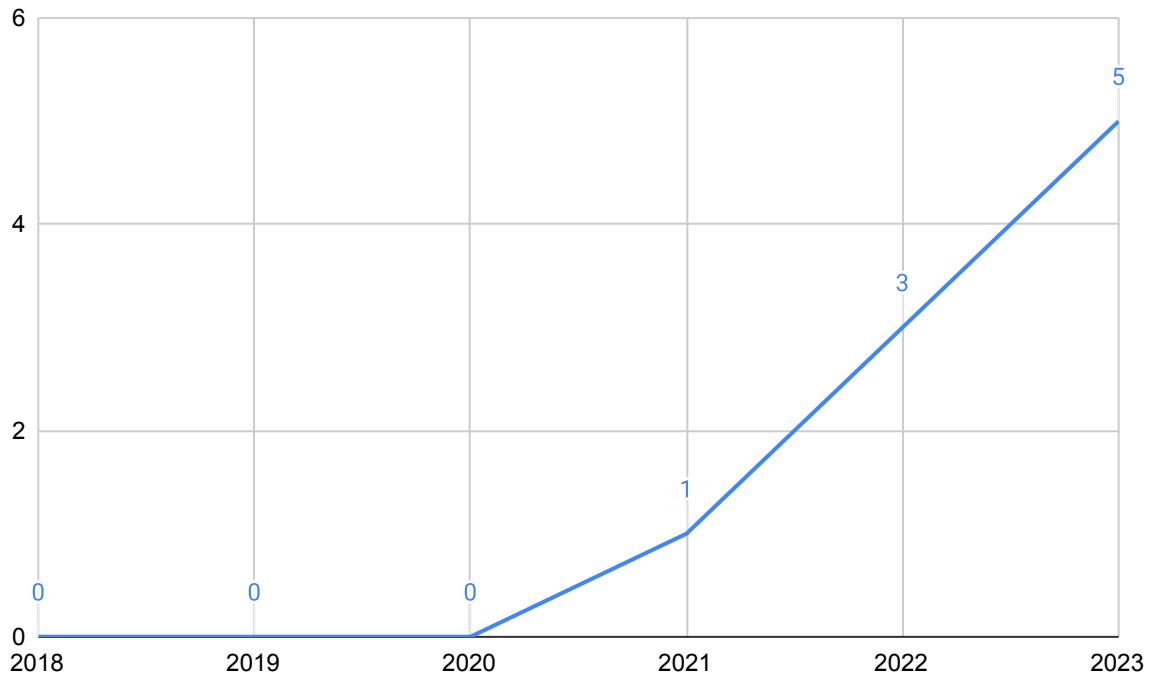
3.4.2 Estudos excluídos e incluídos

Durante a fase de extração, 7 estudos foram excluídos, pois, mesmo após a análise inicial do título e resumo, durante a leitura completa, foi verificado que não atendiam os critérios de inclusão. Estes estudos estão listados no Apêndice B.

Os dados detalhados e referências dos 9 estudos analisados e incluídos na revisão são apresentados no Apêndice A deste trabalho. As próximas seções detalham os dados obtidos após análise destes estudos.

3.4.3 Anos dos estudos

Na Figura 7 pode ser observado o aumento da quantidade de estudos que tratam do tema da presente revisão no passar dos últimos 3 anos (2021, 2022 e 2023), demonstrando a relevância crescente do tema, considerando que o ano de 2024 não foi analisado de forma completa e não retornou estudos.

Figura 7 – Gráfico mostrando a evolução da quantidade de estudos durante os anos.

Fonte — Produzida pelo autor.

3.4.4 Objetivos dos estudos

Por meio da Questão 2 apresentada na Tabela 2, "Qual o objetivo?", buscou-se identificar a principal motivação que levou ao desenvolvimento de cada estudo. Para melhor organização dos dados, os estudos com objetivos semelhantes foram agrupados em categorias. A Tabela 3 representa a divisão por categorias de objetivo.

Tabela 3 – Categorias de objetivos dos estudos.

Categoria	Estudos
Avaliação da evolução, prognóstico e desfecho clínico dos pacientes	5
Avaliação das sessões de reabilitação, considerando dados de movimento	4

Fonte: Produzida pelo autor.

Pode-se observar que a maioria dos estudos (5) tem como objetivo "Avaliação da evolução, prognóstico e desfecho clínico dos pacientes", nesta categoria de objetivos foram agrupados os estudos, que em sua maioria através de dados clínicos, buscaram avaliar o processo de recuperação e a evolução do paciente que sofreu AVC ou alguma outra doença com sintomas neuromotores e foi submetido a tratamentos de reabilitação, a avaliação com método de IA permite a predição com certo grau de certeza do estado futuro do paciente de acordo com sua situação clínica e de como ela se altera.

Já os demais artigos (4) buscam de formas diversas avaliar os movimentos exe-

cutados durante as sessões de reabilitação e as respostas musculares do paciente, estes estudos foram agrupados na categoria "Avaliação das sessões de reabilitação, considerando dados de movimento".

3.4.5 Uso da expressão XAI

Para compreender o alcance do termo XAI e o crescimento da área de pesquisa, buscou-se através da Questão 3 da Tabela 2 avaliar se os estudos analisados usavam de forma direta o termo XAI, o resultado pode ser observado na Tabela 4.

Tabela 4 – Presença da expressão XAI.

Cita diretamente uso de XAI	Estudos
Sim	8
Não	1

Fonte: Produzida pelo autor.

A quase totalidade dos artigos (8) cita o uso de Inteligência Artificial com aspectos de Explicabilidade, Interpretabilidade, Transparência, citando diretamente inclusive a expressão XAI.

3.4.6 Bibliotecas de apoio

A Questão 4 da Tabela 2 buscou identificar a utilização de bibliotecas que implementavam métodos de XAI para facilitar a utilização. Diversos estudos citaram a utilização de bibliotecas já existentes para uso de XAI, como a LIME (RIBEIRO; SINGH; GUESTRIN, 2016) e a SHAP (LUNDBERG; LEE, 2017), entre outras. Na Tabela 5 podem ser verificadas as quantidades de estudos que citaram uso de bibliotecas e na Tabela 6 detalha as bibliotecas citadas e o número de citações.

Tabela 5 – Uso das bibliotecas.

Cita uso de alguma biblioteca XAI	Estudos
Sim	7
Não	2

Fonte: Produzida pelo autor.

Tabela 6 – Bibliotecas citadas.

Bibliotecas	Estudos
SHapley Additive exPlanations (SHAP)	5
Local Interpretable Model-agnostic Explanations (LIME)	1
Accumulated local effects (ALE)	1
Permutation Feature Importance (PFI)	1
Random forest feature importance (RFI)	1
Grad-CAM	1

Fonte: Produzida pelo autor.

Pode ser observada a quantidade relevante (7) de pesquisas que utilizam de alguma forma bibliotecas no uso de XAI, para obter um ganho de velocidade utilizando soluções já implementadas, ao invés de se optar por realizar todo desenvolvimento, sem nenhum tipo de apoio. Métodos independentes de modelo, como SHAP, acabam por ser amplamente utilizados, pois, atendem oferecendo explicabilidade nas mais diversas situações, sejam para tratamento de dados tabulares, textos e imagens, inclusive para modelos de aprendizado profundo ([ARRIETA et al., 2020](#)).

3.4.7 Momento de geração da explicação

Por meio da Questão 5 apresentada na Tabela 2, buscou-se identificar o momento de geração da explicação, podendo ela ser uma explicação posterior, utilizando dados de entrada e saída do modelo ou durante a execução do modelo de IA, se integrando ao modelo e tornando-o transparente.

A quantidade de estudos que demonstraram utilizar explicação posterior (8) mostra a opção por se utilizar os métodos de IA já consolidados e posteriormente agregar explicabilidade a eles, para acelerar o processo de desenvolvimento. A questão não se aplicou a um estudo por conta do mesmo não citar explicabilidade diretamente. Essa prevalência pode ser visualizada também na Tabela 7.

Tabela 7 – Estudos separados por momento de geração da explicabilidade.

Momento de geração da explicabilidade	Estudos
Explicação posterior	8
Técnica desenvolvida com explicabilidade	0
Não se aplica	1

Fonte: Produzida pelo autor.

3.4.8 Fontes de dados

Através da Questão 6 da Tabela 2, examinou-se os conjuntos de dados utilizadas em cada estudo. Todos os estudos analisados indicam a coleta de dados e produção do

seu próprio conjunto de dados, com pacientes que sofreram AVC ou outras doenças com sintomas neuromotores e que estavam passando pelo processo de reabilitação nos mais diversos locais. Isso demonstra a escassez de dados abertos e a opção pela coleta de dados da maneira desejada, mais adequada para cada estudo.

3.4.9 Visualização da informação

Por meio da Questão 7 apresentada na Tabela 2, buscou-se identificar o uso de visualização da informação no processo de explicabilidade, permitindo o aumento da compreensão sobre o funcionamento do método de IA por meio de representações visuais. Os gráficos utilizados são na maioria dos estudos (6) provenientes das bibliotecas de XAI utilizadas em cada estudo. Dos demais estudos, parte (2) utilizam um método de visualização próprio e um estudo não apresenta visualização sendo utilizada para explicabilidade, estes dados podem ser observados também na Tabela 8.

Tabela 8 – Métodos de visualização citados.

Faz uso de visualização da informação para obter explicabilidade?	Estudos
Sim, proveniente da biblioteca XAI utilizada	6
Sim, com representações próprias	2
Não	1

Fonte: Produzida pelo autor.

3.4.10 Avaliação ou validação com usuários

Conforme a Questão 8 apresentada na Tabela 2, buscou-se identificar se o estudo analisado descrevia a realização de um processo de avaliação ou validação da técnica de explicabilidade proposta com usuários, por meio de testes, entrevistas ou formulários de avaliação, buscando assim garantir os resultados positivos da proposta com relação à facilidade de compreensão e uso. Nenhum estudo demonstrou ter realizado um processo como o desejado. Ainda que o processo de teste com usuários seja bastante trabalhoso, a falta dele enfraquece as conclusões propostas.

3.4.11 Estudos secundários

Nesta seção serão apresentados os estudos excluídos, por já proporem uma revisão anterior relacionada ao assunto. Estes estudos podem ser chamados de estudos secundários ou comparativos e apresentam uma boa base para continuidade de pesquisas relacionadas ao presente tema. A seguir podem ser vistos os títulos dos 2 estudos secundários identificados e uma breve descrição do seu objetivo.

- **Wearable Interactions for Users with Motor Impairments: Systematic Review, Inventory, and Research Implications:** revisão sistemática da literatura sobre interações vestíveis para usuários com deficiências motoras, fez a meta-análise de 57 artigos científicos identificados nas bases de dados ACM DL e IEEE Xplore. (SIEAN; VATAVU, 2021);
- **Intelligent Robotics in Pediatric Cooperative Neurorehabilitation: A Review:** busca fornecer uma visão abrangente do cenário em evolução da robótica orientada por IA na neuroreabilitação pediátrica. (TSUR; ELKANA, 2024);

3.5 Discussão

Na presente seção são discutidos os resultados obtidos, de forma a extrair deles novas informações e tendências, expondo também a relevância e as limitações da revisão executada.

3.5.1 Tendências

Para se obter informações que componham uma resposta à questão de pesquisa principal apresentada, "Como se usa XAI na reabilitação neuromotora?", serão abordadas as questões de pesquisa secundárias, apresentando informações obtidas que representem uma resposta para cada uma delas.

- **Quais são as soluções de Inteligência Artificial Explicável (técnicas ou métodos) já existentes em uso na reabilitação neuromotora?** Foi possível averiguar que existe uma diversidade de técnicas, muitas delas baseadas em bibliotecas, com destaque para a biblioteca SHAP, porém, com diferentes abordagens e métodos. A utilização de bibliotecas para apoio indicou um ponto de partida interessante para continuidade do projeto, buscando compreender como pode ser definido um padrão de trabalho na área de pesquisa, que ainda é muito recente. Como relação a integração das ferramentas de XAI às soluções de IA, a explicação posterior foi fortemente citada, indicando que tende a ser mais eficiente utilizar os métodos de IA mais comuns, ainda que sem aspectos de explicabilidade e depois adicioná-los utilizando ferramentas adequadas;
- **Quais os recursos visuais utilizados por essas soluções de XAI na reabilitação neuromotora?** Analisando o uso de recursos visuais e da Visualização da Informação, pode-se constatar que a maioria dos estudos utiliza as visualizações e gráficos fornecidos pelas bibliotecas de explicabilidade. São poucos os estudos que se propõe a criar suas próprias visualizações, gerando novas metáforas visuais, de

forma mais adequada ao contexto de utilização da XAI, em especial no apoio da reabilitação;

- **Quais são as fontes de dados utilizadas para aplicar a XAI na reabilitação neuromotora?** Todos os estudos avaliados indicam a coleta de dados e a criação de conjuntos de dados próprios, indicando a necessidade de alinhamento com profissionais da área de reabilitação para acesso adequado aos dados no que diz respeito a aspectos éticos. Não houve menção ao uso de dados públicos disponíveis de forma aberta; e
- **Quão maduras são essas soluções de XAI na reabilitação neuromotora?** Por meio da informação dos anos de publicação dos estudos, pode-se constatar que tratam de uma área nova, ainda em expansão, que vem ganhando força com o passar dos anos, oferecendo diversas oportunidades de pesquisa. Entretanto, foi averiguado que muitos estudos não apresentam a avaliação das técnicas e ferramentas com usuários reais no que diz respeito a avaliação da explicabilidade, se restringindo a testes simulados e opiniões pouco formalizadas. Isso demonstra que existe um espaço para evolução, ainda que o processo de avaliação com usuários seja um processo complexo, que demanda um esforço considerável.

3.5.2 Descrição dos estudos

Na presente seção, serão apresentadas descrições breves sobre cada um dos estudos incluídos na revisão.

A Transferable Deep Learning Prognosis Model for Predicting StrokePatients' Recovery in Different Rehabilitation Trainings (LIN et al., 2022): estudo avalia o prognóstico de 15 pacientes sobreviventes de AVC antes e depois de diversos tratamentos, através de dados clínicos e fisiológicos coletados.

An Interpretable Deep Learning Optimized Wearable Daily Detection System for Parkinson's Disease (CHEN et al., 2023): estudo busca detectar e avaliar caminhada do paciente com doença de Parkinson, através dos dados de 100 indivíduos utilizando cinco unidades de medição inercial fixadas no pulso, tornozelo e cintura são usadas para coletar dados de movimento dos indivíduos durante um teste de caminhada de 10 metros.

Cross-validation of predictive models for functional recovery after post-stroke rehabilitation (CAMPAGNINI et al., 2022): estudo analisa a contribuição de cada fator prognóstico na alta após reabilitação intensiva pós-AVC em 278 pacientes, utilizando técnicas de validação cruzada.

EEG fractal dimensions predict high-level behavioral responses in minimally conscious patients (LIUZZI et al., 2023): estudo analisa registros de eletroencefalograma, um exame não invasivo que mede a atividade elétrica do cérebro e registra os padrões de

ondas cerebrais, para classificação de estado de 68 paciente com estado de consciência mínima.

eXplainable AI Allows Predicting Upper Limb Rehabilitation Outcomes in Sub-Acute Stroke Patients (GANDOLFI et al., 2023): estudo busca prever as pontuações (escores) finais de recuperação funcional após reabilitação de membros superiores utilizando um grupo de 95 pacientes pós-AVC subagudo internados em uma Unidade de Neuroreabilitação entre janeiro de 2018 e outubro de 2020.

Factors influencing neuromuscular responses to gait training with a robotic ankle exoskeleton in cerebral palsy (CONNER et al., 2023) : estudo busca identificar fatores que influenciam as respostas neuromusculares a uso de um exoesqueleto resistivo de tornozelo em indivíduos com paralisia cerebral, são feitas sessões de exercícios com os pacientes utilizando o exoesqueleto e posterior avaliação dos movimentos.

Implementation of eXplainable Artificial Intelligence Case Study on the Assessment of Movements to Support Neuromotor Rehabilitation (CAMARGO; DIAS; BREGA, 2023): estudo aplica XAI em dados de classificação de movimentos, para apoio a reabilitação neuromotora, utilizando banco de dados próprio, produzido por grupo de pesquisa e um avatar para demonstração das coordenadas e dos movimentos.

Predicting outcome of patients with prolonged disorders of consciousness using machine learning models based on medical complexity (LIUZZI et al., 2022): estudo investiga o impacto das complicações médicas na predição do desfecho clínico, para isso utiliza dados de um grupo de 176 pacientes com desordem prolongada de consciência inscritos em um projeto multicêntrico.

Upper Limb Motion Tracking and Classification: A Smartphone Approach (RODRIGUES et al., 2021): estudo busca capturar, transformar e classificar dados de movimentos capturados com câmera RGB de *smartphone* com aprendizado de máquina, utilizando rastreamento corporal e estimativa de pose.

3.5.3 Relevância do estudo

O presente estudo pode ser considerado relevante por conta da confiabilidade apresentada pelos dados coletados, da robustez da metodologia utilizada e do cuidado aplicado, levando a resultados concretos e verificáveis, afastando-se de informações subjetivas. Isso foi obtido após diversas buscas prévias, com uma variedade de *strings* de busca, para se realizar uma busca focada que retornasse estudos relevantes para o tema em questão.

3.5.4 Limitações do estudo

Como ponto de limitação, pode ser citado a restrição do método de busca automatizado utilizado, não recorrendo a buscas manuais ou outras formas de pesquisa. Pode

ser apontado como limitante também a questão da falta da informação de forma clara em muitos estudos, não informando os nomes das ferramentas desenvolvidas, entre outras informações faltantes ou não tão claras.

3.6 Considerações finais do capítulo

Este capítulo investigou a área de pesquisa do Uso XAI na Reabilitação Neuromotora por meio da execução de uma RSL que foi realizada em diversas bases de dados e que, depois das devidas fases de seleção, resultou na inclusão e análise de 9 estudos primários.

A presente revisão contribuiu mostrando tendências que contribuem no processo de construção e uso de uma ferramenta de XAI como apoio na reabilitação neuromotora, as possibilidades de explicação posterior ao processo de aprendizado, a utilização de Visualização da Informação e a possibilidade de realização de testes com usuários para avaliação da XAI.

Com base nos resultados obtidos e discutidos neste capítulo do trabalho, foi dada continuidade no estudo, considerando o uso de bibliotecas para uso de XAI e a possibilidade de aplicação das técnicas apoiando o uso de IA na classificação de movimentos terapêuticos na reabilitação. No próximo capítulo, são detalhadas e avaliadas algumas ferramentas de XAI disponíveis para uso.

4 Ferramentas de Inteligência Artificial Explicável

4.1 Considerações iniciais do capítulo

Neste capítulo estão apresentadas algumas ferramentas que buscam permitir o uso de explicabilidade nos métodos de IA. Uma vez que a explicabilidade provida por essas ferramentas é adicionada aos *pipelines* de IA, serão tratados métodos que em sua maioria utilizam a chamada explicabilidade posterior ou *post-hoc*, utilizando os dados de entrada e saída do modelo e não necessitando de um conhecimento interno tão profundo deste, sendo assim agnósticas, também chamadas de *Model-agnostic*.

4.2 *Frameworks*, bibliotecas e métodos

Uma diferenciação complexa, que acaba por ainda não ser muito clara e definida, porém que necessita ser considerada é a questão de *framework* ou biblioteca para explicabilidade e método para explicabilidade. *Framework* ou biblioteca diz respeito a implementação, geralmente pensando na integração com a IA e suas bibliotecas. O termo método está mais ligado a solução matemática utilizada para gerar a explicação. Embora existam diversos métodos e também diversas bibliotecas, a maioria das bibliotecas e implementações utilizam os métodos LIME (RIBEIRO; SINGH; GUESTRIN, 2016) e valores de Shapley (ROTH, 1988). Cabe aqui observar que os autores do método LIME o propõe juntamente com uma biblioteca que leva o mesmo nome. Existem outros métodos matemáticos utilizados para geração de explicabilidade em IA, porém, são menos utilizados. A biblioteca LIME e seu método são detalhados na Seção 4.3 e a biblioteca SHAP juntamente com os valores de Shapley são detalhados na Seção 4.4.

4.3 LIME

O projeto LIME (abreviação de *Local Interpretable Model-agnostic Explanations*, em português, explicações interpretáveis locais agnósticas de modelo) foi disponibilizado em 2016, tendo seu código distribuído por meio do repositório na plataforma GitHub¹. O projeto busca explicar o que os classificadores ou modelos de aprendizado de máquina estão fazendo ao realizar suas predições. Atualmente suporta a explicação de previsões individuais para classificadores de texto, para classificadores que atuam em tabelas (matrizes com dados

¹ <https://github.com/marcotcr/lime>

numéricos ou categóricos) ou para classificadores de imagens, a biblioteca disponibilizada é utilizada junto a linguagem Python por meio de um pacote chamado `lime` (RIBEIRO; SINGH; GUESTRIN, 2016). O desenvolvimento da ferramenta está pausado, com a última versão disponibilizada em 03/04/2020.

Detalhando os termos que formam a sigla LIME, *model-agnostic* se refere à propriedade que possibilita a geração de explicações para qualquer modelo de aprendizagem supervisionada, tratando-o como uma "caixa preta", e *local explanations* está relacionado ao fato do LIME fornecer explicações que são localmente fiéis nos arredores ou nas proximidades da uma amostra que será explicada.

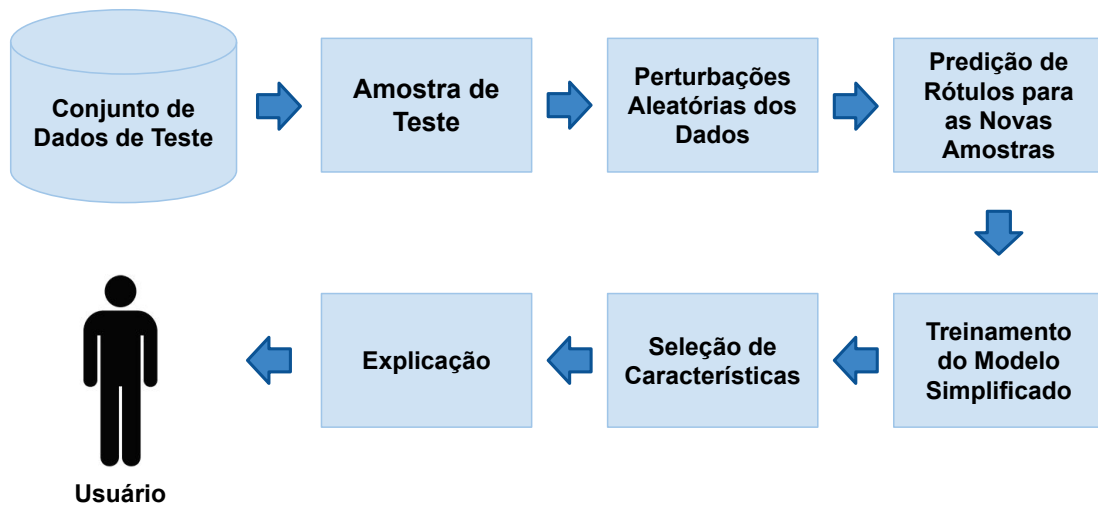
O LIME pressupõe que todo modelo, por mais complexo que seja, é linear em escala local, dessa forma ele tenta identificar um modelo simples que possa se ajustar a uma única observação, se aproximando do modelo global naquela localidade, esse modelo mais simplificado serve para gerar as explicações das previsões do modelo complexo.

Sendo um modelo local, o LIME é aplicado em uma única amostra quando é executado, seguindo os seguintes passos:

- O método LIME recebe o modelo de IA e uma amostra de teste;
- Através de perturbações nas características de entrada, com leves alterações, o método gera novas amostras próximas da amostra analisada, chamadas vizinhos ou conjunto de dados substituto. Por padrão, ele produz 5.000 amostras seguindo a distribuição normal;
- Através do modelo de IA que está sendo analisado, o valor da variável de destino é calculado para cada vizinho;
- As distâncias entre a amostra analisada e os vizinhos são medidas, através desse processo se obtém informações sobre as características de entrada mais relevantes para o modelo;
- O conjunto de dados formado pelos vizinhos é utilizado no treinamento de um modelo linear mais simplificado; e
- Para realizar a interpretação da amostra analisada, o modelo linear é interpretado, considerando que seu comportamento no local daquela amostra é semelhante ao modelo real, mais complexo.

O fluxo de dados para realização de uma explicação através da biblioteca LIME por de observado na Figura 8.

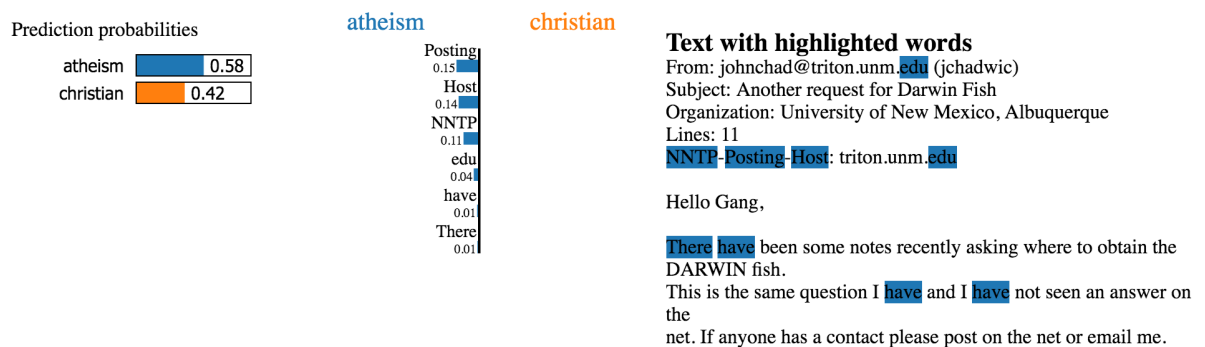
Figura 8 – Diagrama demonstrando o fluxo de dados para realização de uma explicação através da biblioteca LIME.



Fonte — Adaptado de Nigar et al. (2022).

Nas Figuras 9 e 10 podem ser vistas aplicações de técnicas de XAI por meio da biblioteca LIME.

Figura 9 – Biblioteca de XAI LIME sendo utilizada para explicar classificação de texto nas classes "ateísmo" ou "cristianismo". Palavras destacadas em azul colaboram com a classificação como "ateísmo", palavras destacadas em laranja colaboram com a classificação como "cristianismo".



Fonte — Lundberg (2021).

Figura 10 – Biblioteca de XAI LIME sendo utilizada para explicar classificação de imagem de gato, indicando áreas que colaboraram positiva (área verde) e negativamente (área vermelha) para classificação.



Fonte — [Lundberg \(2021\)](#).

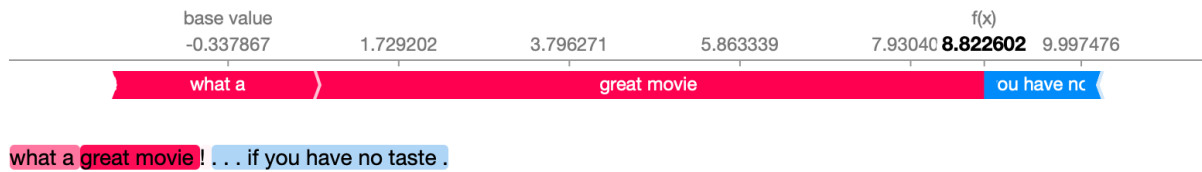
4.4 SHAP

SHAP (SHapley Additive exPlanations) é uma biblioteca com abordagem baseada em teoria dos jogos, que busca explicar a saída de dados de qualquer modelo de aprendizado de máquina. Baseado nos valores de Shapley, a ferramenta SHAP busca compreender o quanto cada característica de entrada colabora para a produção da saída do modelo ([LUNDBERG; LEE, 2017](#)). O projeto teve sua divulgação realizada a partir de 2017 por meio da plataforma GitHub ². Possui o desenvolvimento ativo no momento, com a última versão disponibilizada em 27/06/2024.

Valor de Shapley é um conceito utilizado nas teorias dos jogos cooperativos, essa teoria foi criada para garantir a distribuição justa de dinheiro para todos os jogadores que contribuíram de forma variada para o resultado de um determinado jogo. Por meio da chamada justiça de Shapley, se busca garantir para cada jogador um pagamento justo. Para determinar a contribuição de um jogador, o resultado é calculado com e sem o jogador presente no subconjunto de jogadores, a diferença observada é utilizada na obtenção dos valores de Shapley para cada jogador. No *framework* SHAP esses valores são traduzidos em quantidades e direções (positiva ou negativa) do efeito de cada característica ou recurso nas previsões do modelo. Na Figura 11 pode ser vista uma aplicação de uma técnica de XAI por meio da SHAP.

² <https://github.com/slundberg/shap>

Figura 11 – Biblioteca de XAI SHAP sendo utilizada para explicar classificação da opinião sobre um filme como positiva ou negativa, indicando trechos que colaboraram positiva (cor rosa) e negativamente (cor azul) para classificação.



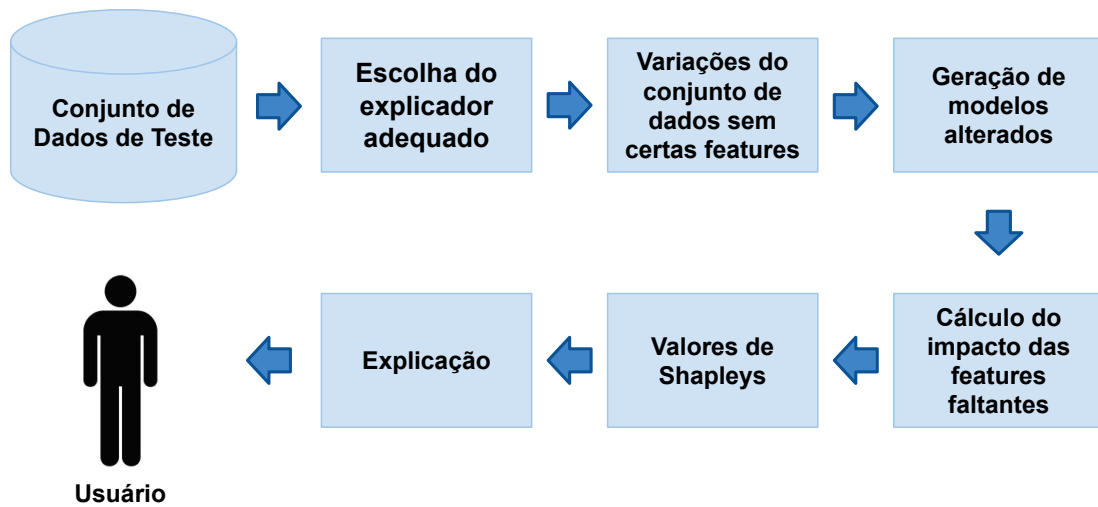
Fonte — Ribeiro (2021).

O funcionamento do SHAP se baseia nos seguintes passos:

- Definir um *explainer*, em português, explicador, a biblioteca SHAP fornece diferentes opções para diversos modelos de IA;
- Calcular os valores SHAP para o conjunto de treino;
- Valores SHAP são contribuições para o modelo definir entre uma das classes;
- O somatório dos valores SHAP não podem ultrapassar de 50%, em um problema com duas classes;
- Valores positivos considerando uma classe de referência indicam ser favoráveis aquela classe na predição; e
- Valores negativos indicam que a classe correta não é aquela de referência, mas sim outra classe.

O fluxo de dados para realização de uma explicação através da biblioteca SHAP por de observado na Figura 12.

Figura 12 – Diagrama demonstrando o fluxo de dados para realização de uma explicação através da biblioteca SHAP.



Fonte - Produzido pelo autor

4.5 Outras ferramentas

Dando continuidade a apresentação de ferramentas para uso de XAI, serão abordadas ferramentas de menor relevância, mas que propõe a busca por explicabilidade em diferentes contextos.

4.5.1 ELI5

ELI5 é uma biblioteca desenvolvida na linguagem de programação Python que busca fornecer depuração e visualização de diversos modelos de aprendizado de máquina, gerando explicações para modelos considerados caixa preta. O nome ELI5 vem da expressão “*Explain like I’m 5*”, que significa que a explicação deve ser feita como se fosse para uma criança de 5 anos. Utiliza LIME para gerar alguns tipos de explicação. As últimas alterações no repositório datam de outubro de 2022, mostrando pouca atividade no projeto (KOROBOV, 2017). Na Figura 13 pode ser observada a explicação gerada pela biblioteca ELI5 para a classificação de uma imagem.

Figura 13 – Exemplo de uso da biblioteca ELI5, indicando a área mais relevante para a classificação através de um mapa de calor sobreposto a imagem original.

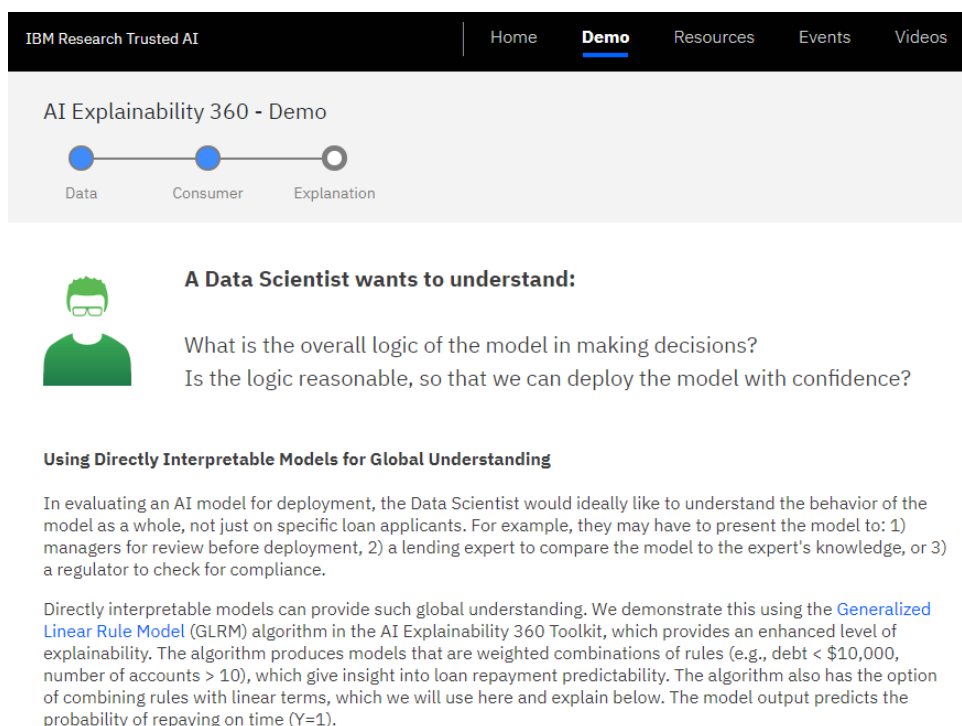


Fonte - [Korobov \(2017\)](#)

4.5.2 AI Explainability 360

O AI Explainability 360 é uma biblioteca de código aberto baseada na linguagem Python cujo objetivo é oferecer explicabilidade para diversos modelos de aprendizado de máquina. Ela abrange diversos algoritmos de XAI, entre eles LIME e SHAP, para cobrir diferentes dimensões e métricas de explicabilidade. Este projeto recebe apoio da empresa de tecnologia IBM. O desenvolvimento da ferramenta está ativo, com a última versão disponibilizada a cerca de 1 ano, em julho de 2023, mas os arquivos do repositório possuem alterações em março de 2024 ([ARYA et al., 2019](#)). Na Figura 14 pode ser observada a interface baseada em web disponível com estudos de caso para demonstração do funcionamento da biblioteca.

Figura 14 – Exemplo de uso da biblioteca AI Explainability 360, gerando uma explicação textual, obtido executando a demonstração da ferramenta disponível na página do projeto.

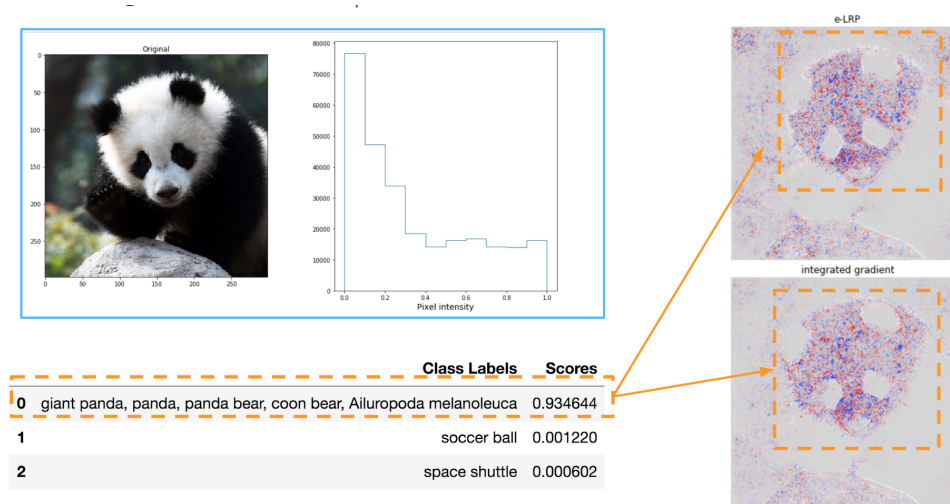


Fonte - Produzido pelo autor

4.5.3 Skater

Skater é uma biblioteca para interpretação de modelos de Inteligência Artificial mantida com o apoio da empresa Oracle, de código aberto, baseada na linguagem Python. Suporta modelos de caixa preta globalmente (analisando o comportamento do modelo em um conjunto de dados completos) quanto localmente (analisando uma única previsão), utiliza diversos algoritmos, entre eles cita o LIME. Foi criado originalmente por Aaron Kramer, Pramit Choudhary e o restante da equipe do site DataScience.com para ajudar cientistas e entusiastas de dados a obter uma melhor visão dos seus modelos. Entretanto, aparentemente o projeto teve seu desenvolvimento pausado, pois o repositório de código sofreu as últimas alterações em março de 2023 (SKATER, 2017). Na Figura 15 pode ser observada a explicação gerada pela biblioteca Skater para a classificação de uma imagem.

Figura 15 – Exemplo de uso da biblioteca Skater, indicando as áreas da imagem que levaram a classificação da imagem como "panda".

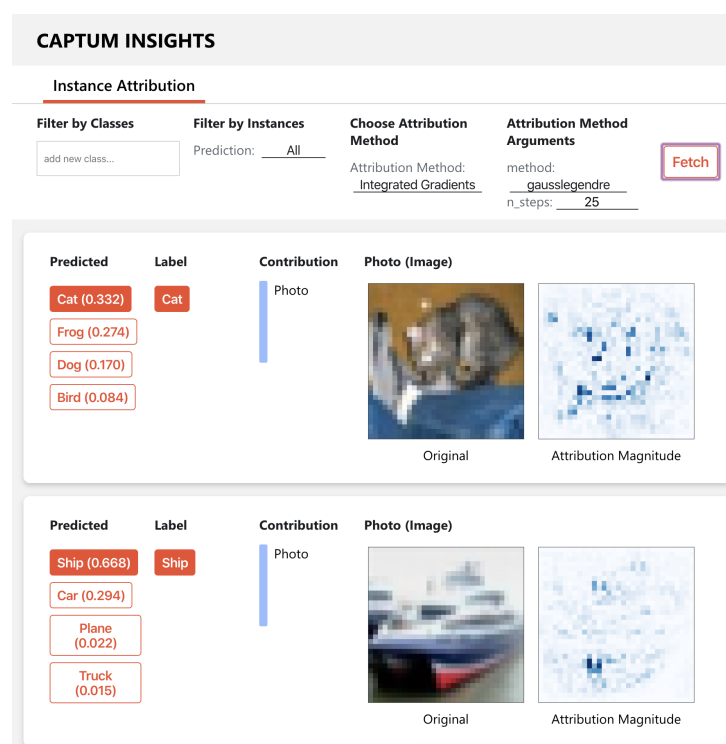


Fonte - [Skater](#) (2017)

4.5.4 Captum

Captum, que em latim significa compreensão, é uma biblioteca de código aberto para interpretabilidade de modelos feita para utilização conjunta com o *framework* de IA PyTorch. Se encontra em versão de testes, beta, e com desenvolvimento ativo, tendo sua última versão disponibilizada na plataforma GitHub em 05/12/2023 e alterações constantes nos arquivos do repositório. Tem como primeira audiência os desenvolvedores de IA e utiliza variações do SHAP para gerar alguns tipos de explicação ([KOKHLIKYAN et al., 2020](#)). Na Figura 16 pode ser observada a explicação gerada pela biblioteca Captum para a classificação de duas imagens.

Figura 16 – Exemplo de uso da biblioteca Captum, com 2 exemplos, indicando a classificação das imagens como "gato" e "navio", após o devido tratamento das imagens.

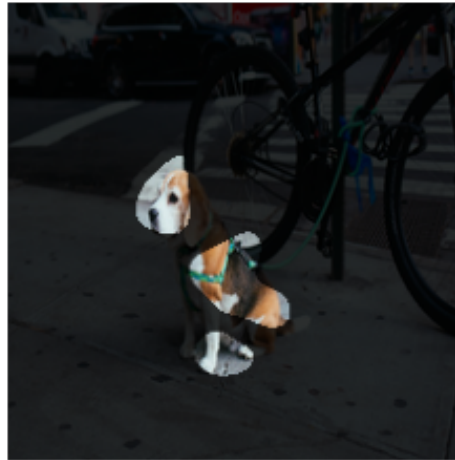


Fonte - [Kokhlikyan et al. \(2020\)](#)

4.5.5 TruLens

O TruLens é uma biblioteca de explicabilidade que se propõe a ser *cross-framework*, fornecendo uma camada de abstração sobre os *frameworks* TensorFlow (versão 1 e versão 2 com Keras) e PyTorch, focando em aprendizado profundo e redes neurais. Seu desenvolvimento está ativo, com a última versão disponibilizada em 24/06/2024 e com alterações frequentes nos arquivos do seu repositório de código, é mantido pelo grupo TruEra, originário da Universidade Carnegie Mellon ([LEINO et al., 2021](#)). Na Figura 17 pode ser observada a biblioteca sendo utilizada para geração de um mapa de saliência na classificação de uma cachorro da raça Beagle.

Figura 17 – Exemplo de uso da biblioteca TruLens, aplicando uma máscara sobre a imagem original, indicando as áreas mais relevantes para a classificação através de janelas nessa máscara.



Fonte - [Leino \(2021\)](#)

4.5.6 VisualDL

VisualDL é uma ferramenta para explicabilidade e visualização ligada ao *framework* de IA chinês PaddlePaddle, permite visualizar diversos gráficos de tendências de parâmetros e outras estruturas do modelo, além de amostras de dados, curvas e etc. Possui apoio da empresa de tecnologia chinesa Baidu. Suporta a execução em tempo real, por meio de navegadores web e utiliza a linguagem de programação Python. Seu desenvolvimento está ativo, com a última versão disponibilizada em 17/01/2023 e atualizações posteriores nos arquivos do repositório ([VISUALDL, 2021](#)). Na Figura 18 pode ser observada a interface da biblioteca VisualDL mostrando alguns parâmetros de um modelo.

Figura 18 – Exemplo de uso da biblioteca VisualDL, demonstrando um *dashboard*, com informações sobre modelo e seus parâmetros.

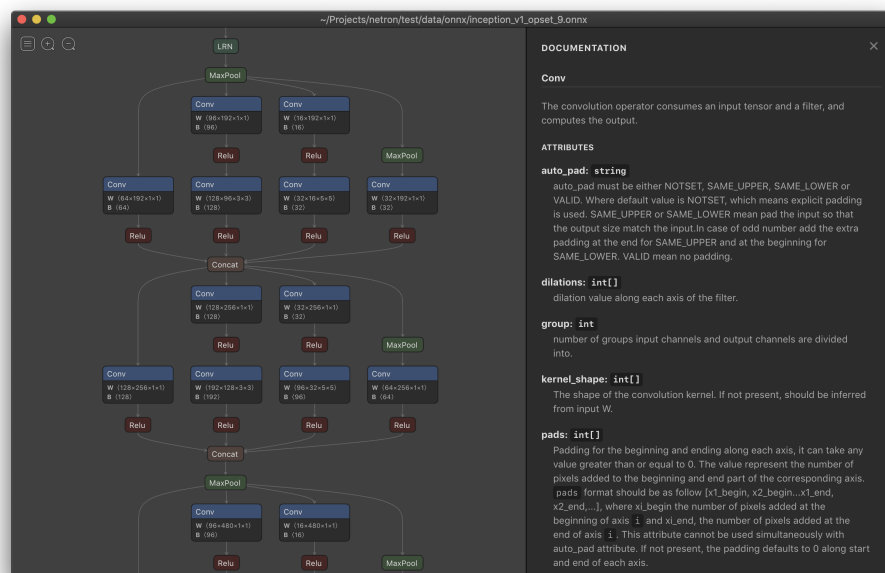


Fonte - [VisualDL \(2021\)](#)

4.5.7 Netron

Netron é um visualizador de redes neurais, busca acrescentar um grau de explicabilidade para esse tipo de modelo de aprendizado de máquina. Compatível com uma gama enorme de *frameworks* de IA, é limitado em recursos, fornecendo somente a visualização da estrutura interna da rede neural, não oferecendo informações sobre as predições e sobre o comportamento do modelo como um todo, está com seu desenvolvimento ativo, com a última versão disponibilizada em 01/07/2024 e com alterações constantes no repositório, possui também exemplos para testes disponibilizados para acesso por meio do navegador. (NETRON, 2021). Na Figura 19 pode ser observada a interface da ferramenta Netron mostrando uma visualização de uma rede neural.

Figura 19 – Exemplo de uso da ferramenta Netron, exibindo a estrutura da rede neural.

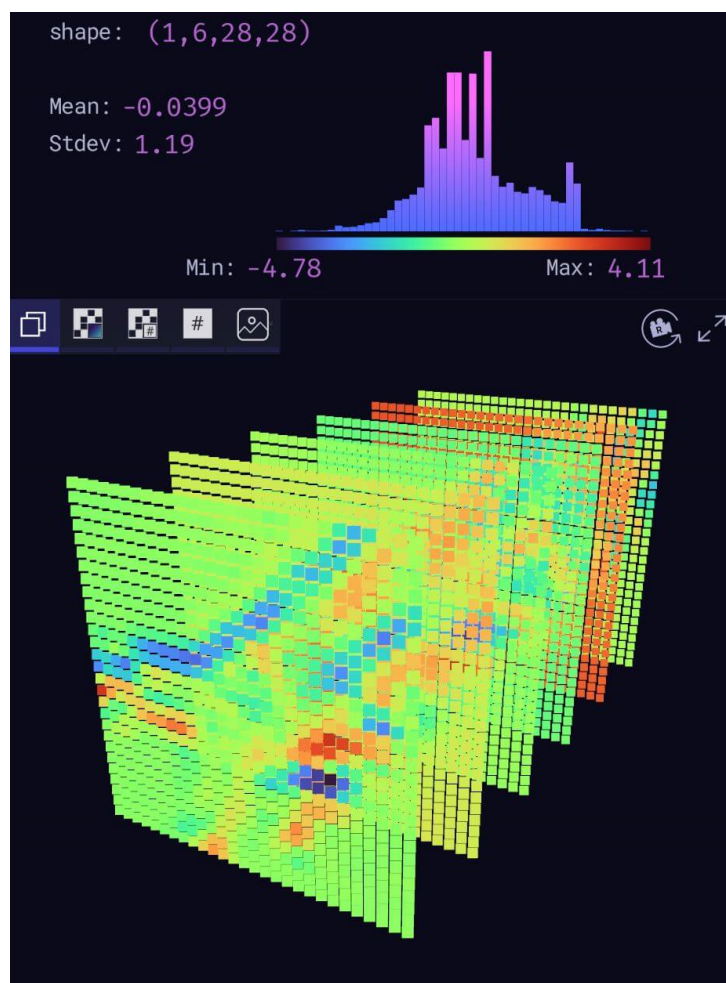


Fonte - Netron (2021)

4.5.8 Zetane Viewer

Zetane Viewer é outra ferramenta que atua como visualizador de redes neurais, ajudando a entender o aprendizado de máquina nas redes neurais artificiais. Busca entender a arquitetura do modelo de rede neural e seus dados internos (mapas de recursos, pesos, vieses e tensores de saída de camadas), permitindo inclusive visualizações em três dimensões. Atualmente com o desenvolvimento pausado, com a última versão disponibilizada em 01/06/2022, oferece opções de versões comerciais com custo de licenciamento (JON, 2021). Na Figura 20 pode ser observada a interface da ferramenta Zetane Viewer mostrando uma visualização de uma rede neural.

Figura 20 – Exemplo de uso da ferramenta Zetane Viewer, exibindo a estrutura da rede neural através de uma visualização 3D, permitindo a análise de cada camada.



Fonte - [Jon \(2021\)](#)

4.6 Comparativo

A Tabela 9 busca sintetizar as informações obtidas sobre as ferramentas de XAI analisadas.

Tabela 9 – Informações obtidas sobre as ferramentas de XAI analisadas.

Ferramenta	Explicação global ou local	Transparente ou posterior/ <i>post-hoc</i>	Específico ou <i>model-agnostic</i>	Linguagem de construção	<i>Framework</i> compatível	Organização apoiadora	Data da última alteração no repositório
LIME	Local	Posterior/ <i>post-hoc</i>	<i>Model-agnostic</i>	JavaScript, Python	Diversos	Nenhuma	29/07/2021
SHAP	Ambas	Posterior/ <i>post-hoc</i>	<i>Model-agnostic</i>	Python	Diversos	Nenhuma	01/07/2024
ELI5	Local	Posterior/ <i>post-hoc</i>	<i>Model-agnostic</i>	Python	Diversos	Nenhuma	12/10/2022
AI Explainability 360	Ambas	Ambos	<i>Model-agnostic</i>	Python	Diversos	IBM	05/03/2024
Skater	Ambas	Posterior/ <i>post-hoc</i>	<i>Model-agnostic</i>	JavaScript, Python	Diversos	Oracle	31/03/2023
Captum	Ambas	Posterior/ <i>post-hoc</i>	<i>Model-agnostic</i>	JavaScript, Python	PyTorch	Facebook	25/06/2024
TruLens	Local	Posterior/ <i>post-hoc</i>	Aprendizado profundo e redes neurais	Python	TensorFlow, PyTorch	Grupo TruEra - Univ. Carnegie Mellon	01/07/2024
VisualDL	Global	Posterior/ <i>post-hoc</i>	<i>Model-agnostic</i>	TypeScript, JavaScript, Python	PaddlePaddle	Baidu	20/09/2023
Netron	Global	Posterior/ <i>post-hoc</i>	Redes Neurais	JavaScript, HTML, Python	Diversos	Nenhuma	02/07/2024
Zetane Viewer	Vi-Global	Posterior/ <i>post-hoc</i>	Redes Neurais	Não informado	Diversos	Zetane - Comercial	Versão em 01/06/2021, sem código disponível

Fonte: Produzida pelo autor.

Como opções com grande flexibilidade, permitindo a utilização com tipo de dados diversos: imagens, textos, dados tabulares, podem ser citadas as ferramentas LIME e SHAP, com destaque para o uso maior da ferramenta SHAP. A ferramenta Eli5 oferece opções mais restritas, focadas em dados textuais, o uso com outros tipos de dados ainda é experimental.

A ferramenta AI Explainability 360 se denomina um toolkit, em português kit de ferramentas, de XAI, unificando diversos algoritmos. Utilizando de forma combinada diversas abordagens, as ferramentas Skater, Captum, TruLens possuem a capacidade de tratar dados de imagem, com diversos algoritmos, porém sendo menos utilizados.

A ferramenta VisualDL oferece visualização dos parâmetros de treinamento e execução do modelo de IA. As ferramentas Netron e Zetane focam na visualização da estrutura das redes neurais.

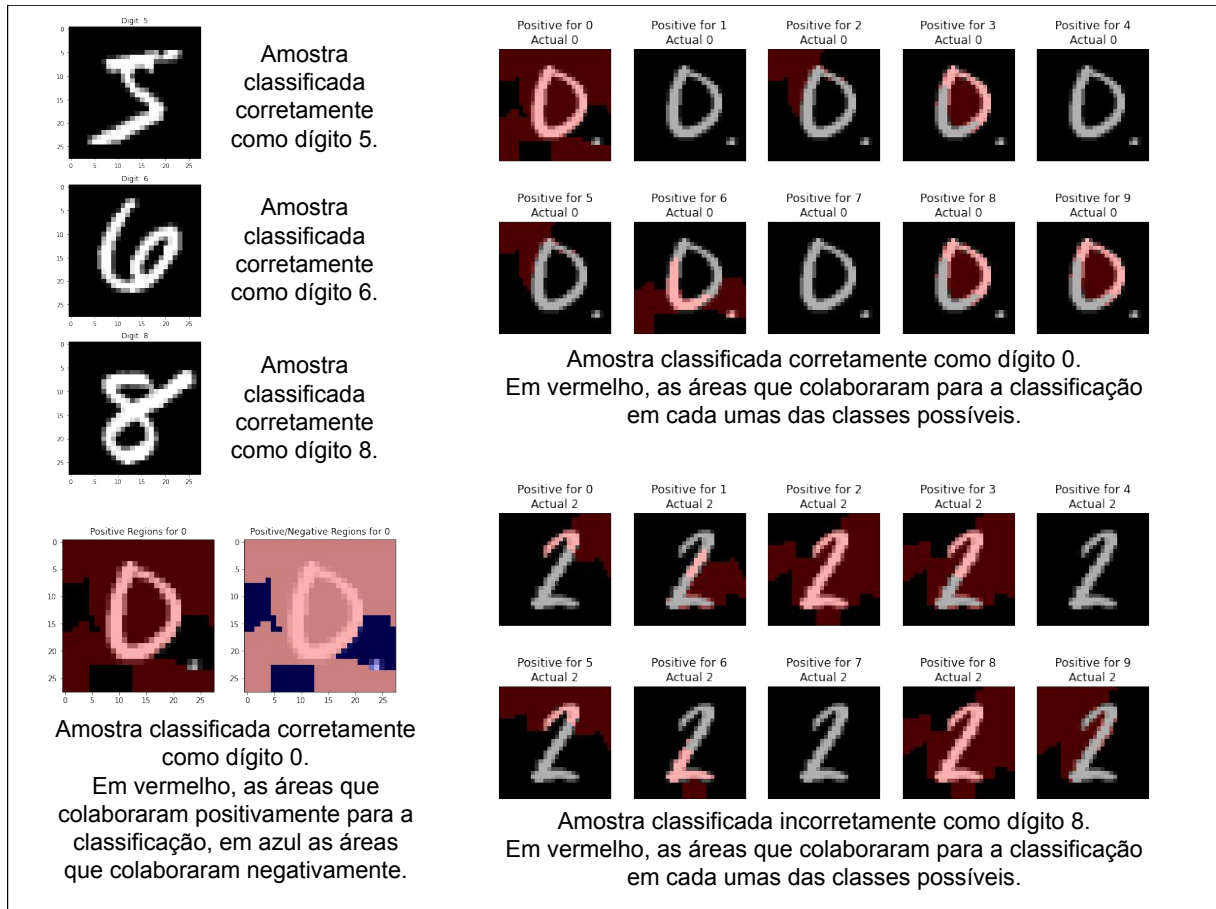
4.7 Exemplos de uso

A partir da identificação dessas diversas ferramentas para uso de XAI, optou-se por realizar testes executando exemplos disponibilizados nos repositórios das bibliotecas LIME e SHAP, pois ambas são as mais populares, têm seu desenvolvimento ativo e são compatíveis com uma grande diversidade de modelos. Os testes foram realizados utilizando a plataforma Google Colab, onde é possível executar os chamados *notebooks*, arquivos que unem anotações a trechos de código executáveis, com código Python em servidores da empresa Google de forma gratuita por meio de um navegador de internet.

4.7.1 Exemplo 1 - LIME

O primeiro exemplo executado como teste foi com a biblioteca LIME, foi utilizado o conjunto de dados MNIST, que possui imagens de caracteres numéricos escritos à mão, com 28 pixels de altura por 28 pixels de largura, totalizando 784 pixels utilizados como características de entrada. O conjunto de dados possui 70.000 amostras, das quais 55%, 38.500 amostras, foram utilizadas como conjunto de treino e 45%, 31.500 amostras, foram utilizadas como conjunto de testes. Como modelo foram utilizadas as florestas aleatórias a partir da biblioteca scikit-learn. O objetivo final foi a classificação de cada amostra em uma das 10 categorias (dígitos de 0 a 9). Na Figura 21 podem ser observados os resultados obtidos nesse cenário.

Figura 21 – Resultados dos experimentos realizados no Exemplo 1, utilizando a biblioteca LIME em um modelo de identificação de caracteres.



Fonte — Produzida pelo autor

O exemplo 1 demonstra uma aplicação de explicabilidade em IA em uma tarefa de classificação de imagens, com a visualização das regiões das imagens que colaboram com as decisões tomadas no processo de classificação efetuado pelo modelo, permitindo à audiência desejada compreender as características das imagens de entrada que levam as classificações, sejam elas corretas ou incorretas.

4.7.2 Exemplo 2 - LIME

O segundo exemplo executado utilizou a biblioteca LIME para gerar explicações de uma rede neural. Como modelo foi utilizado o modelo pré-treinado Inception_V3/GoogleNetv3, disponibilidade no *framework* PyTorch, é composto por uma rede neural convolucional treinada na base de dados de imagem Imagenet a partir de 2015. O objetivo do experimento é a classificação de imagens, para isso foi aplicado o modelo na classificação de uma amostra de imagem contendo um cachorro e um gato. Na Figura 22 podem ser observados os resultados obtidos nesse cenário.

Figura 22 – Resultados dos experimentos realizados no Exemplo 2, utilizando a biblioteca LIME em um modelo de identificação de imagens.



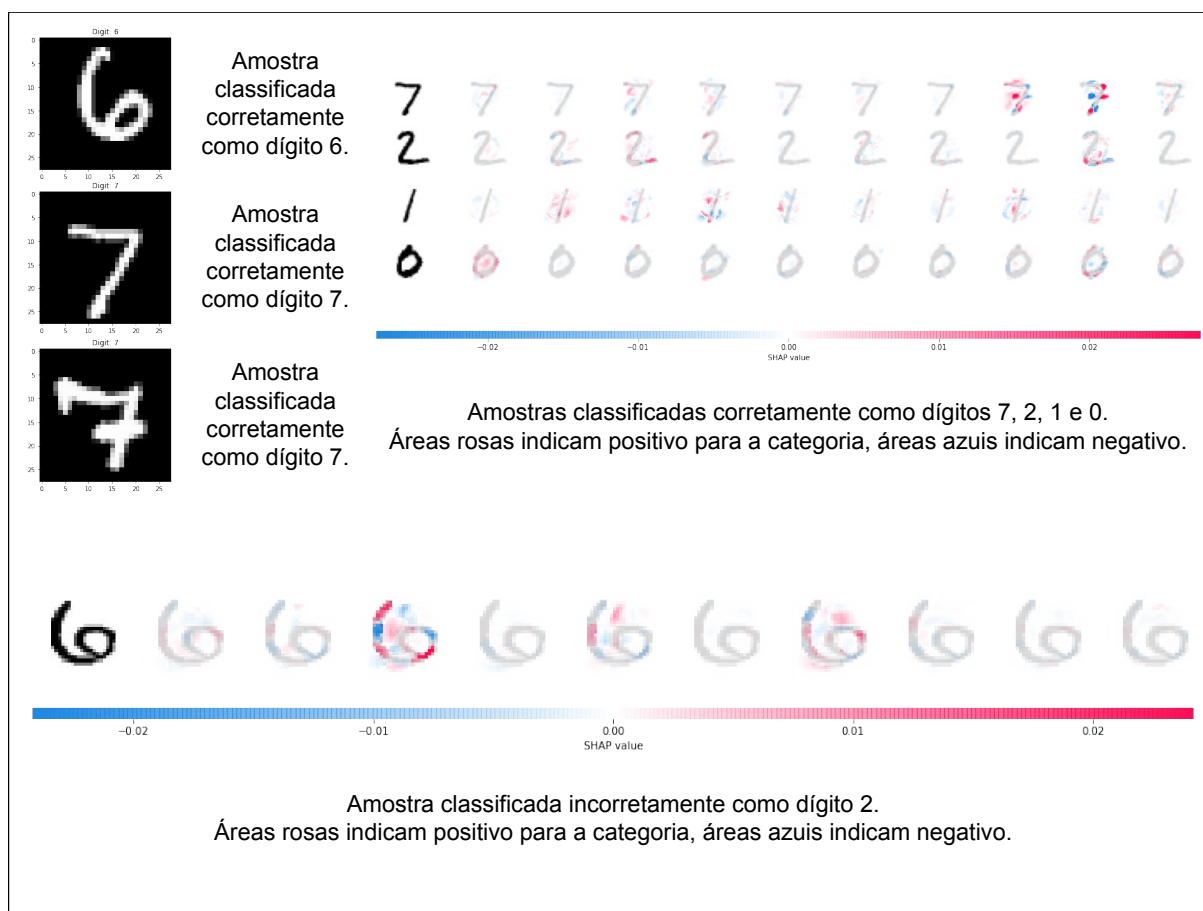
Fonte — Produzida pelo autor

O exemplo 2 trata da classificação das imagens, obtendo uma classe que identifica o cachorro presente na imagem e sua raça, por meio do modelo pré-treinado. Com relação à explicabilidade, são demonstradas as áreas que colaboraram para a classificação, acrescentando informações sobre o processo de decisão do modelo.

4.7.3 Exemplo 3 - SHAP

O terceiro exemplo executado como teste foi com a biblioteca SHAP, foi utilizado o conjunto de dados MNIST, novamente com 70.000 amostras, das quais 85,71%, 60.000 amostras, foram utilizadas como conjunto de treino e 14,28%, 10.000 amostras, foram utilizadas como conjunto de testes. Como modelo foi utilizada uma rede neural convolucional a partir da biblioteca keras. O objetivo final era a classificação da amostra em uma das 10 categorias (dígitos de 0 a 9). Na Figura 23 podem ser observados os resultados obtidos nesse cenário.

Figura 23 – Resultados dos experimentos realizados no Exemplo 3, utilizando a biblioteca SHAP em um modelo de identificação de caracteres.



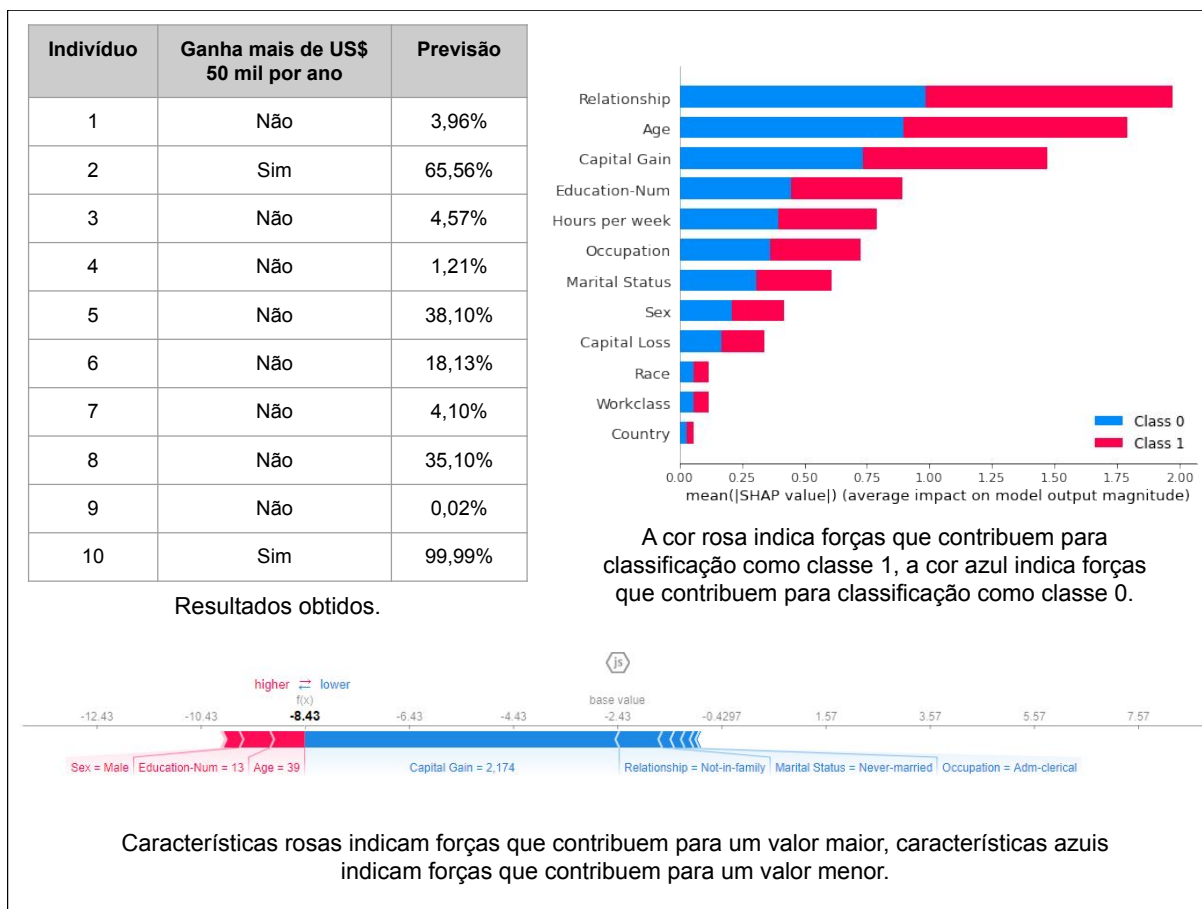
Fonte — Produzida pelo autor

Para este terceiro exemplo, são obtidas informações de explicabilidade semelhantes ao do primeiro exemplo, com a diferença da utilização da biblioteca SHAP ao invés de LIME. Ainda que o foco do SHAP seja a explicação global do modelo, é possível explicar cada amostra individualmente, como neste exemplo.

4.7.4 Exemplo 4 - SHAP

O quarto exemplo utiliza a biblioteca SHAP e busca demonstrar uma aplicação de XAI além das classificações de dados tabulares. Foi utilizado um conjunto de dados de censo fornecidos pela própria biblioteca SHAP como exemplo. O conjunto contém 32.561 amostras com 12 características, foram utilizadas 80%, 26.048 amostras, no conjunto de treino e 20%, 6.513 amostras, no conjunto de testes. O objetivo é determinar a probabilidade do indivíduo de cada amostra ganhar mais de US\$ 50 mil por ano, para isso foi utilizado o modelo LightGBM (*Light Gradient Boosting Machine*), baseado em árvores de decisão. Na Figura 24 podem ser observados os resultados obtidos nesse cenário.

Figura 24 – Resultados dos experimentos realizados no Exemplo 4, utilizando a biblioteca SHAP em um modelo de previsão que utiliza dados tabulares.



Fonte — Produzida pelo autor

É possível obter detalhes sobre o comportamento geral do modelo, identificando quais características colaboram para as classificações, positiva e negativamente, para cada uma das classes.

Tanto o LIME quanto o SHAP são independentes do modelo e podem fornecer a explicação para qualquer modelo. Eles constroem seus próprios modelos substitutos e fazem previsões sobre eles. Ambos tratam os modelos como uma caixa preta. As explicações são fornecidas sobre como o modelo se comporta para diferentes entradas em torno do ponto de dados, sem entrar nas partes internas do modelo.

LIME é mais rápido em comparação com SHAP, pois pode levar algum tempo para que os valores de Shapley sejam calculados. Como o SHAP usa todos os subconjuntos dos dados de entrada, considera-se que tem uma melhor precisão e consistência local.

Zhou et al. (2021) afirmam: "Embora vários métodos de explicação tenham sido explorados, há uma necessidade de avaliações para quantificar a qualidade dos métodos de explicação para determinar se e em que medida a explicabilidade oferecida atinge o objetivo definido e comparar os métodos de explicação disponíveis e sugerir a melhor

explicação a partir da comparação para uma tarefa específica."

4.7.5 Considerações finais do capítulo

Neste capítulo foram apresentadas as principais opções de ferramentas para uso de XAI e os experimentos com exemplos para familiarização com as bibliotecas LIME e SHAP. A seguir, será aprofundado o estudo no processo de avaliação de XAI, buscando identificar se as explicações são efetivas e satisfatórias.

5 Avaliação de Métodos de Inteligência Artificial Explicável

5.1 Considerações iniciais do capítulo

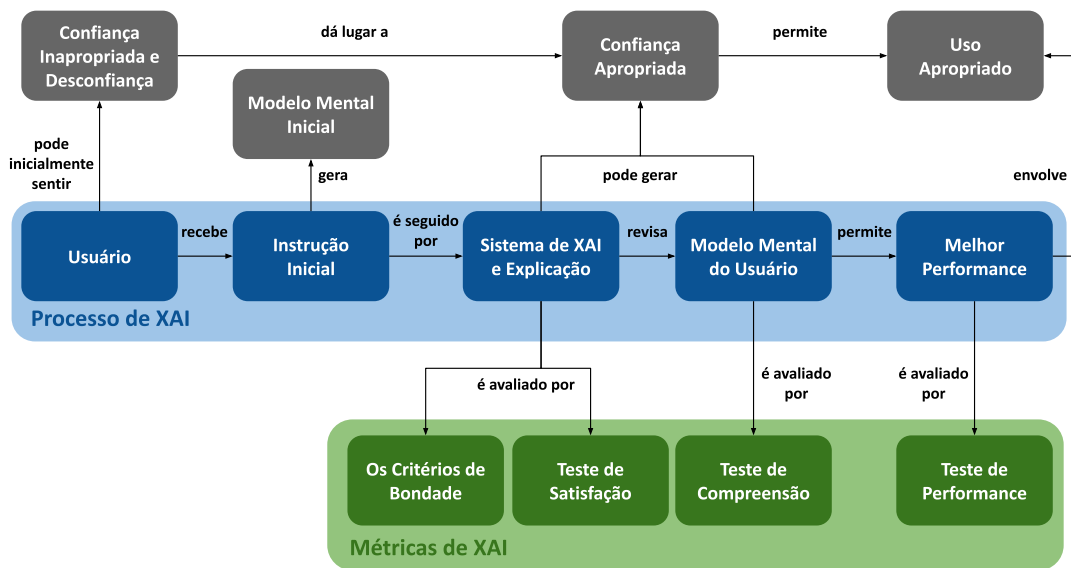
No Capítulo 2, o desafio de propor métricas de avaliação para XAI foi brevemente apresentado, em especial para modelos de aprendizado profundo. No presente capítulo o assunto será detalhado, buscando nortear o desenvolvimento do presente projeto, direcionando ele para essa limitação da área de pesquisa.

No cenário de pesquisas de XAI, percebe-se que diferentes pesquisadores, de diversas disciplinas, acabam por ter objetivos distintos ao tratar sobre XAI. Ainda que todos sigam aspectos gerais que definem a XAI, cada disciplina usa diferentes métricas para avaliar se uma XAI atende os objetivos desejados. Pesquisadores de aprendizado de máquina tendem a utilizar métodos analíticos numéricos, enquanto pesquisadores de visualização de dados e interface humano-computador tendem a realizar avaliação mais ligadas aos aspectos humanos e de usabilidade. Por mais que possa parecer que essas abordagens sejam incompatíveis, para um benefício maior ao utilizar XAI, deve-se buscar uma convergência nesses aspectos de avaliação (MOHSENI; ZAREI; RAGAN, 2018) (LOPES et al., 2022).

5.2 O processo de entendimento

O trabalho de Hoffman et al. (2018) se aprofunda nesta questão da falta de métricas, propondo a seguinte questão: "Se apresentarmos a um usuário um sistema de IA que explica como funciona, como saberemos se a explicação funciona e se o usuário alcançou uma compreensão pragmática da IA?".

O processo de entendimento começa geralmente com uma pequena instrução inicial que leva o usuário a formar um modelo mental inicial, em seguida, por meio do sistema de XAI, esse modelo pode ser aprimorado, aumentando a confiança e o desempenho na ferramenta de IA. Na Figura 25 pode ser observado um modelo conceitual para todo processo de geração da explicação por meio de XAI.

Figura 25 – Diagrama mostrando um modelo conceitual do processo de explicação, no contexto do XAI.

Fonte — Adaptado de [Hoffman et al. \(2018\)](#).

Boas e satisfatórias explicações permitem que usuários aprimorem seu modelo mental, que por consequência, permite que os usuários desenvolvam a confiança adequada na IA e essa passe a ter um bom desempenho. Podem ser destacadas quatro classes de medidas para o processo de XAI: os critérios de bondade, que determinam o quão boa é a explicação e os testes de satisfação, compreensão e desempenho.

O processo de explicação se mostra interativo, dependendo das necessidades do usuário, dos conhecimentos prévios que o mesmo possui e principalmente dos seus objetivos, com isso chega-se a seguinte pergunta: "Por que um determinado usuário precisa de uma explicação?". Ao se olhar para toda literatura e pesquisa sobre explicação e explicabilidade produzida anteriormente, encontram-se muitas afirmações sobre como é formada uma boa explicação, a partir dos mais diversos pontos de vista, entretanto é de comum acordo que boas explicações estão relacionadas a clareza e precisão.

Ainda que existam características comuns a todas as explicações consideradas boas, determinada explicação pode não ser considerada satisfatória ou adequada em um determinado contexto, ou para um determinado grupo de usuários. O conceito de satisfação é definido como o grau onde os usuários sentem que entenderam o sistema ou processo de IA alvo da explicação, dessa forma o julgamento de satisfação é posterior a explicação e depende do contexto.

5.3 A curiosidade

Diversas razões levam a curiosidade a ser considerada um fator importante ao se tratar de XAI. Na maioria das situações, o ato de buscar uma explicação é movido pela curiosidade, da mesma forma que explicações podem levar o usuário a ter curiosidade, buscando melhorar sua percepção e desenvolver um modelo mental mais aprimorado. Dessa forma, a avaliação da curiosidade pode ser bem-vinda na avaliação de sistemas com XAI (HOFFMAN et al., 2018).

Porém, as explicações podem suprimir a curiosidade reforçando modelos mentais falhos, isso pode acontecer com uma explicação que sobrecarrega a pessoa com detalhes, com um sistema de XAI que não permite ou dificulta a realização de perguntas, com explicações que fazem a pessoa hesitar e ficar desconfortável com sua falta de conhecimento ou com explicações abertas demais, causando confusão e aumentando a complexidade da situação (HOFFMAN et al., 2018).

5.4 Medidas de avaliação de Inteligência Artificial Explicável

Segundo Mohseni, Zarei e Ragan (2018), 5 aspectos podem ser medidos na avaliação de XAI, nas seções a seguir eles serão brevemente detalhados.

5.4.1 Modelo mental

Pode-se considerar a explicabilidade uma metodologia para obter, representar e analisar modelos mentais de usuários de IA. A psicologia cognitiva define os modelos mentais como representações ou expressões de como um indivíduo entende algum tipo de evento, processo ou sistema (KLEIN; HOFFMAN, 2008). Dessa forma, no contexto de XAI, o modelo mental pode ser definido como a compreensão do usuário sobre o sistema de IA.

O sistema de XAI deve permitir que o avaliador possa compreender o que é positivo ou negativo no modelo mental do usuário, de forma que o usuário também possa compreender o que é bom ou limitado dentro do seu entendimento (MOHSENI; ZAREI; RAGAN, 2018).

5.4.2 Explicação de utilidade e satisfação

A utilidade da explicação e a satisfação do usuário final são importantes ao avaliar explicação de XAI. Existem diversas medidas, tanto objetivas quanto subjetivas, que os pesquisadores usam para avaliar compreensão, utilidade e quantidade de detalhes de uma explicação e seu valor para os usuários. Ainda que existam métodos implícitos para medir a satisfação do usuário, a literatura utiliza quase sempre uma avaliação qualitativa da

satisfação em explicações, como questionários e entrevistas ([MOHSENI; ZAREI; RAGAN, 2018](#)).

5.4.3 Confiança e confiabilidade do usuário

Certos usuários confiam em máquina, dessa forma consideram dados e afirmações vindas dos computadores como válidas e verdadeiras simplesmente por serem originados de computadores, esse tipo de usuário geralmente já possui um certo grau de afinidade com tecnologia. Porém, outros usuários podem exigir uma justificativa para confiarem em computadores. Sendo assim, existem diferentes formas e graus de confiança e desconfiança e isso se aplica também à tecnologia. A confiança em sistemas automatizados pode se perder rapidamente em condições com muitas falhas ou alta frequência de alarmes falsos e ainda a confiança em máquinas pode ser difícil de ser recobrada quando perdida ([HOFFMAN et al., 2018](#)).

5.4.4 Desempenho de tarefa Humano-Inteligência Artificial

Como um dos objetivos da XAI é auxiliar os usuários finais a serem bem-sucedidos em tarefa que utilizam sistemas com aprendizado de máquina, dessa forma, o desempenho em uma tarefa Humano-IA é relevante independente do tipo do usuário, seja ele um desenvolvedor de IA, especialista no contexto ou usuário final. O desempenho de uma explicação pode se relacionar ao julgamento da exatidão da saída de um sistema. As explicações também ajudam os usuários a depurar programas de aprendizado de máquina que sejam interativos ([MOHSENI; ZAREI; RAGAN, 2018](#)).

5.4.5 Medidas computacionais

Medidas chamadas de computacionais são frequentemente utilizadas para avaliar a exatidão e integridade das técnicas de interpretabilidade, buscando demonstrar o que o modelo aprendeu. Esse tipo de método de avaliação acaba sendo utilizado por especialistas em aprendizado de máquina e ciência de dados para avaliar a precisão de métodos de explicabilidade e modelos treinados, incluindo a qualidade do treinamento, extraindo dados estatísticos durante os processos de treino e teste ([MOHSENI; ZAREI; RAGAN, 2018](#)).

5.5 Como avaliar

A forma mais direta e efetiva de se avaliar o desempenho de um sistema de XAI é medir quão fácil ou difícil é fazer a parte interessada (potencial usuário) adotar o uso do sistema. Discutindo sistemas que buscavam realizar um diagnóstico médico precoce, [Lent, Fisher e Mancuso \(2004\)](#) afirmaram: "No início, os desenvolvedores desses sistemas

perceberam que os médicos não estavam dispostos a aceitar o diagnóstico do sistema especialista pela fé", essa questão levou ao desenvolvimento dos primeiros sistemas de IA com explicabilidade (HOFFMAN et al., 2018).

Por mais que existam muitas propostas para avaliação de XAI, a maioria dos estudos trata de propostas para situações bem específicas, ainda que todo estudo colabore para área de pesquisa e possa trazer reflexões úteis. Entretanto, antes de propor uma nova técnica, devem ser avaliadas as condições em que a técnica proposta se mostra superior as demais técnicas existentes, ou se uma modificação feita em um método particular melhora ou piora algum aspecto da técnica ou sistema existente (SISK et al., 2022).

Muitas formas de avaliação de ferramentas XAI tem origem e base no campo de sistemas interativos, interação humano-computador e sistemas de visualização de dados e informações. Um ponto a se levar em conta é que uma única métrica dificilmente servirá para contar uma história e para que a história, ou explicação no caso, seja relevante e completa, um conjunto de abordagens e métricas deve ser utilizado, principalmente pela complexidade do contexto (WARD; GRINSTEIN; KEIM, 2010) (LOPES et al., 2022).

5.6 Considerações finais do capítulo

Neste capítulo buscou-se aprofundar a discussão sobre avaliação da explicabilidade e das ferramentas de XAI. A partir de todos os pontos abordados até o momento, toda fundamentação e revisão, as análises e testes com ferramentas de XAI, a discussão sobre avaliação, no próximo capítulo será apresentado um cenário para desenvolvimento de uma estratégia de uso de XAI.

6 Estratégia de Uso

6.1 Considerações iniciais do capítulo

No presente capítulo é apresentada a estratégia de uso de XAI que compõe este estudo, será inicialmente descrito os estudos preliminares realizados e posteriormente o cenário onde a ela está inserida e o trabalho a ser utilizado como base, para com isso apresentar a proposta da estratégia que foi desenvolvida.

6.2 Estudos preliminares

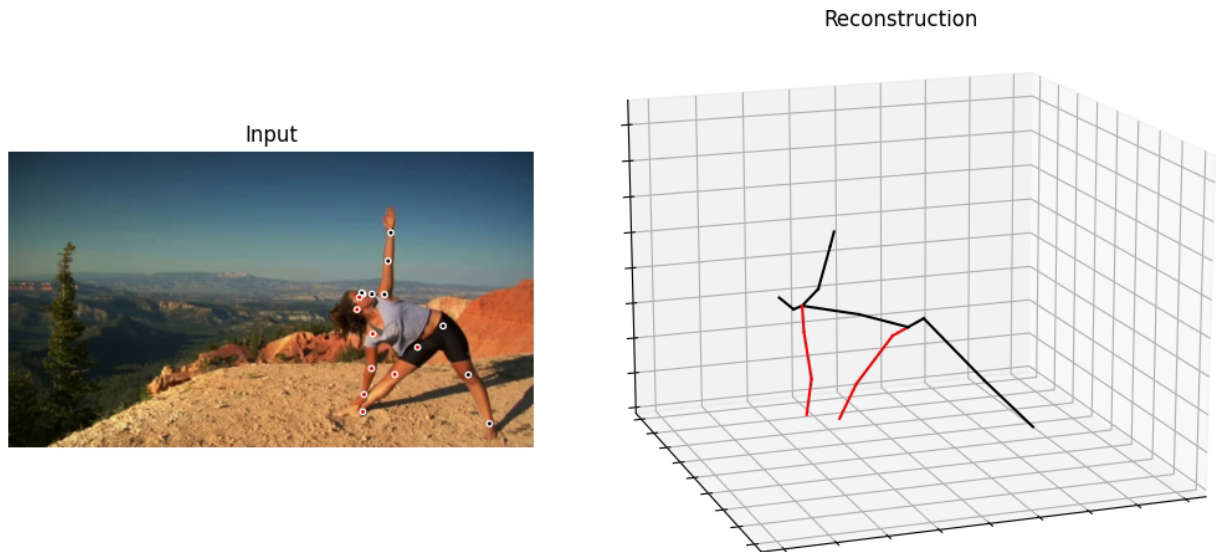
Foram localizados estudos que buscassem implementar tarefas semelhantes as que se desejava realizar e através aumentar o conhecimento sobre as soluções existentes.

Diversos autores buscaram aplicar IA na estimativa, classificação e avaliação de movimentos. No trabalho de [Chen e Yang \(2020\)](#), foi proposta a ferramenta Pose Trainer, utilizando redes neurais convolucionais e técnicas heurísticas para detectar a pose de um usuário e, em seguida, avaliar a geometria vetorial da pose. O Pose Trainer funciona em quatro exercícios comuns e suporta qualquer computador Windows ou Linux com uma GPU. [Dittakavi et al. \(2022\)](#) desenvolveram o Pose Tutor, um sistema de reconhecimento e correção de pose explicável baseado em IA, esse sistema utiliza diversos modelos de IA para estimar a pose, classificar o movimento e indicar por ferramentas de explicabilidade os pontos a serem corrigidos na pose, para a estimativa de pose a ferramenta usa o modelo TransPose, baseado na solução de redes neurais com Transformers ([YANG et al., 2020](#)). Pose Tutor mostra bons resultados dados de pose de Yoga, Pilates e Kung Fu.

Para os estudos de estimativa de pose, foram realizados experimentos com as soluções propostas por [Yang et al. \(2020\)](#) e [Pavlo et al. \(2019\)](#). A solução de [Pavlo et al. \(2019\)](#) aplica estimativa de pose em dados de vídeo através da biblioteca Detectron2, que utiliza um modelo de rede neural convolucional para diversas tarefas de reconhecimento visual, entre elas estimativa de pose ([WU et al., 2019](#)). Ambas as soluções, VideoPose3D e Detectron2, foram desenvolvidas e são mantidas pela equipe de pesquisa Meta Research.

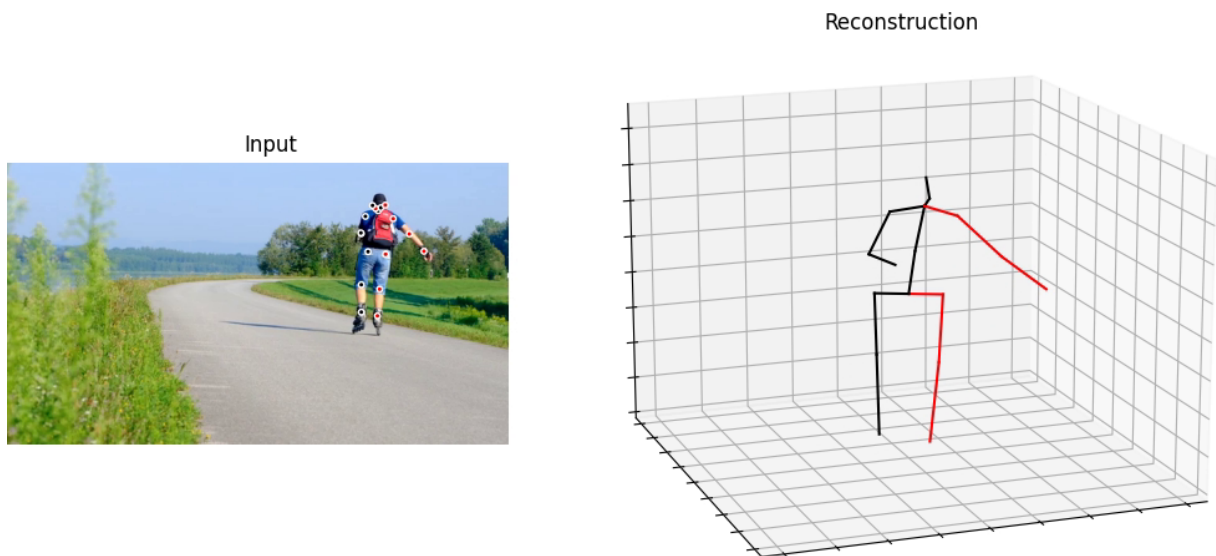
Nas Figuras 26 e 27 podem ser observados resultados da aplicação VideoPose3D em dois vídeos selecionados, obtendo assim a estimativa de posição das articulações, pose e reconstrução em 3D da pose.

Figura 26 – Exemplo de teste do modelo de estimativa de pose em 3D, com a exibição de um avatar executando os movimentos detectados.



Fonte — Produzida pelo autor.

Figura 27 – Exemplo de teste do modelo de estimativa de pose em 3D, com a exibição de um avatar executando os movimentos detectados.



Fonte — Produzida pelo autor.

Foram realizados estudos sobre medição de ângulos das articulações detectadas em uma pose, buscando a partir dessa informação avaliar a qualidade do movimento executado. Foram realizados testes com um protótipo baseado nas tecnologias OpenCV e MediaPipe, usadas respectivamente para visão computacional e detecções e rastreamentos em entradas de vídeos (TEAM, 2023).

Uma prova de conceito criada permitiu a entrada de informações de vídeo e após a estimativa de pose, fornece o valor do ângulo de flexão do tronco do indivíduo, como

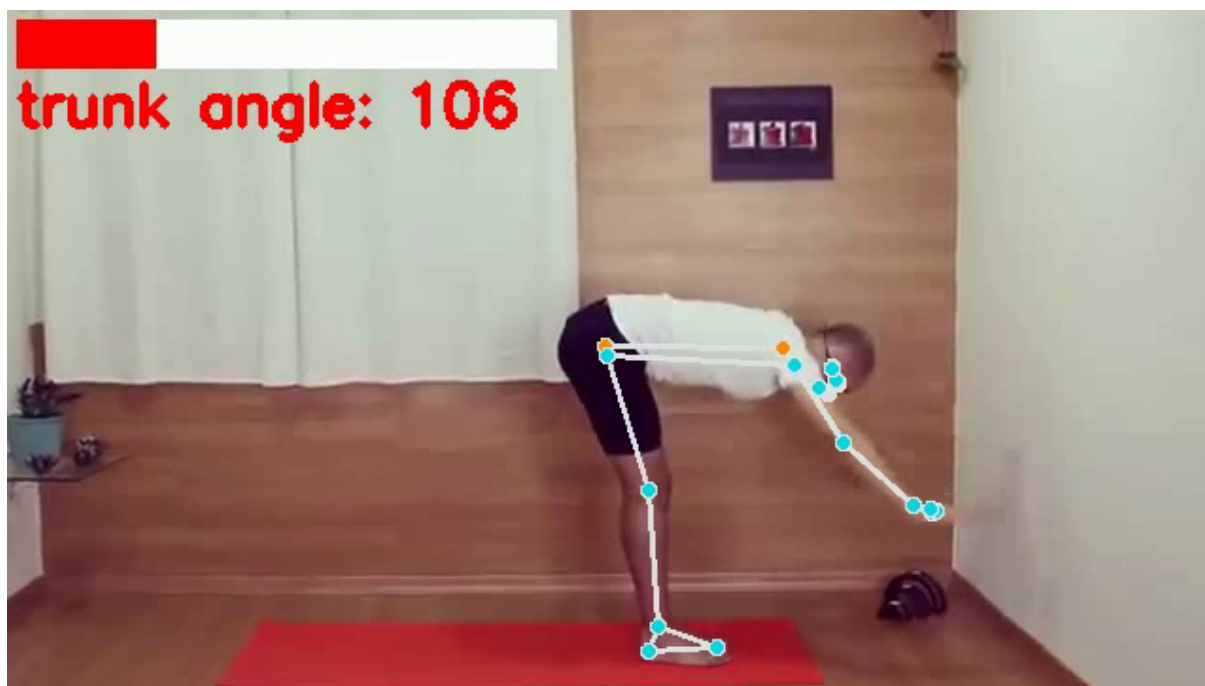
Figura 28 – Exemplo de teste do modelo de avaliação de ângulo, com a exibição de um avatar sobreposto ao vídeo original e a estimativa do ângulo de flexão do tronco.



Fonte — Produzida pelo autor.

pode ser observado nas Figuras 28 e 29.

Figura 29 – Exemplo de teste do modelo de avaliação de ângulo, com a exibição de um avatar sobreposto ao vídeo original e a estimativa do ângulo de flexão do tronco.



Fonte — Produzida pelo autor.

6.3 Cenário

O cenário onde será aplicada a explicabilidade é a aplicação de IA na classificação de movimentos obtidos por meio de rastreamento corporal, para utilização de dados na reabilitação de pacientes, em especial os que sofreram acidente vascular cerebral.

A aplicação de XAI se mostra relevante nesse contexto por se tratar de uma aplicação na área de saúde, que demanda confiança nas decisões. Considerando a implantação clínica de um processo de decisão baseado em IA, é importante a capacidade de se compreender e validar o método de IA, nesse contexto, a XAI visa fornecer explicações compreensíveis (LENT; FISHER; MANCUSO, 2004 apud NAYEBI et al., 2022).

Interações em ambientes virtuais e jogos são geralmente gravadas em vídeos, para avaliação posterior do fisioterapeuta, entretanto armazenar somente vídeos dificulta o tratamento de dados, inclusive utilizando IA, para geração de conhecimento que auxiliem os profissionais da área da saúde. Diante disso, foi desenvolvida uma solução para produção de dados de movimentação por meio de dispositivos com câmeras, como celulares, extraindo dados relativos aos movimentos e armazenando em um servidor remoto, sem a necessidade do armazenamento do vídeo (TROTТА et al., 2021). Além da solução com a captura através de imagens, o sistema de armazenamento possibilita a aquisição de dados através de sensores de captura de movimento, chamados biomecânicos, que geram rastreamento

corporal e podem interagir com ambientes de realidade virtual (RIBEIRO et al., 2021).

Por meio dos dados extraídos, é possível realizar a avaliação da amplitude dos movimentos dos membros do paciente, e conforme essa avaliação é realizada durante a evolução do tratamento, pode servir para auxiliar o fisioterapeuta nas suas decisões e avaliar os resultados obtidos. Para auxiliar no tratamento desses dados, foram propostas estratégias utilizando aprendizado de máquina para geração de conhecimento, auxiliando em diversos aspectos o tratamento fisioterápico. Do ponto de vista do fisioterapeuta, ao auxílio da tecnologia pode auxiliar na comparação do estado atual do paciente com outros pacientes disponíveis no banco de dados, podendo trazer até sugestões para o tratamento (RODRIGUES et al., 2021a).

Os autores Rodrigues et al. (2021a) vêm trabalhando nesse contexto junto ao Instituto Brasileiro de Neurociências e Neurotecnologia - o CEPID BRAINN, utilizando IA na classificação de movimentos, porém ainda sem envolver aspectos de XAI, dessa forma, um dos objetivos deste estudo é adicionar essa abordagem para colaborar no desenvolvimento da solução e acrescentar explicabilidade em diversos pontos, fornecendo novas informações para diferentes públicos, entre eles: usuários da solução, especialistas no domínio, desenvolvedores da IA, obtendo assim aspectos de explicabilidade bem abrangentes, com diversas possibilidades.

6.4 Classificação de movimentos com IA

Foram realizados testes estendendo as funcionalidades do trabalho realizado pelo autor Rodrigues et al. (2022) em conjunto com o CEPID BRAINN, onde originalmente se realizava estimativa de pose e classificação de movimentos, para se obter informações relacionadas a explicabilidade. Os testes se mostraram complexos, porém com caminhos promissores, com possibilidade de se associar as articulações da pose ao impacto que cada uma causa na classificação.

O processo de captura inicial das informações pode ser realizado por meio de diversos dispositivos para captura de movimento (Microsoft Kinect, pulseiras, webcams, câmeras de celulares). As informações de vídeo são primeiramente tratadas pela ferramenta Mediapipe, transformando elas em dados de articulações (LUGARESI et al., 2019). Existe um processo de normalização dos dados, uma vez que o movimento é medido em *pixels* e a escala dessa medida varia conforme a resolução do dispositivo de captura. Outro processo de normalização é realizado, definindo um período padrão para o movimento, com uma taxa de 30 ou 90 quadros por segundo. Como resultado desse processo de pré-processamento, é obtido um vetor armazenando a informação de um movimento, que será utilizado como entrada para o modelo (RODRIGUES et al., 2021b). Os dados de movimento padronizados são armazenados em um banco de dados desenvolvido especificamente para esse padrão,

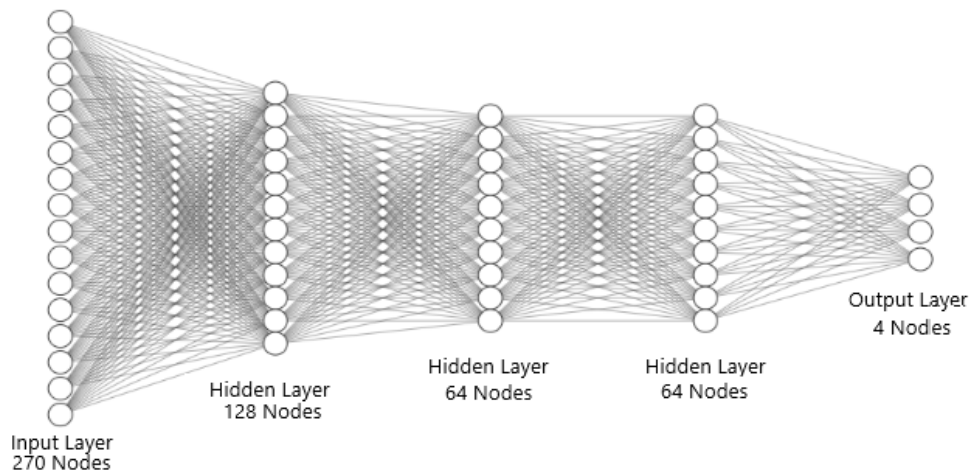
compondo todo um conjunto de aplicações para dados de reabilitação (TROTTA et al., 2021).

A solução de IA utilizada para classificação de movimentos é um modelo de rede neural artificial baseado na biblioteca TensorFlow versão 2 por meio dos métodos disponíveis pela biblioteca Keras. A rede é composta por uma primeira camada totalmente conectada, conforme o tamanho dos dados de entrada, para essa camada foi utilizada a função de ativação ReLU (*rectified linear unit*), buscando tornar a rede esparsa e eficiente e de fácil computação, reduzindo a necessidade de processamento. A camada de saída possui o número de nós conforme o número de classes de saída possíveis, nela é utilizada a função de ativação Softmax, que transforma os vetores resultantes em probabilidades ligadas a cada classe, sendo muito utilizada em cenários de classificação.

Entre as camadas de entrada e saída, existem as camadas ocultas com *dropout*, que buscam prevenir o *overfitting*, situação em que existe o sobreajuste do modelo, prejudicando os resultados alcançados. Essas camadas atuam nos dados de treino, retirando uma fração dos neurônios aleatória e temporariamente, resultando em um processo de regularização (HINTON et al., 2012). Para essas camadas também foi utilizada a função de ativação ReLU. Para a camada de entrada e camadas ocultas, os pesos iniciais foram inicializados de forma aleatória por meio do parâmetro "he_uniform".

Para o processo de treinamento se definiu a aproximação dos resultados por meio de gradiente descendente estocástico (SGD) parametrizado com taxa de aprendizado de 0,001 e coeficiente de *momentum* de 0,9, ajustando a aceleração do gradiente descendente. Como função de custo foi utilizada a entropia cruzada, indicada para situação onde se deseja obter categorias. A métrica utilizada para avaliação do modelo foi a precisão. O treinamento foi definido para durar 100 épocas, o número de amostras por atualização de gradiente foi definido para 10 por meio do parâmetro "batch_size". A Figura 30 demonstra a arquitetura da rede neural e suas camadas para um problema de classificação de 4 movimentos.

Figura 30 – Diagrama mostrando a arquitetura de um dos modelo de rede neural utilizados nos estudos de (RODRIGUES et al., 2021a).



Fonte — Rodrigues et al. (2021a)

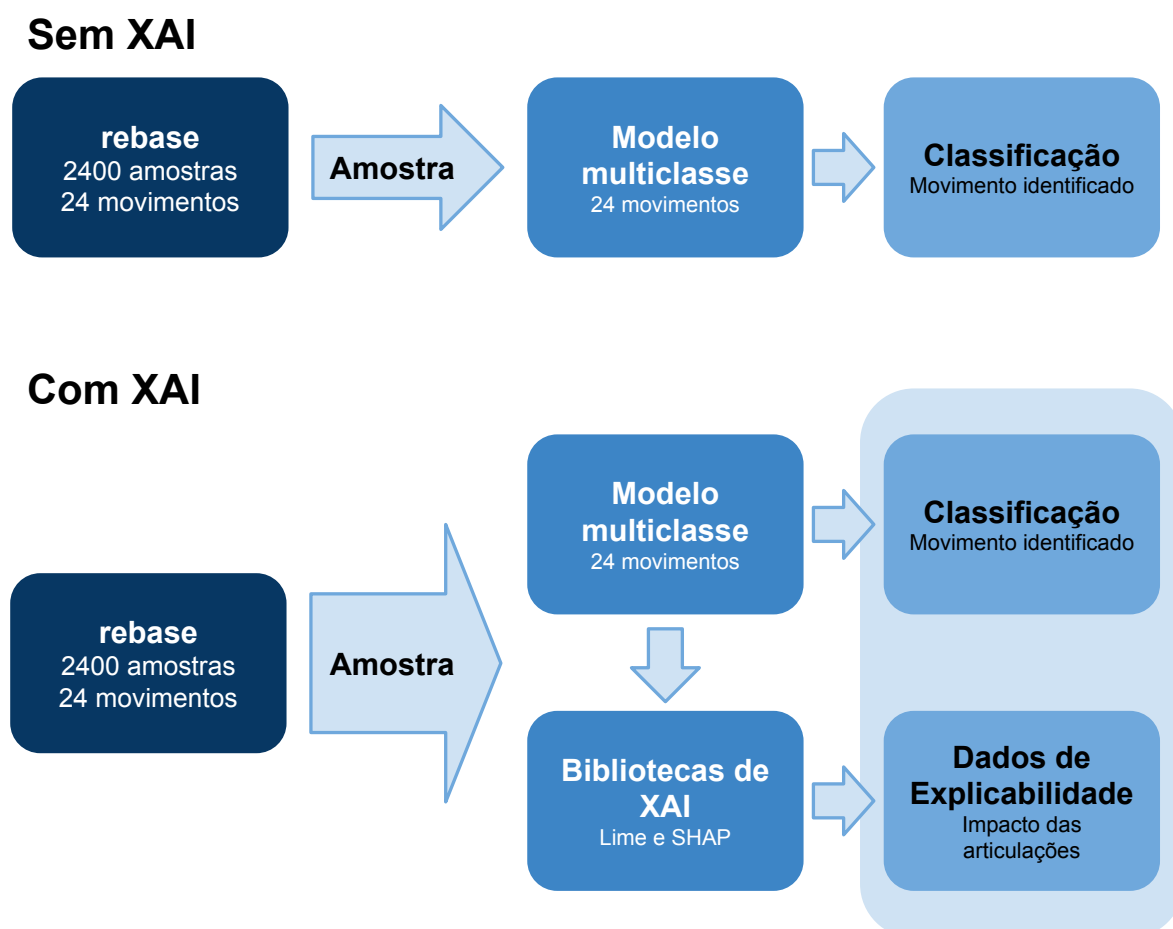
6.5 Estratégia

Considerando o uso da explicabilidade como ferramenta de apoio ao usuário final da aplicação, como nos exemplos dos trabalhos apresentados anteriormente, o objetivo da estratégia é utilizar a explicabilidade como ferramenta para avaliação dos movimentos apresentados nas atividades de reabilitação, permitindo a avaliação do movimento, comparando com a proposta do profissional de saúde. Para desenvolvimento desta solução, foi utilizada a classificação de movimentos, juntamente com aspectos de explicabilidade, buscando obter um retorno visual com a avaliação do movimento executado.

De modo a explorar a explicabilidade por meio da aplicação de XAI, foram realizados experimentos utilizando as bibliotecas LIME e SHAP junto a solução de IA desenvolvida para classificação de movimentos na plataforma de programação Jupyter Notebook, que permite o desenvolvimento interativo através da linguagem Python direto do navegador. Optou-se pela utilização de bibliotecas que utilizam a chamada explicabilidade *post-hoc* para se obter a explicabilidade de maneira posterior, por meio de modelos que não possuem transparência em sua concepção. Por possuírem abordagens diferentes que resultam por vezes em aspectos de explicabilidade igualmente importantes, optou-se em utilizar LIME e SHAP de forma conjunta, ao invés de definir uma única ferramenta para ser utilizada.

Na Figura 31 é demonstrado o fluxo de dados na solução existente e a proposta inicial de integração de XAI no fluxo. O objetivo é associar a saída do modelo dados de explicabilidade além da classificação retornada.

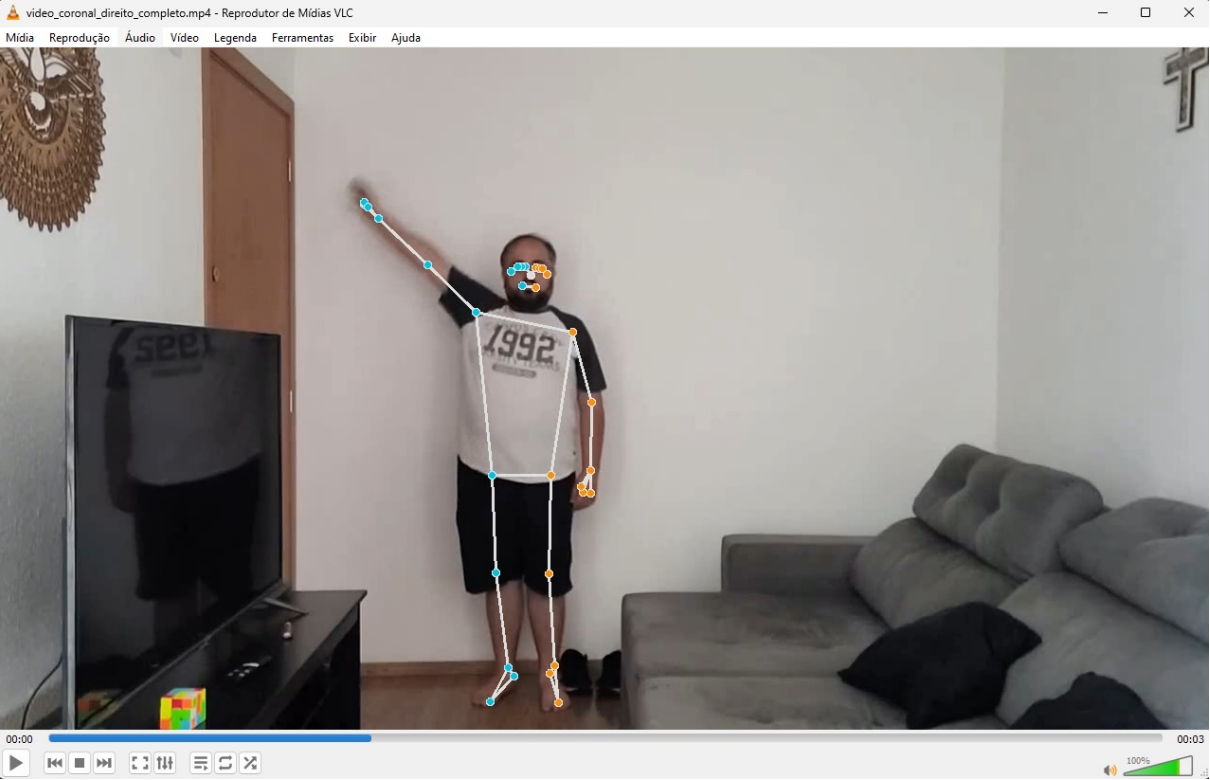
Figura 31 – Diagramas mostrando o fluxo de dados para classificação de movimentos com a ferramenta existente e proposta inicial de utilização de XAI.



Fonte — Produzida pelo autor.

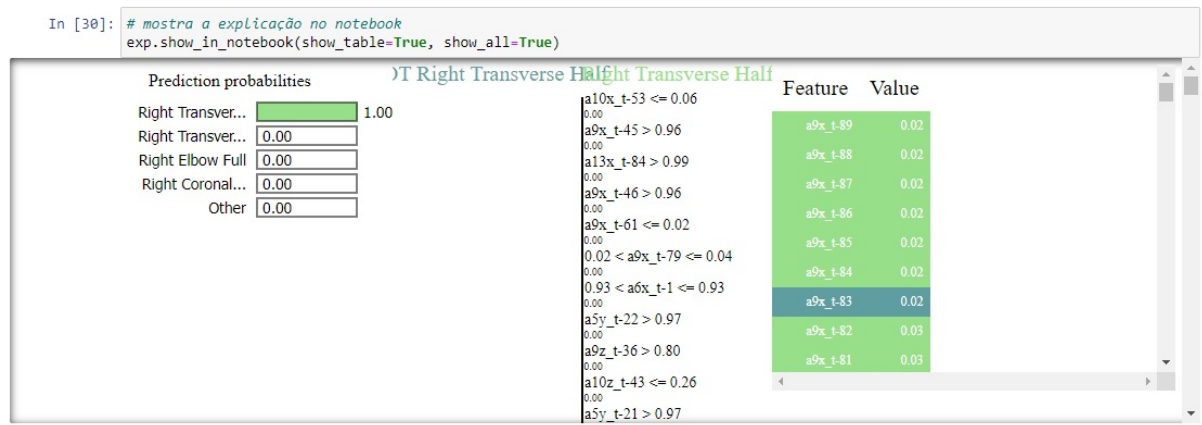
Nas Figuras 32, 33, 34 e 35 podem ser observados diversos testes executados, desde o rastreamento de novos vídeos com o *framework* Mediapipe, aplicação das bibliotecas de XAI LIME e SHAP no modelo de IA selecionado e a manipulação dos dados obtidos através dessas bibliotecas.

Figura 32 – Captura de tela mostrando a aplicação do rastreamento de pose em novo vídeo produzido para o estudo atual.



Fonte — Produzida pelo autor.

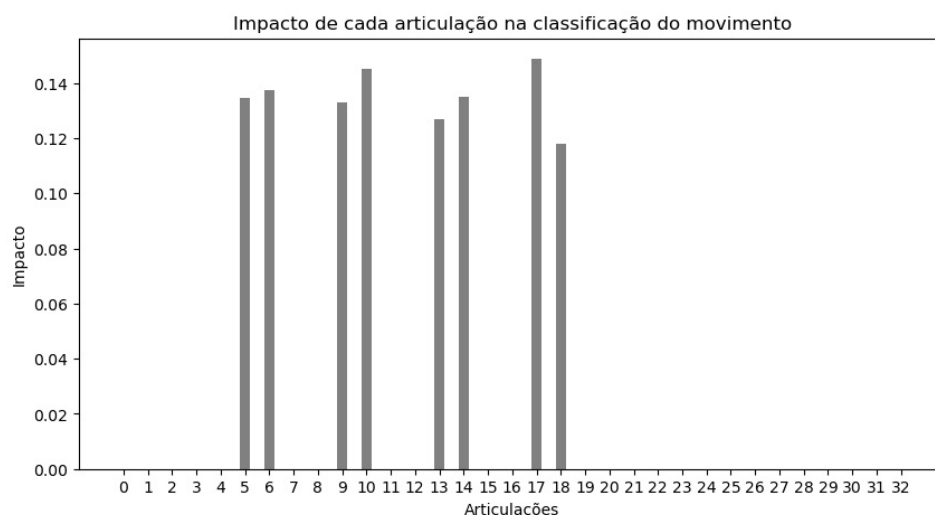
Figura 33 – Captura de tela mostrando os resultados dos testes executados com a biblioteca LIME, com o impacto de cada articulação em cada eixo e em cada quadro do vídeo para a classificação correta do movimento em uma amostra.



Fonte — Produzida pelo autor.

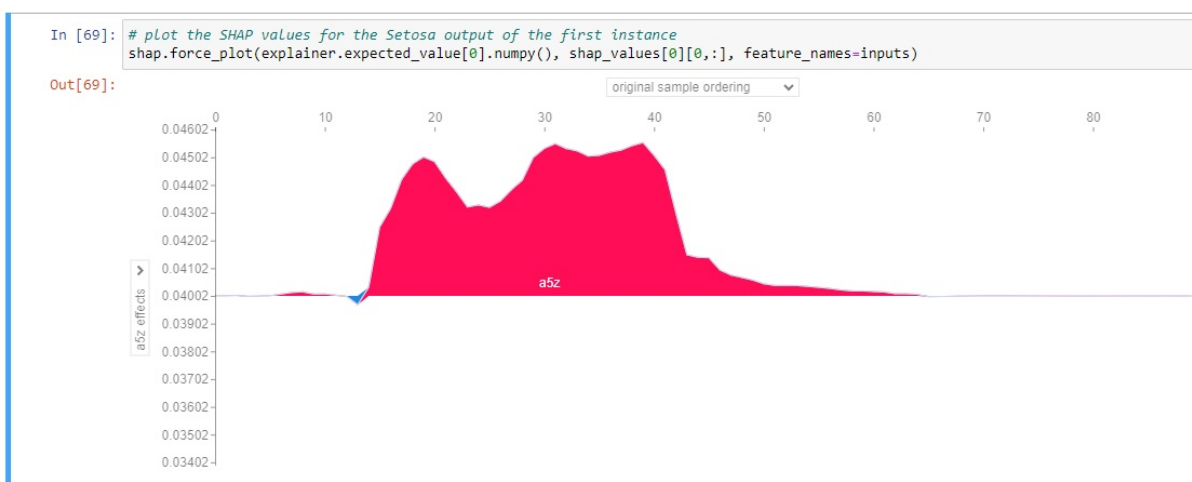
Figura 34 – Captura de tela mostrando a manipulação dos dados obtidos com a biblioteca LIME, agregando os diferentes valores relacionados a mesma articulação.

```
In [34]: plot_lime_org(organiza_lime_art(exp, prediction[0]))
```



Fonte — Produzida pelo autor.

Figura 35 – Captura de tela mostrando os resultados dos testes executados com a biblioteca SHAP, com o impacto de cada articulação em cada eixo no tempo para a classificação correta do movimento de forma global.



Fonte — Produzida pelo autor.

Para a estratégia desenvolvida, o público alvo selecionado foi o de profissionais da área de reabilitação, com o contato inicial mediante um pesquisador da área de reabilitação que já possui experiência com aplicações computacionais. Outras possibilidades de público alvo possíveis serão citadas nos trabalhos futuros. O objetivo da escolha é que o profissional compare a explicabilidade obtida com o conhecimento de especialista na área que ele possui, indicando se a explicabilidade é relevante.

6.6 Considerações finais do capítulo

Este capítulo buscou compreender o cenário para o estudo de caso, as soluções de IA que já estão sendo utilizadas e como a XAI pode auxiliar nesse contexto, apresentando uma estratégia de uso de XAI, com exposição dos primeiros resultados obtidos com experimentos.

7 Desenvolvimento

7.1 Considerações iniciais do capítulo

No presente capítulo é apresentado o processo de desenvolvimento da estratégia de uso e da ferramenta como estudo de caso, serão apresentadas as etapas de desenvolvimento, ainda que o processo tenha sido realizado de maneira constante, sem divisão clara de etapas. As etapas são: as tentativas com soluções inadequadas, o encontro da solução adequada, o desenvolvimento da interface, a integração com um novo conjunto de dados, a avaliação do protótipo por um colaborador pesquisador da área de reabilitação com ajustes para atendimento as suas demandas.

7.2 Soluções inadequadas

Com a solução de IA para classificação de movimentos em funcionamento, através das bibliotecas de XAI foram realizados testes e a partir desse estágio buscou-se compreender como os diferentes componentes dos dados de treinamento, as posições das articulações, impactam nos resultados de classificação.

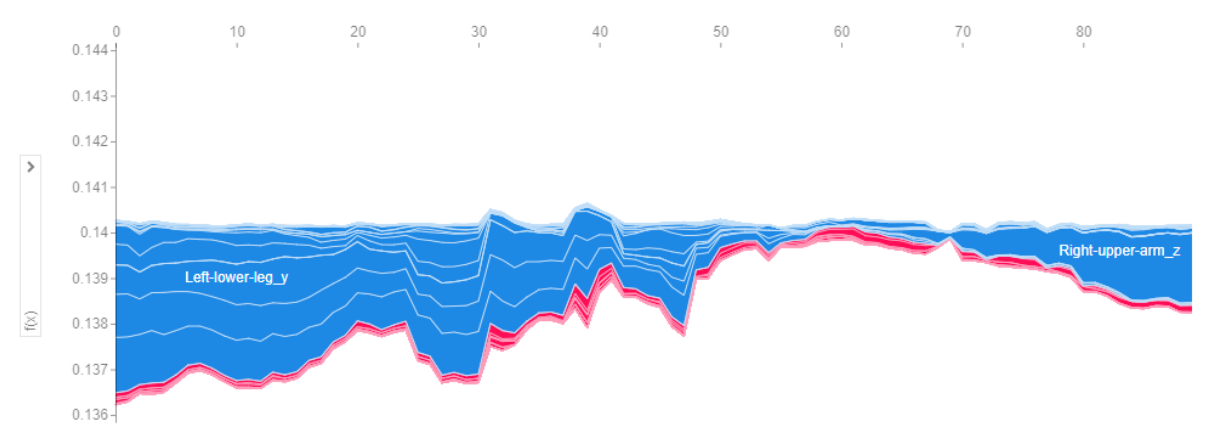
Foram realizados testes com as bibliotecas LIME e SHAP, porém os resultados não foram satisfatórios, para a biblioteca LIME os valores de relevância foram produzidos em grande quantidade e com indicação dos *frames*, não realizando uma análise adequada de dados temporais, não possibilitando uma visão relevante da explicabilidade como pode ser observado na Figura 36. Já os resultados para a biblioteca SHAP geraram gráficos complexos com dados sobrepostos, como visto na Figura 37.

Figura 36 – Captura de tela mostrando que ao aplicar a biblioteca LIME no modelo de IA, os dados eram retornados com indicação de *frame* e não forneciam uma explicação consistente.



Fonte — Produzida pelo autor.

Figura 37 – Captura de tela mostrando que ao aplicar a biblioteca SHAP no modelo de IA, cada *frame* era tratado como um elemento a ser explicado, gerando uma visualização sobreposta e confusa.



Fonte — Produzida pelo autor.

Os resultados não foram adequados devido a dimensionalidade dos dados, uma vez que eles dentro de cada amostra variavam no tempo, onde cada *frame* registrado gerava um novo registro, porém que pertencia a mesma amostra. Ao detectar esse problema, foram realizadas manipulações dos dados fornecidos pela biblioteca LIME e foram exploradas as diversas possibilidades de geração de gráficos com a biblioteca SHAP, buscando obter uma visualização consistente do impacto das articulações como desejado, porém a falta de mecanismos nas bibliotecas para tratar a temporalidade na estrutura de cada amostra impossibilitou a evolução destas abordagens.

Assim sendo, foram buscadas soluções de XAI que pudessem tratar especificamente de dados deste tipo, as buscas retornaram algumas opções, entre ela se destacou a

TimeSHAP, uma implementação da biblioteca SHAP adaptada para tratamento de dados temporais, o destaque se deu pela avaliação e manutenção do repositório de código-fonte da solução na plataforma GitHub ([BENTO et al., 2021](#)) ([NAYEBI et al., 2023](#)).

7.3 TimeSHAP

A biblioteca TimeSHAP tem uma abordagem baseada em eventos para tratar dados multidimensionais com séries temporais, cada amostra pode ser avaliada de 3 formas: por recurso, considerando todos os *frames* que compõem a amostra, por evento, considerando todos os recursos em momento escolhido, e por célula, considerando um certo recurso em um certo momento. Essa abordagem é ilustrada na Figura 38.

Figura 38 – Diagrama mostrando a análise de uma amostra considerando recurso, evento e célula no qual se baseia a biblioteca TimeSHAP.

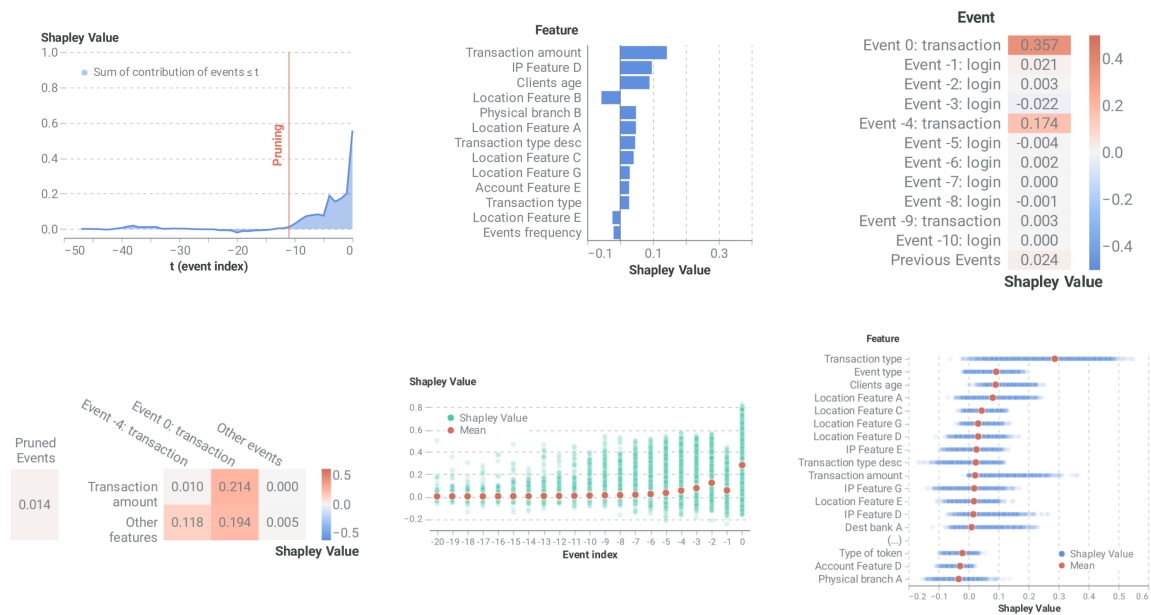
Amostra 1					
Recurso	Célula	Recurso 1 Momento 0	Recurso 1 Momento 1	Recurso 1 Momento 2	Recurso 1 Momento 3
		Recurso 2 Momento 0	Recurso 2 Momento 1	Recurso 2 Momento 2	Recurso 2 Momento 3
		Recurso 3 Momento 0	Recurso 3 Momento 1	Recurso 3 Momento 2	Recurso 3 Momento 3
		Recurso 4 Momento 0	Recurso 4 Momento 1	Recurso 4 Momento 2	Recurso 4 Momento 3
		Recurso 5 Momento 0	Recurso 5 Momento 1	Recurso 5 Momento 2	Recurso 5 Momento 3
	Evento				

Fonte — Produzida pelo autor.

No repositório da biblioteca TimeSHAP estavam disponíveis exemplos de uso, um exemplo de uso foi executado através da plataforma Jupyter Notebook para servir como prova de conceito. Após o funcionamento do exemplo, o mesmo foi adaptado para integração ao cenário de classificação de movimentos. Durante esse processo foram encontrados

problemas de compatibilidade da versão disponível no momento da biblioteca TimeSHAP com a biblioteca SHAP, os problemas foram resolvidos atualizando as referências a biblioteca SHAP nos arquivos da biblioteca TimeSHAP. Na Figura 39 podem ser vistos gráficos do TimeSHAP obtidos executando um exemplo de utilização da explicabilidade em dados bancários disponível no repositório.

Figura 39 – Gráficos diversos fornecidos pelo TimeSHAP obtidos executando um exemplo disponível no repositório.

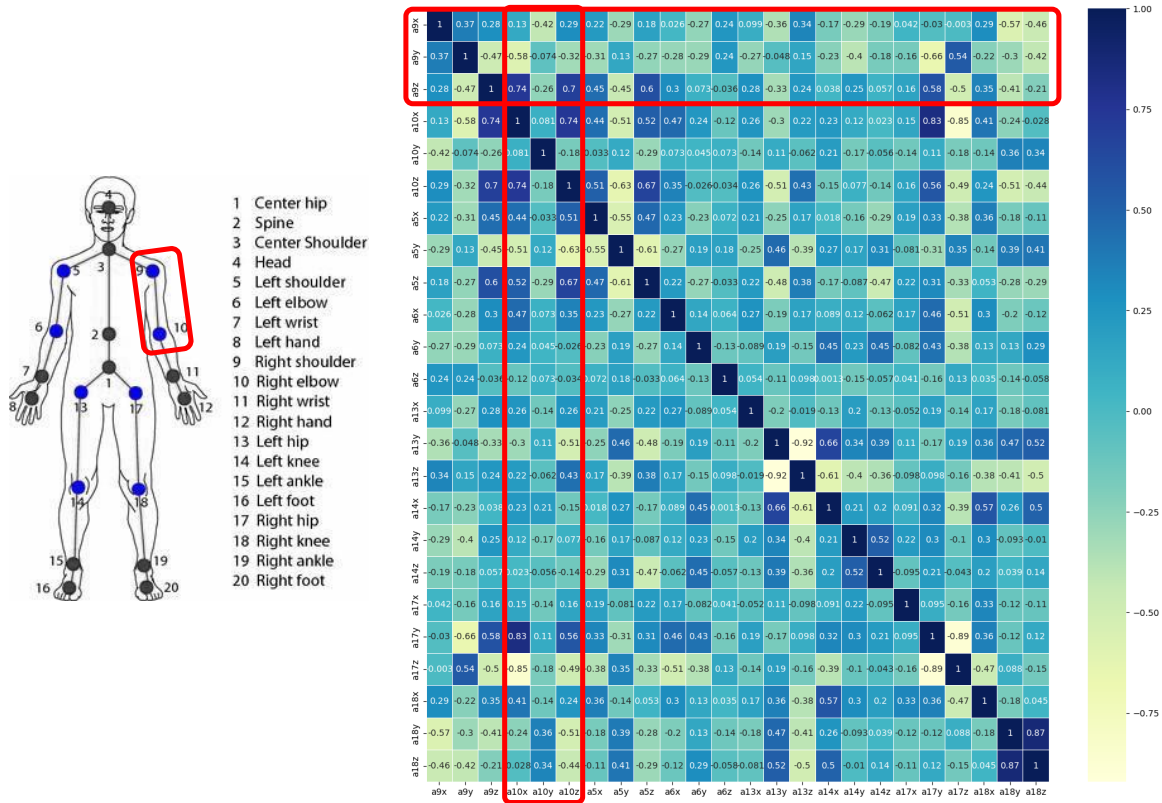


Fonte — Produzida pelo autor.

Entretanto, novos problemas foram encontrados no processo de integração das soluções. Acabou-se detectando que os novos problemas eram causados pela compatibilidade da biblioteca TimeSHAP com modelos de classificação binária (2 classes de saída) e a tentativa de integração com um modelo de classificação multiclasse (diversas classes de saída). A solução encontrada foi a geração de modelos binários individuais ajustados para cada classe para utilização em conjunto com a biblioteca TimeSHAP, além do modelo inicial multiclasse.

Analisando a evolução do desenvolvimento e os primeiros resultados obtidos, foi identificada a possibilidade de se obter informações relevantes entre as variáveis de entrada aplicando técnicas estatísticas de correlação, além de ajustar as amostras de 90 para 30 *frames*. Na Figura 40 pode ser observado um estudo realizado sobre a correlação dos valores das articulações.

Figura 40 – Imagem mostrando um estudo sobre a correlação dos valores das articulações, os eixos mostram as articulações, cada célula indica a correlação de dois valores de articulações, quanto mais escura a cor, mais correlacionados estão os valores.

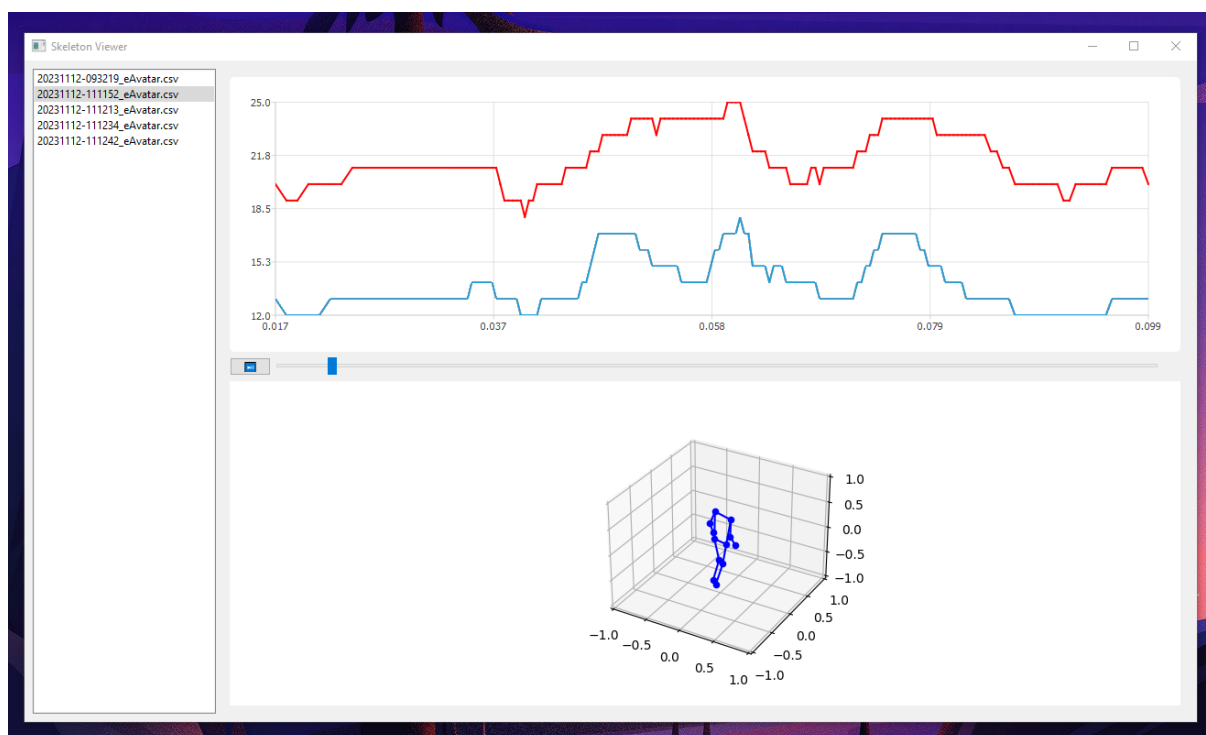


Fonte — Produzida pelo autor.

7.4 Interface

A partir dos resultados obtidos com a biblioteca TimeSHAP, os estudos de correlação, o desenvolvimento foi direcionado para a produção de um *dashboard* integrando as informações obtidas. Decidiu-se que o *dashboard* deveria contar com um *player*, capaz de reproduzir os movimentos através de um avatar e integrar os dados de explicabilidade. Optou-se pela padronização dos dados no formato CSV de forma a utilizar um protótipo de *player* já existente para esse formato, que pode ser observado na Figura 41.

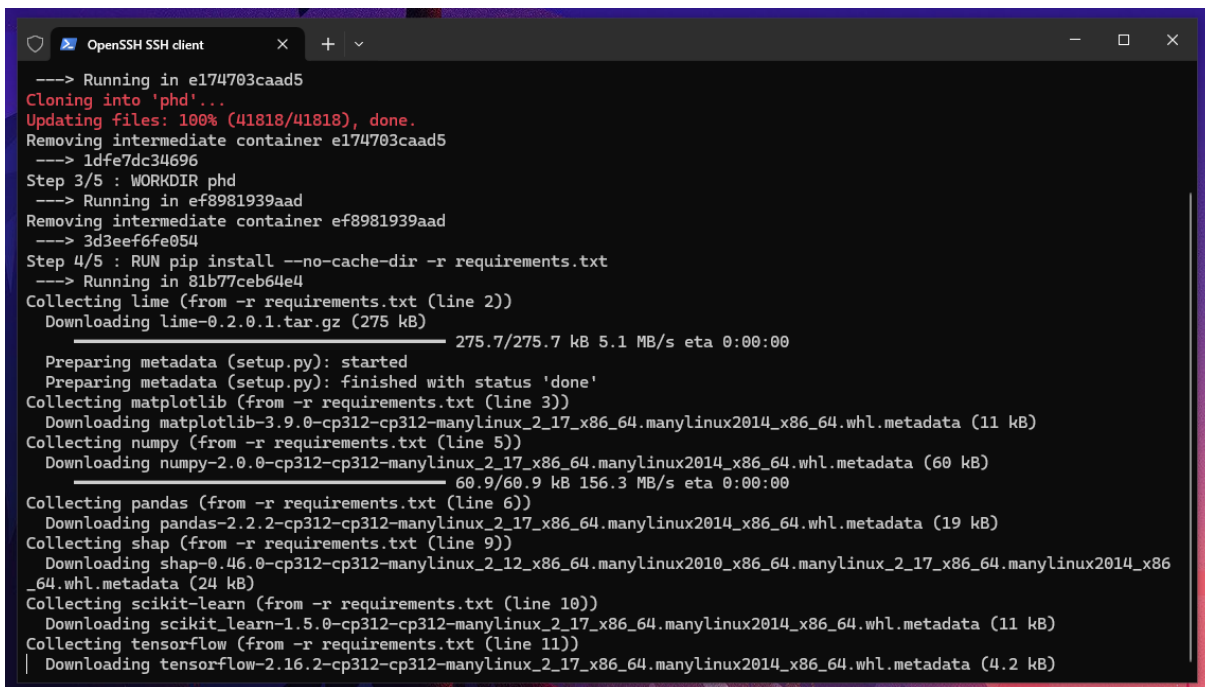
Figura 41 – Captura de tela mostrando o protótipo de *player* compatível com dados no formato CSV que já existia.



Fonte — Produzida pelo autor.

Foi utilizada a biblioteca Voilà que se integra com a plataforma Jupyter Notebook gerando um *dashboard* que pode ser acessado com uma aplicação web independente para disponibilização de acesso à ferramenta. A hospedagem foi realizada em um servidor da universidade através da tecnologia de contêineres Docker, permitindo atualizações instantâneas da aplicação em execução através do repositório de código-fonte na plataforma GitHub. Na Figura 42 é demonstrado parte do processo de criação do contêiner com o código da aplicação atualizado através de *download* do repositório na plataforma GitHub e a instalação de todas as bibliotecas necessárias de forma automatizada.

Figura 42 – Imagem mostrando a captura de tela durante execução de *script* para atualização da aplicação no contêiner em execução no servidor.



```

----> Running in e174703caad5
Cloning into 'phd'...
Updating files: 100% (41818/41818), done.
Removing intermediate container e174703caad5
----> 1dfe7dc34696
Step 3/5 : WORKDIR phd
----> Running in ef8981939aad
Removing intermediate container ef8981939aad
----> 3d3eef6fe054
Step 4/5 : RUN pip install --no-cache-dir -r requirements.txt
----> Running in 81b77ceb64e4
Collecting lime (from -r requirements.txt (line 2))
  Downloading lime-0.2.0.1.tar.gz (275 kB)
    _____ 275.7/275.7 kB 5.1 MB/s eta 0:00:00
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Collecting matplotlib (from -r requirements.txt (line 3))
  Downloading matplotlib-3.9.0-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (11 kB)
Collecting numpy (from -r requirements.txt (line 5))
  Downloading numpy-2.0.0-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (60 kB)
    _____ 60.9/60.9 kB 156.3 MB/s eta 0:00:00
Collecting pandas (from -r requirements.txt (line 6))
  Downloading pandas-2.2.2-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (19 kB)
Collecting shap (from -r requirements.txt (line 9))
  Downloading shap-0.46.0-cp312-cp312-manylinux_2_12_x86_64.manylinux2010_x86_64.manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (24 kB)
Collecting scikit-learn (from -r requirements.txt (line 10))
  Downloading scikit_learn-1.5.0-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (11 kB)
Collecting tensorflow (from -r requirements.txt (line 11))
  Downloading tensorflow-2.16.2-cp312-cp312-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.2 kB)

```

Fonte — Produzida pelo autor.

Após um período de avaliação do protótipo foi verificada uma constância nos resultados, mostrando um retorno positivo. Foram revistas e ajustadas as cores da interface, indicando de forma mais precisa os impactos positivos e negativos na classificação, conforme demonstrado na Figura 43. Foi detectado um problema de desempenho, decorrente da necessidade de se realizar os cálculos para análise de explicabilidade cada vez que uma nova amostra era selecionada. Para mitigar esse problema, foram realizados os cálculos para todas as amostras disponíveis no conjunto de dados, somente carregando as informações e produzindo a interface conforme o usuário interage com a ferramenta, sem a necessidade de cálculos demorados. Certas amostras apresentaram incompatibilidade com o processo de explicabilidade, não foi possível identificar a causa do problema, sendo assim foi dada continuidade no projeto, visto que o percentual de amostras com problemas era pequeno.

Figura 43 – Escala de cores escolhida para exibição dos dados de explicabilidade no *player*, variando as cores nas articulações, onde vermelho indica contribuição positiva e azul negativa.

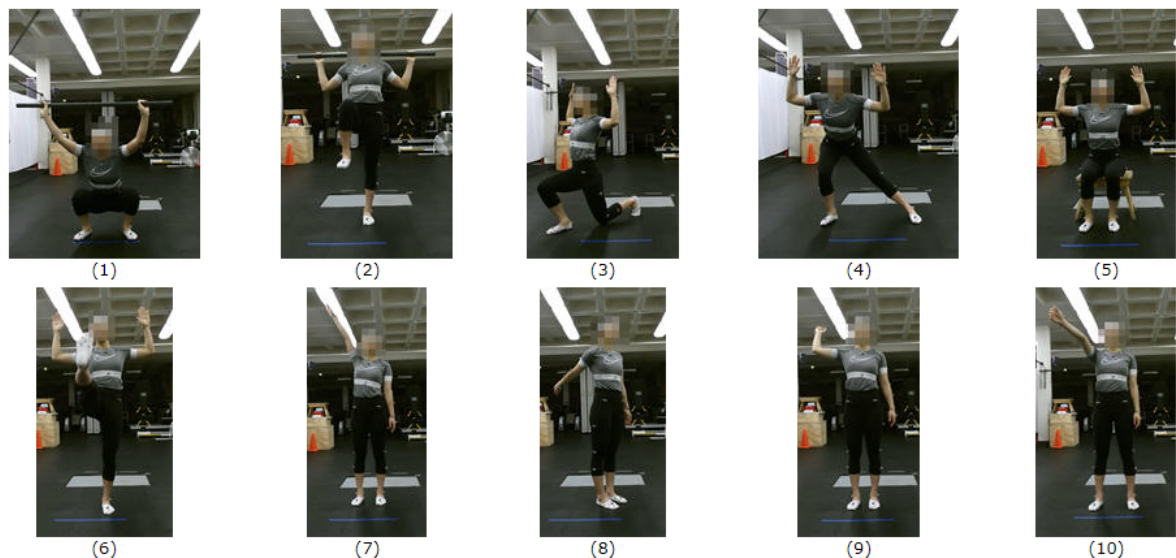


Fonte — Produzida pelo autor.

7.5 Conjunto de dados UI-PRMD

Durante o processo de desenvolvimento do protótipo foi identificado um conjunto de dados aberto disponível na Internet composto por gravações de movimentos através dos dispositivos Kinect e Vicon. Chamado UI-PRMD (University of Idaho – Physical Rehabilitation Movement Dataset), o conjunto é formado por 1.000 amostras, 10 exercícios comuns realizados por pacientes em programas de fisioterapia e reabilitação, repetidos 10 vezes por 10 indivíduos saudáveis, em frente a dois sistemas sensoriais para captura de movimento: um rastreador óptico Vicon e uma câmera Kinect ([VAKANSKI et al., 2018](#)). Na Figura 44 pode ser observado o processo de criação desse conjunto de dados.

Figura 44 – Imagem mostrando a execução dos exercícios para produção do conjunto de dados UI-PRMD.



Fonte — Obtido na página do projeto.

Após estudos com o conjunto de dados UI-PRMD, buscando a oportunidade de validar a estratégia proposta com outro conjunto de dados, de modo a fortalecer as evidências de funcionamento da estratégia, foi então elaborada uma nova versão da ferramenta para análise dos dados do conjunto UI-PRMD seguindo a mesma abordagem, para que isso fosse possível, os dados do novo conjunto foram convertidos ao formato CSV compatível e pequenos ajustes foram realizados na ferramenta, principalmente no *player*, porém buscando alterar minimamente a interface, de modo a não interferir na experiência do usuário.

7.6 Pesquisador da área de reabilitação

A versão inicial, chamada de "rebase" a partir desse ponto, em referência ao banco de dados utilizado para gravação do primeiro conjunto de dados (TROTTA et al., 2021), e versão adaptada para o conjunto UI-PRMD foram unificadas em uma única interface. Ao constatar certa maturidade no protótipo, o mesmo foi encaminhado para testes por um colaborador pesquisador da área de reabilitação, inicialmente para explorar a interface, foi posteriormente realizada uma apresentação da proposta e foram sanadas as dúvidas, o pesquisador demonstrou interesse na ferramenta e se comprometeu a testá-la. Após mais um período de testes, o pesquisador indicou interesse na implementação de duas novas funcionalidades: a aplicação em um conjunto de dados coletados por ele e outros colegas da área através de aplicações de reabilitação que produziam dados no formato CSV e a possibilidade de realizar *upload* de uma amostra CSV de forma individual para análise.

Novos testes foram realizados com o conjunto de dados fornecido pelo colaborador pesquisador da área de reabilitação, porém, esse conjunto precisou ser organizado, as amostras adequadas foram selecionadas e rotuladas, resultando em um conjunto de 225 amostras chamado de "rgbtracker", em referência ao método de captura integrado as aplicações que geraram os dados. As classes do conjunto de dados são duas:

- Puzzle sentado - 80 amostras: execução de movimentos através da aplicação de um jogo de quebra-cabeça (puzzle) com o paciente sentado.
- Puzzle em pé - 145 amostras: execução de movimentos através da aplicação de um jogo de quebra-cabeça (puzzle) com o paciente em pé.

Após diversos testes, optou-se por realizar um novo treinamento para obtenção de um novo modelo, com ambos os conjunto de dados "rebase" e "rgbtracker", obtendo assim maior consistência nos retornos do modelo e na análise de XAI. Ajustando as amostras para 90 *frames* devido a maior duração das amostras do conjunto de dados "rgbtracker" e utilizando o novo modelo treinado, foi gerada uma nova versão da aplicação, explorando os dados fornecidos pelo colaborador pesquisador da área de reabilitação.

Por fim foi adaptada a interface para possibilidade de *upload* de uma nova amostra, o campo de *upload* permite o envio de arquivos no formato CSV e a ferramenta se ajusta para utilizar o modelo compatível com os conjuntos de dados "rebase" e "rgbtracker", caso a amostra contenha dados de movimentos compatíveis com esses conjuntos de dados, a ferramenta consegue retornar uma análise de XAI consistente. O processo de análise de amostra enviada através de *upload* tende a ser mais demorado, visto que não é possível realizar nenhum tipo de cálculo de forma antecipada, executando toda análise em tempo real.

Foram integradas todas as versões produzidas, permitindo análises nos conjuntos de dados "rebase", "UI-PRMD" e "rgbtracker", além o *upload* de uma nova amostra. Na Figura 45 pode ser observada a interface final da ferramenta que foi denominada "MoveXAI", destacando o campo de seleção do conjunto de dados.

Figura 45 – Captura de tela mostrando a interface inicial do MoveXAI com o campo de seleção de conjunto de dados.

MoveXAI

Dataset: rebase ▼

- rebase:** UI-PRMD
- UI-PRMD:** rgbtracker
- rgbtracker:** upload
- upload:** opção de enviar para análise um arquivo CSV padronizado com 10 pontos. (90 frames)

Sessão (0-2398): Analisar Sessão

Player	Linha do Tempo	Frames	Articulações	Células	Correlação
--------	----------------	--------	--------------	---------	------------

Fonte — Produzida pelo autor.

7.7 Considerações finais do capítulo

Foram demonstradas as etapas percorridas até o ponto atual de desenvolvimento da estratégia, citando as interações do desenvolvedor com os interessados, buscando o melhor resultado ao desenvolver uma ferramenta que permita utilizar a abordagem de XAI no auxílio à reabilitação neuromotora. No capítulo a seguir serão demonstrados os resultados obtidos com o desenvolvimento da estratégia e da ferramenta e todo o processo de pesquisa.

8 Resultados

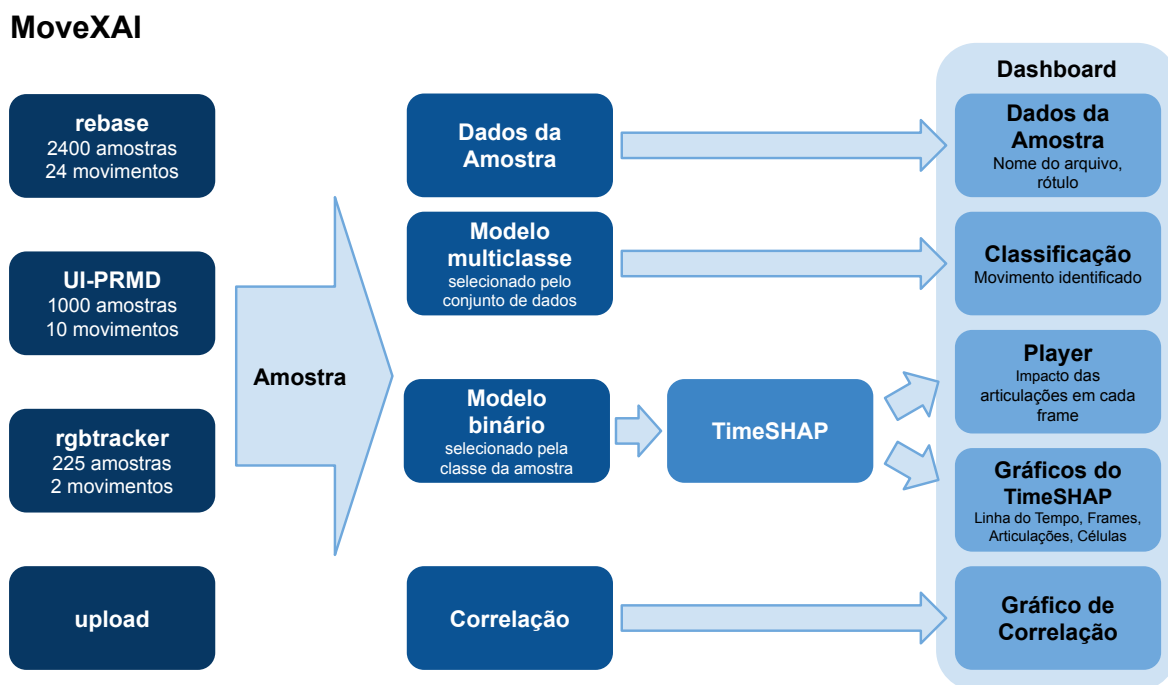
8.1 Considerações iniciais do capítulo

Neste capítulo são apresentados os resultados do estudo desenvolvido, incluindo as funcionalidades da ferramenta e os avanços na área de XAI.

A estratégia de uso de XAI desenvolvida unifica a utilização de IA, aspectos de XAI e Visualização da Informação, fornecendo ao profissional de saúde que trabalha com reabilitação neuromotora uma visão dos dados de movimento ampliada, com estimativas dos momentos e das articulações mais relevantes de cada amostra.

Essa nova metáfora visual foi apresentada em um *dashboard* juntamente com os gráficos padrões fornecidos pela biblioteca TimeSHAP e com uma visualização das correlações existentes na amostra. Na Figura 46, pode ser observado o fluxo de dados entre os componentes que compõe a ferramenta.

Figura 46 – Diagrama mostrando o fluxo de dados da ferramenta MoveXAI.



Fonte — Produzida pelo autor.

Pode-se observar que o resultado obtido com o desenvolvimento da estratégia e da ferramenta difere do proposto inicialmente e apresentado na Figura 31, isso se deve a necessidade de adequação aos dados temporais, das oportunidades de aplicação

da estratégia em novas bases de dados e implementação de funções demandadas pelos interessados.

8.2 A ferramenta

Nesta seção são demonstrados os componentes da ferramenta, detalhando suas funcionalidades.

A parte superior da interface é composta pelo seletor de conjunto de dados, chamado de *Dataset* na interface, com uma breve descrição de cada conjunto, o seletor de amostras, chamadas de sessões na interface, e o espaço para exibição dos dados da amostra selecionada. A opção de *upload* pode ser acessada no seletor de conjunto de dados, ao selecionar essa opção, a parte superior da interface é alterada, se adaptando a função de *upload*. Os modos da parte superior da interface podem ser visualizados nas Figuras 47 e 48.

Figura 47 – Captura de tela mostrando a parte superior da interface após seleção de uma amostra.

MoveXAI

Dataset: ▼

rebase: dataset com os dados gravados na base de dados Rebase, porém em formato CSV, com 100 amostras para da um dos 24 tipos de movimento presentes no conjunto. (30 frames)

UI-PRMD: dataset com dados do dataset público UI-PRMD, porém em formato CSV, com 100 amostras para da um dos 10 tipos de movimento presentes no conjunto. (30 frames)

rgbtracker: dataset com os dados gravados através da aplicação Puzzle sendo executada com o usuário sentado e em pé. (90 frames)

upload: opção de enviar para análise um arquivo CSV padronizado com 10 pontos. (90 frames)

Sessão (0-2398):

Nome do Arquivo: datasets/rebase/Left Circle Clockwise/20231112-093445_eAvatar.csv

Número da Sessão: 25 - **Rótulo da Sessão:** Left Circle Clockwise

Previsão do Modelo: Left Circle Clockwise - **Precisão:** 100.0% - **Previsão Correta!!!**

Fonte — Produzida pelo autor.

Figura 48 – Captura de tela mostrando a parte superior da interface após seleção do modo de *upload*.

MoveXAI

Dataset:

rebase: dataset com os dados gravados na base de dados Rebase, porém em formato CSV, com 100 amostras para da um dos 24 tipos de movimento presentes no conjunto. (30 frames)

UI-PRMD: dataset com dados do dataset público UI-PRMD, porém em formato CSV, com 100 amostras para da um dos 10 tipos de movimento presentes no conjunto. (30 frames)

rgbtracker: dataset com os dados gravados através da aplicação Puzzle sendo executada com o usuário sentado e em pé. (90 frames)

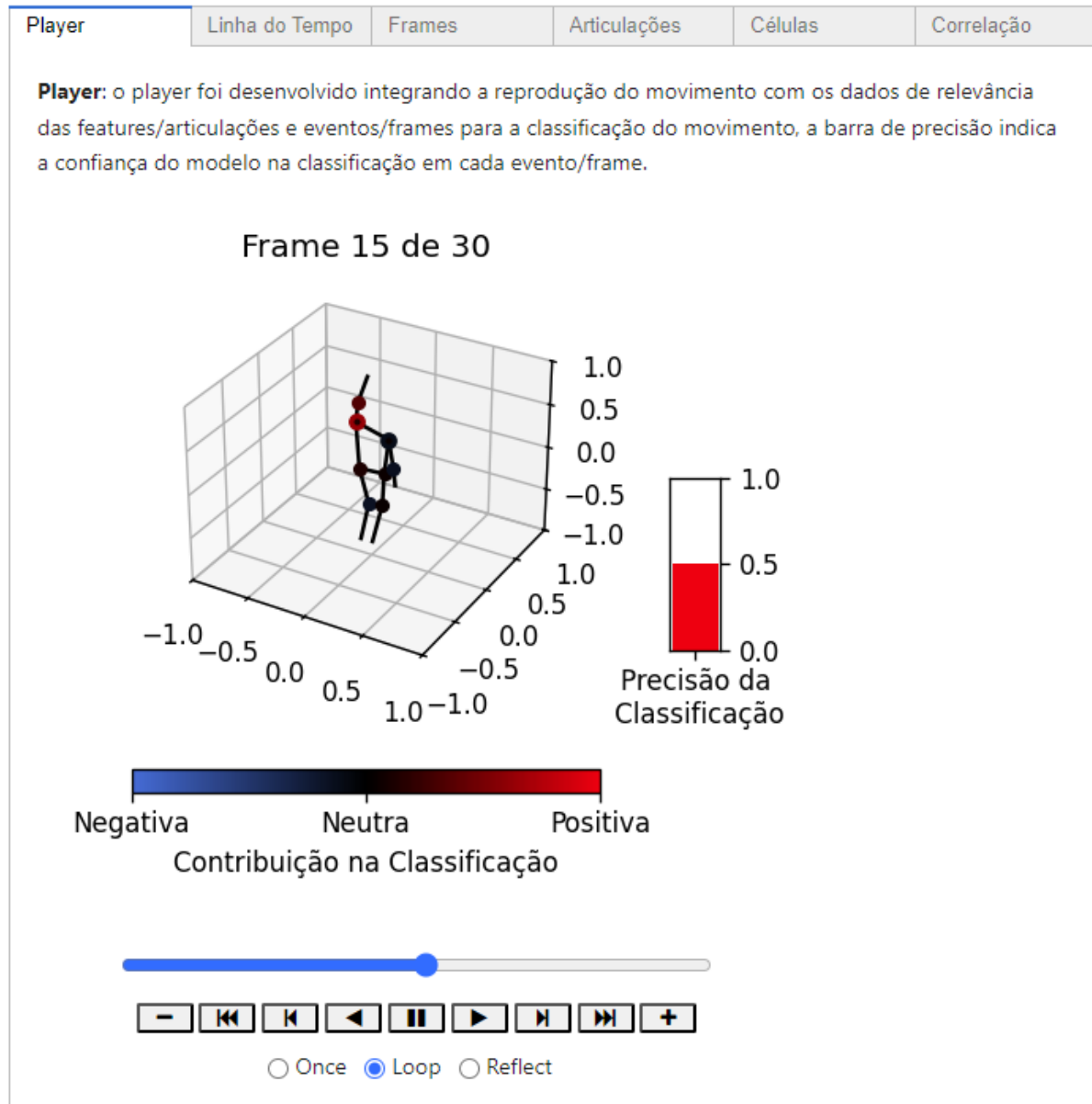
upload: opção de enviar para análise um arquivo CSV padronizado com 10 pontos. (90 frames)

 Upload (1)

Fonte — Produzida pelo autor.

A Figura 49 mostra o *player*, elemento que associa um avatar executando os movimentos registrados nos dados da amostra, com suas articulações variando de cor conforme o impacto que cada articulação gera em cada momento na classificação do movimento. Associado ao *player* se tem a barra de precisão, que indica o crescimento da precisão do modelo em sua predição no decorrer da amostra.

Figura 49 – Captura de tela mostrando a interface do MoveXAI na aba "Player", com a articulação do ombro destacada em vermelho, indicando o impacto positivo dela na classificação e a barra indicando a precisão da classificação no *frame* indicado.



Fonte — Produzida pelo autor.

Dentre os gráficos padrões da biblioteca TimeSHAP se tem o gráfico de *pruning*, utilizado juntamente com o recurso de poda de dados da biblioteca, ele permite otimizar a análise XAI eliminando dados de eventos pouco relevantes para o modelo. Esse recurso não foi utilizado no atual estudo, porém visualizar o gráfico de *pruning* permite um ganho de informação relativa à confiança do modelo em sua previsão. A contagem de eventos é invertida, sendo o último indicado como 0, conforme padrão de biblioteca de XAI utilizada. Esses dados foram utilizados na produção da barra de precisão do *player*. O gráfico de

pruning foi utilizado na aba "Linha do Tempo" conforme pode ser observado na Figura 50

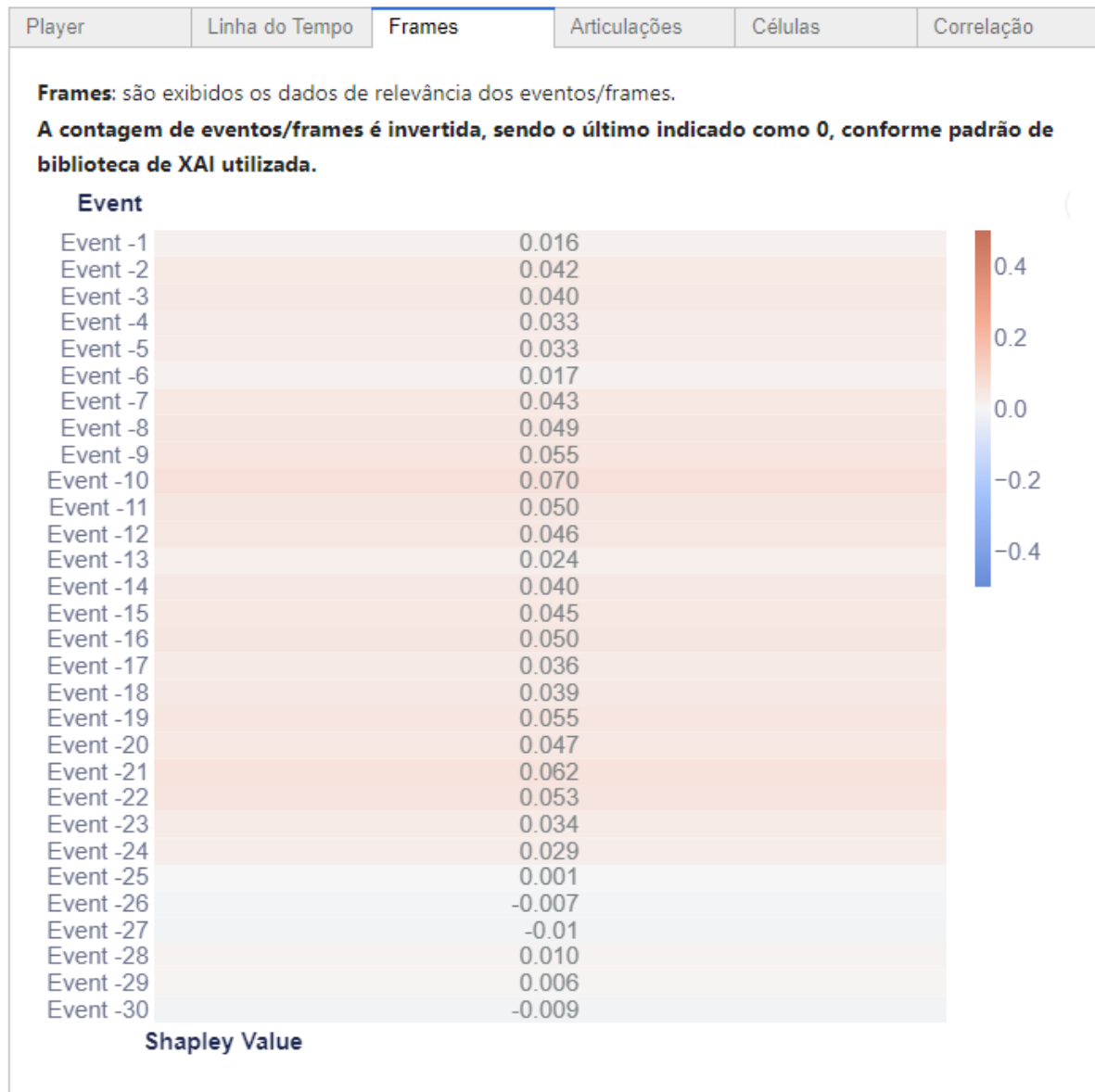
Figura 50 – Captura de tela mostrando a interface do MoveXAI na aba "Linha do Tempo", onde o gráfico de *pruning* da Biblioteca XAI é utilizado para demonstrar o crescimento da precisão do modelo no decorrer dos *frames*.



Fonte — Produzida pelo autor.

A aba "Frames" da ferramenta utiliza o gráfico de eventos da biblioteca TimeSHAP, onde os *frames* são considerados os eventos sendo exibidos no gráfico, com a coloração variando conforme o impacto do evento na classificação, conforme pode ser visto na Figura 51. A ordem dos eventos é invertida, sendo o último indicado como 0, como também acontece no gráfico de *pruning*.

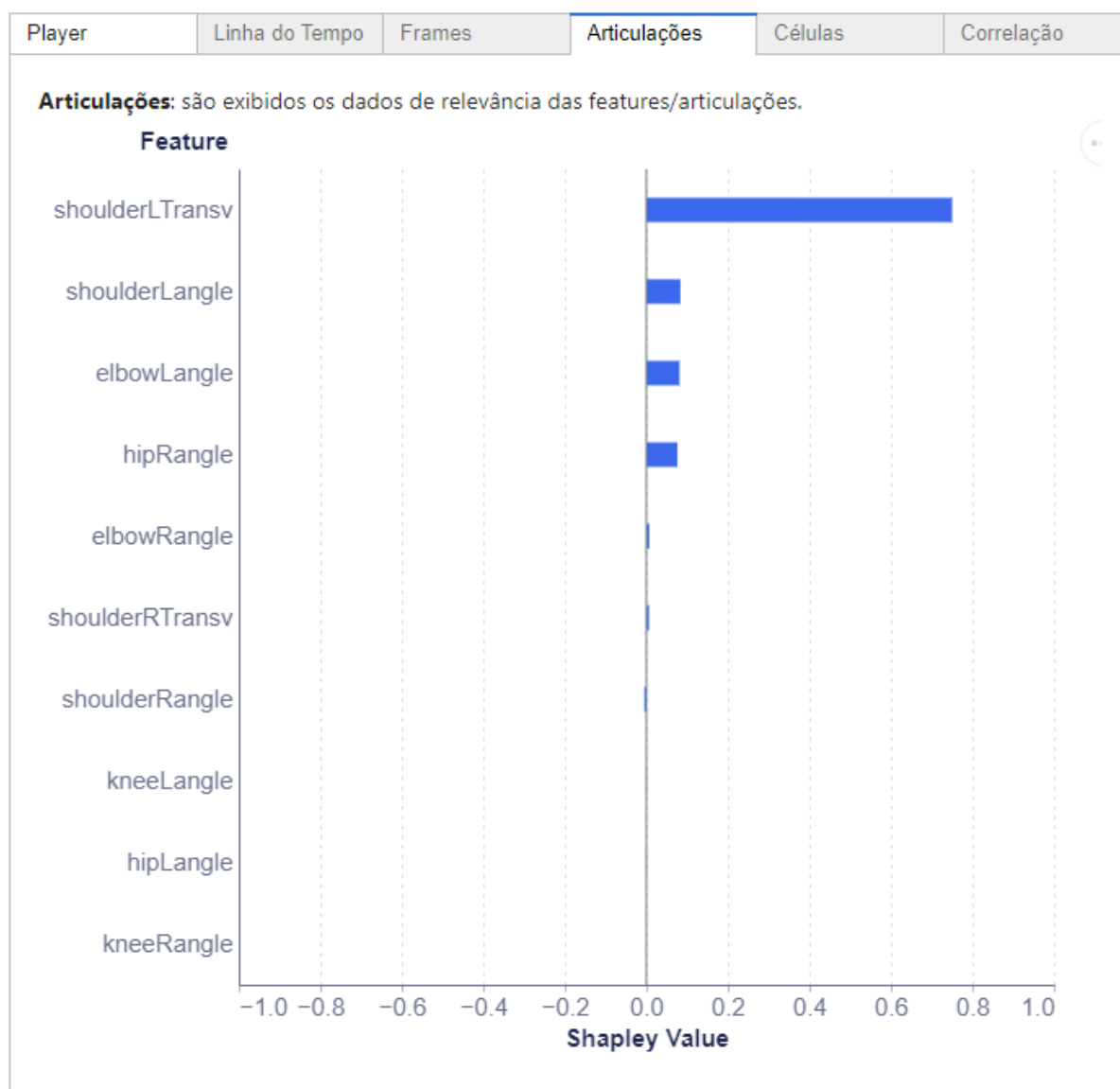
Figura 51 – Captura de tela mostrando a interface do MoveXAI na aba "Frames", onde o gráfico de eventos da Biblioteca XAI é utilizado para demonstrar o impacto de cada *frame* na classificação. Os *frames* são indicado na lateral esquerda, com os valores de Shapley calculados, onde quanto mais intensa a cor, maior o impacto do *frame*.



Fonte — Produzida pelo autor.

A aba "Articulações" da ferramenta utiliza o gráfico de características da biblioteca TimeSHAP, onde as articulações são consideradas as características e são exibidos no gráfico, com as barras indicando o impacto na classificação, conforme pode ser visto na Figura 52.

Figura 52 – Captura de tela mostrando a interface do MoveXAI na aba "Articulações", onde o gráfico de características da Biblioteca XAI é utilizado para demonstrar o impacto de cada articulação na classificação. As articulações são mostradas na lateral esquerda e os valores de impacto são exibidos através de um gráfico de barras.



Fonte — Produzida pelo autor.

O gráfico de células da biblioteca TimeSHAP é utilizado na aba "Células", exibindo no gráfico uma intersecção das 10 articulações mais relevantes com os 10 eventos mais relevantes na classificação, além disso, existe uma coluna demonstrando o somatório dos demais eventos para cada articulação exibida, conforme visto na Figura 53.

Figura 53 – Captura de tela mostrando a interface do MoveXAI na aba "Células", onde o gráfico de células da Biblioteca XAI é utilizado para demonstrar o impacto das principais articulações e *frames* na classificação. É apresentada uma matriz com o eixo horizontal mostrando as articulações e o eixo vertical os *frames*, com a cor variando de acordo com o impacto da célula (articulação/*frame*) na classificação.

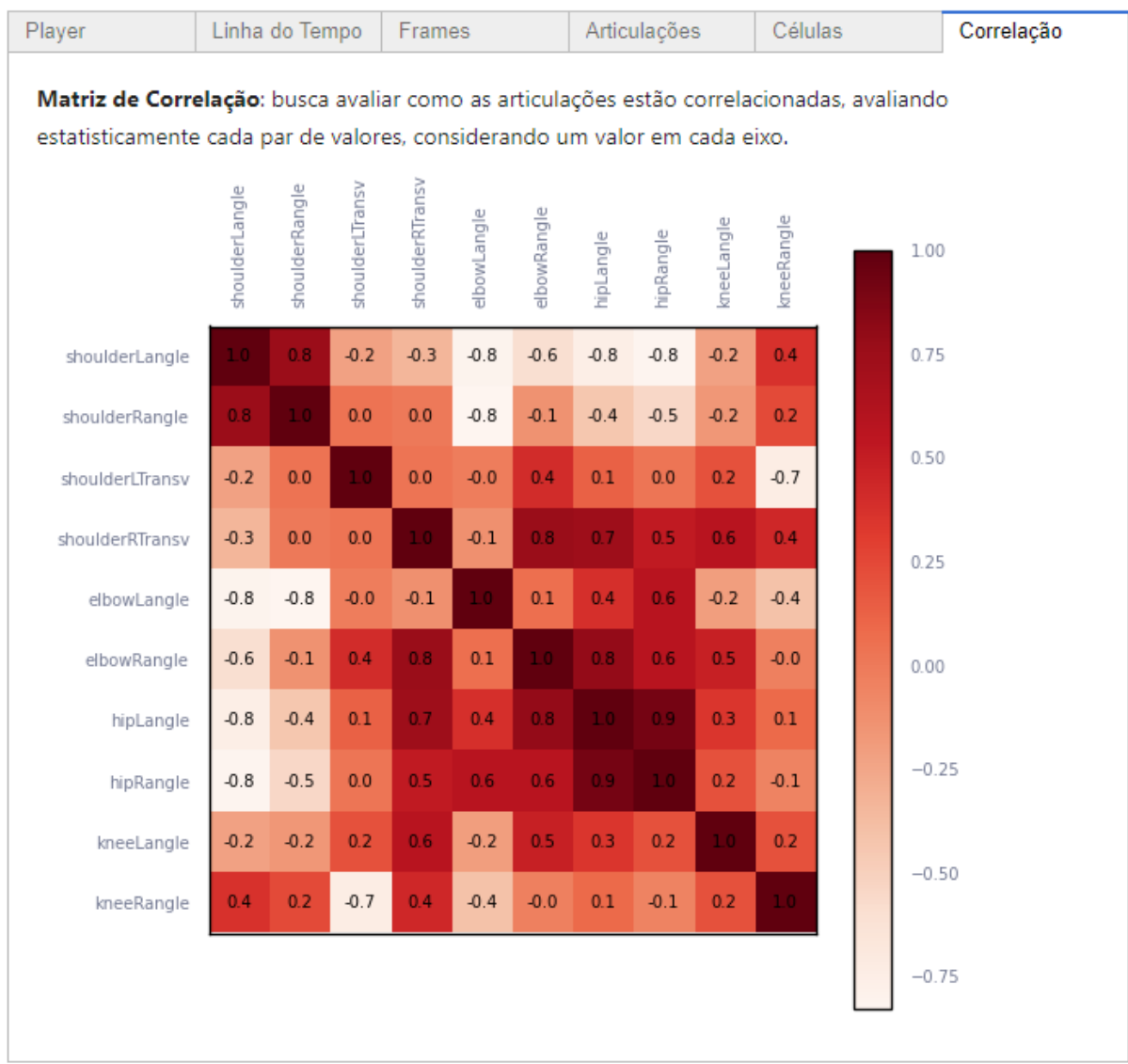


Fonte — Produzida pelo autor.

Os dados de uma versão ampliada dessa análise, sem filtrar os eventos e articulações mais relevantes, são usados para agregar explicabilidade no *player*.

Na aba "Correlação" foi utilizada uma análise estatística de correlação entre os dados da amostra, analisando a correlação entre as articulações. Através dessa análise pode-se observar como articulações próximas, como ombro e cotovelo, acabam por serem movimentadas em conjunto dependendo do exercício executado, inclusive indicando movimentações conjuntas não desejadas do ponto de vista terapêutico. O gráfico de correlação pode ser visto na Figura 54

Figura 54 – Captura de tela mostrando a interface do MoveXAI na aba "Correlação", onde um gráfico demonstra a correlação da articulações. Os eixos mostram as articulações, cada célula indica a correlação de dois valores de articulações, quanto mais escura a cor, mais correlacionados estão os valores.



Fonte — Produzida pelo autor.

8.3 Infraestrutura computacional

Atualmente a ferramenta se encontra hospedada em um servidor da universidade localizado na Faculdade de Medicina de Botucatu, por conta do vínculo do pesquisador principal como servidor técnico-administrativo dessa unidade da Unesp. O servidor em questão é da marca Dell, modelo PowerEdge R740, com processador Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz e 32GB de memória RAM. O servidor executa o sistema de virtualização VMware ESXi 6 e através dele uma máquina virtual para execução de contêineres através da ferramenta Docker. Atualmente, após ajustes e otimizações necessárias, a ferramenta é executada em um contêiner que consome uma quantidade de memória de cerca de 1 GB e limitando a velocidade de execução dentro do potencial disponível do processador. A função de *upload* tende a aumentar o consumo de recursos. As demandas de armazenamento da ferramenta dependem dos conjuntos de dados utilizados, atualmente não representam um fator de preocupação.

Esse cenário foi organizado para a fase inicial de testes, para ampliação de uso da ferramenta será necessário a hospedagem em um servidor com mais poder de processamento, inclusive com potencial adequado para inteligência artificial. A ampliação de uso deve ser realizada em conjunto com o monitoramento de recursos do servidor de hospedagem, para escalonamento adequado e dessa forma evitar problemas de indisponibilidade.

8.4 A área de explicabilidade

Ainda que a ferramenta desenvolvida tenha papel importante, os resultados deste estudo não se resumem ao desenvolvimento da ferramenta, mas estão presentes em todo processo, desde a revisão de literatura realizada, o estudo comparativo das ferramentas de XAI, o aprofundamento na questão de avaliação da explicabilidade, por fim a definição da estratégia de uso e desenvolvimento da ferramenta.

A RSL realizada além de mostrar um panorama da área de pesquisa e permitir que futuros pesquisadores interessados na área possam utilizar e se necessário atualizar os dados coletados, mostrou um caminho para uma oportunidade de pesquisa, utilizando dados de exercícios de reabilitação juntamente com aspectos de XAI e Visualização da Informação. Mostrou também que mesmo com área de pesquisa evoluindo nos últimos anos, existe uma lacuna no processo da avaliação da explicabilidade buscada.

O comparativo de diversas ferramentas de XAI disponíveis serve como base para escolha de qual abordagem ou biblioteca se deseja utilizar, no presente projeto e em projetos desenvolvidos por outros pesquisadores no futuro.

O aprofundamento na questão de avaliação da explicabilidade visa explorar essa área nova de pesquisa, indicada na RSL, visou compreender como o público alvo selecionado

entende e considera a explicabilidade alcançada, considerando o processo de entendimento humano.

A definição da estratégia de uso inclui como resultado relevante a uso de uma nova metáfora visual, utilizando um avatar para demonstração dos movimentos associado aos dados de impacto das articulações e dos *frames* na classificação do movimento. É uma abordagem que une diversas áreas de pesquisa: IA, XAI, Visualização da Informação, Reabilitação Neuromotora, de forma a auxiliar o profissional da área de reabilitação.

E por fim, o desenvolvimento da ferramenta com o apoio de um colaborador pesquisador da área de reabilitação, validando a utilidade da estratégia e demandando recursos que direcionaram o desenvolvimento da ferramenta, inclusive permitindo o *upload* de uma amostra escolhida, permitindo um nível maior de interação com o usuário, através do envio de dados próprios, desde que padronizados em CSV com a estrutura adequada.

8.5 Considerações finais do capítulo

O presente capítulo buscou expor os resultados obtidos no desenvolvimento do estudo, atendendo o objetivo definido e demonstrando novas possibilidades de uso de XAI, em especial na área de reabilitação neuromotora.

9 Conclusão

9.1 Considerações iniciais do capítulo

No presente capítulo, buscando concluir esta tese, será apresentada uma breve discussão considerando os pontos relevantes e as limitações do presente estudo. Em seguida, serão apresentadas as possibilidades de continuidade do projeto através de trabalhos futuros.

9.2 Relevância

A presente tese se mostra relevante por tratar de uma área nova, XAI, porém, que vem ganhando força nos últimos anos, juntamente com as discussões sobre a utilização justa da IA nos processos de classificação e decisão. Colaborando com essa relevância, tem-se o estudo de caso na área da saúde, uma das áreas que mais demandam um aumento na confiança em decisões auxiliadas por computador.

A estratégia apresentada demonstra resultados consistentes que podem ser observados com o uso da ferramenta na análise das sessões de reabilitação. É possível analisar as articulações e os momentos mais relevantes em cada amostra, obtendo informações sobre o impacto destes fatores na classificação do movimento executado na amostra. O interesse do colaborador pesquisador da área de reabilitação nos testes da ferramenta comprovam que a estratégia de explicabilidade pode fornecer dados úteis no apoio ao processo de reabilitação.

A metáfora visual apresentada na estratégia desenvolvida e implementada para avaliação através da ferramenta, associando um avatar representando os movimentos aos dados de explicabilidade é um ponto de destaque do estudo, visto que uma abordagem combinando explicabilidade e Visualização da Informação dessa forma ainda não havia sido proposta.

A utilização da biblioteca TimeSHAP no contexto de saúde, em especial na reabilitação, expande o uso da biblioteca e também as possibilidades de utilização de XAI em dados com características temporais.

A utilização de diversos conjunto de dados, inicialmente rebase e UI-PRMD e posteriormente o conjunto de dados fornecido pelo colaborador pesquisador da área de reabilitação e a opção de *upload* demonstra um bom nível de flexibilidade e adaptação da ferramenta desenvolvida, possibilitando uma iteração e exploração maior do usuário.

Outro ponto que colabora para relevância deste estudo é sua consistência, realizando uma revisão de literatura, avaliando as opções de ferramentas disponíveis, se aprofundando na questão de avaliação e por fim apresentando uma estratégia de uso de XAI na reabilitação neuromotora.

A RSL realizada forneceu dados relevantes e úteis durante o restante do desenvolvimento da estratégia. A comparação de ferramentas de XAI proporciona um conhecimento maior da área de pesquisa, avaliação diferentes opções e abordagens no uso de XAI. O aprofundamento na questão de avaliação de XAI identifica uma área pouco explorada, gerando oportunidades de pesquisas futuras, com a definição de métricas e métodos.

9.3 Limitações

Uma vez que XAI é um campo de pesquisa novo, existe a limitação do processo de uso e em especial de avaliação, não estarem bem definidos na literatura disponível, sendo necessário um aprofundamento nessas temáticas, buscando resultados positivos de maneira empírica.

As possibilidades de aplicação de XAI são enormes, porém, por se tratar de um projeto ligado a um curso de Doutorado, com prazo e escopo limitados, optou-se por focar os esforços no estudo de caso envolvendo a classificação de movimentos para auxílio na reabilitação neuromotora, sem buscar novas áreas para desenvolvimento de aplicações práticas no momento.

Uma limitação apresentada está relacionada com a necessidade de se realizar testes com usuários, principalmente durante o processo de avaliação. Os testes com usuários são possíveis, porém podem gerar uma grande carga de trabalho burocrático e podem não ser aprovados pelos comitês responsáveis por avaliar esse tipo de procedimento, dessa forma optou-se por não incluir essa etapa no presente estudo, porém a questão será tratada nos trabalhos futuros.

9.4 Trabalhos futuros

Para continuidade do presente estudo, a ferramenta que compõe a estratégia desenvolvida deverá ganhar novos recursos conforme demandados pelo colaborador pesquisador da área de reabilitação que está realizando os testes, expandindo assim os usos da estratégia.

Após constatado um nível satisfatório de maturidade na ferramenta, a mesma será disponibilizada para um grupo maior de profissionais, permitindo inclusive um estudo de avaliação da explicabilidade como aprofundado no decorrer do presente estudo. Para esse estudo de avaliação, pretende-se aprofundar a discussão sobre métricas e ferramentas para

avaliação de XAI, realizando experimentos com questionários e outras ferramentas que se mostrarem adequadas. Com a disponibilização da ferramenta para mais profissionais, existe a possibilidade de realização de uma avaliação quantitativa simples, indicando objetivamente se a explicação obtida para cada amostra é correta do ponto de vista do profissional, considerando a importância de cada articulação no movimento executado, por exemplo.

Para fortalecer a validação da estratégica, serão buscados novos conjuntos de dados para aplicação da estratégia, de forma a demonstrar a flexibilidade e capacidade de adaptação da mesma.

No desenvolvimento do presente estudo, percebeu-se que a estratégia proposta se mostra flexível, podendo ser utilizada em todo contexto de movimentos, não necessariamente se restringindo a movimentos de reabilitação. Dessa forma, como possibilidade de extensão da pesquisa, pode-se utilizar a estratégia de forma mais abrangente, adaptando a mesma para outros contextos.

A explicabilidade obtida possui foco no público alvo dos profissionais da área de reabilitação, ainda que possa ser interessante para outros públicos: pesquisadores da área de Computação com ênfase em Inteligência Artificial, pacientes que possam ter seus dados avaliados e se interessem pelo processo. Como trabalhos futuros podem ser realizados novos estudos, expandido a estratégia para abranger novos públicos alvos.

Com os resultados atuais e futuros, se espera produzir material para elaboração de artigos científicos relevantes.

9.5 Considerações finais

Tratar de XAI, seja no uso ou na avaliação, envolve um esforço multidisciplinar, envolvendo aspectos de computação, no domínio de aprendizado de máquina e na interação humano-computador, tratando a interface do sistema com os usuários, aspectos de visualização da informação, gerando imagens e gráficos para colaborar na compreensão, além de aspectos psicológicos e de comportamento. Espera-se que esta tese tenha abordado um pouco estes aspectos, mediante todas suas etapas, gerando uma visão global do contexto e finalizando com a estratégia de uso de XAI no apoio à reabilitação neuromotora.

9.6 Publicações no período do doutorado

Esta seção apresenta as publicações científicas produzidas durante o período do doutorado pelo candidato. Nem todas foram diretamente relacionadas à proposta apresentada nesta tese e sua temática, mas todas estiveram relacionadas aos seus interesses de pesquisa.

CAMARGO, L. F.; DIAS, D. R. C ; BREGA, J. R. F. . Implementation of eXplainable Artificial Intelligence. Lecture Notes in Computer Science. 1ed.: Springer International Publishing, 2023, v. 13956, p. 564-580. Link: https://link.springer.com/chapter/10.1007/978-3-031-36805-9_37 - Evento: Computational Science and Its Applications - ICCSA 2023: 23rd International Conference, Athens, Greece, July 3–6, 2023

CAMARGO, L. F.; BREGA, J. R. F. . Estratégia de Uso de Inteligência Artificial Explicável: Estudo de Caso na Classificação de Movimentos para Apoio a Reabilitação Neuromotora. In: XII Workshop do Programa de Pós-Graduação em Ciência da Computação da Unesp (WPPGCC 2023), 2023, Online. XII Workshop do Programa de Pós-Graduação em Ciência da Computação da Unesp (WPPGCC 2023), 2023.

CAMARGO, L. F.; FEITOSA, J. C. ; BONATTI, E. ; SIMIONI, G. B. ; BREGA, J. R. F. . eXplainable Artificial Intelligence - A Study of Sentiments About Vaccination in Brazil. Lecture Notes in Computer Science. 1ed.: Springer International Publishing, 2023, v. 13956', p. 617-634. Link: https://link.springer.com/chapter/10.1007/978-3-031-36805-9_40

CAMARGO, LUIZ FELIPE DE; REIS, CARLOS ; PAIOLA, PEDRO HENRIQUE ; PAPA, JOÃO PAULO ; BREGA, JOSÉ REMO F. ; COSTA, KELTON A. P. DA . Métodos de Aprendizado de Máquina Adversariais na Detecção de Anomalias em Redes de Computadores. In: Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais, 2021, Brasil. Anais do XXI Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg 2021), 2021. p. 169. Link: <https://sol.sbc.org.br/index.php/sbseg/article/view/17314>

REQUENA, ADRIANO ; DA COSTA FEITOSA, JULIANA ; DE CAMARGO, LUIZ FELIPE ; FERREIRA BREGA, JOSE REMO . Using Information Visualization Techniques for Fish Genetic Management. In: 2022 26th International Conference Information Visualisation (IV), 2022, Vienna. 2022 26th International Conference Information Visualisation (IV), 2022. p. 204. Link: <https://ieeexplore.ieee.org/document/10017926>

Referências

- ABUSELIDZE, G.; MAMALADZE, L. The impact of artificial intelligence on employment before and during pandemic: A comparative analysis. *Journal of Physics: Conference Series*, IOP Publishing, v. 1840, n. 1, p. 012040, mar 2021. ISSN 1742-6596. Disponível em: <<https://iopscience.iop.org/article/10.1088/1742-6596/1840/1/012040https://iopscience.iop.org/article/10.1088/1742-6596/1840/1/012040/meta>>. Citado na página 20.
- ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 6, p. 52138–52160, sep 2018. ISSN 21693536. Citado na página 34.
- ALI, S. et al. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, v. 99, p. 101805, 2023. ISSN 1566-2535. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1566253523001148>>. Citado na página 30.
- ARRIETA, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, v. 58, p. 82–115, 2020. ISSN 1566-2535. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1566253519308103>>. Citado 8 vezes nas páginas 28, 30, 31, 33, 34, 35, 36 e 48.
- ARYA, V. et al. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. 2019. Disponível em: <<https://arxiv.org/abs/1909.03012>>. Citado na página 60.
- AXEL, C. et al. Engineering rehabilitation: Blending two tool-supported approaches to close the loop from tasks-based rehabilitation to exercises and back again. *Proc. ACM Hum.-Comput. Interact.*, Association for Computing Machinery, v. 7, 6 2023. Disponível em: <<https://doi.org/10.1145/3593229>>. Citado na página 132.
- BENTO, J. et al. Timeshap: Explaining recurrent models through sequence perturbations. In: . [S.l.]: Association for Computing Machinery, 2021. p. 2565–2573. ISBN 9781450383325. Citado na página 92.
- BORYS, K. et al. Explainable ai in medical imaging: An overview for clinical practitioners – saliency-based xai approaches. *European Journal of Radiology*, v. 162, p. 110787, 2023. ISSN 0720-048X. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0720048X23001018>>. Citado na página 37.
- BRERETON, P. et al. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, Elsevier Science Inc., v. 80, n. 4, p. 571–583, 4 2007. ISSN 01641212. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S016412120600197X>>. Citado na página 39.
- BROWN, T. B. et al. *Language Models are Few-Shot Learners*. 2020. Disponível em: <<https://arxiv.org/abs/2005.14165>>. Citado na página 28.

- BRUNO, F.; PEREIRA, P. C.; FALTAY, P. Inteligência artificial e saúde: ressituar o problema. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, Instituto de Comunicacao e Informacao Cientifica e Tecnologica em Saude, v. 17, p. 235–242, 6 2023. Citado na página 37.
- CAMARGO, L. F. de; DIAS, D. R. C.; BREGA, J. R. F. Implementation of explainable artificial intelligence case study on the assessment of movements to support neuromotor rehabilitation. In: GERVASI, O. et al. (Ed.). [S.l.: s.n.], 2023. v. 13956, p. 564–580. ISBN 978-3-031-36804-2; 978-3-031-36805-9. ISSN 0302-9743. 23rd International Conference on Computational Science and its Applications (ICCSA), Natl Tech Univ Athens, Athens, GREECE, JUL 03-06, 2023. Citado 2 vezes nas páginas 52 e 130.
- CAMBRIA, E. et al. Xai meets llms: A survey of the relation between explainable ai and large language models. 7 2024. Disponível em: <<http://arxiv.org/abs/2407.15248>>. Citado na página 28.
- CAMPAGNINI, S. et al. Cross-validation of predictive models for functional recovery after post-stroke rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, v. 19, 2022. Cited by: 14; All Open Access, Gold Open Access, Green Open Access. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137468804&doi=10.1186%2fs12984-022-01075-7&partnerID=40&md5=eaf14f57a1aca98b22056e1bcf6c2ae7>>. Citado 2 vezes nas páginas 51 e 128.
- CARBONERA, J.; GONÇALVES, B.; SOUZA, C. de. O problema da explicação em inteligência artificial: considerações a partir da semiótica. v. 17, p. 59–75, 2023. ISSN 1984-3585. Citado na página 28.
- CARVALHO, A. C. P. de Leon FERREIRA de. Inteligência artificial: riscos, benefícios e uso responsável. *Estudos Avançados*, Instituto de Estudos Avançados da Universidade de São Paulo, v. 35, p. 21–36, 4 2021. ISSN 1806-9592. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142021000100021&tlng=pt>. Citado na página 20.
- CHEN, M. et al. An interpretable deep learning optimized wearable daily detection system for parkinson’s disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, v. 31, p. 3937 – 3946, 2023. Cited by: 3; All Open Access, Gold Open Access. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171751984&doi=10.1109%2fTNSRE.2023.3314100&partnerID=40&md5=a969394cb8e22b208943ab990edf6d6f>>. Citado 2 vezes nas páginas 51 e 128.
- CHEN, S.; YANG, R. R. Pose trainer: Correcting exercise posture using pose estimation. *arXiv: Computer Vision and Pattern Recognition*, 6 2020. Disponível em: <<http://arxiv.org/abs/2006.11718>>. Citado na página 79.
- CHEN, Y. et al. Dropoutseer: Visualizing learning patterns in massive open online courses for dropout reasoning and prediction. In: *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. [S.l.: s.n.], 2016. p. 111–120. Citado na página 24.
- CONNER, B. C. et al. Factors influencing neuromuscular responses to gait training with a robotic ankle exoskeleton in cerebral palsy. *ASSISTIVE TECHNOLOGY*, v. 35, p. 463–470, 11 2023. ISSN 1040-0435. Citado 2 vezes nas páginas 52 e 130.

DECAMP, M.; LINDVALL, C. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*, v. 27, n. 12, p. 2020–2023, 06 2020. ISSN 1527-974X. Disponível em: <<https://doi.org/10.1093/jamia/ocaa094>>. Citado na página 21.

DITTAKAVI, B. et al. Pose tutor: An explainable system for pose correction in the wild. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.]: IEEE, 2022. p. 3539–3548. ISBN 978-1-6654-8739-9. Citado na página 79.

FJELD, J. et al. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*, Elsevier BV, jan 2020. ISSN 1556-5068. Disponível em: <<https://papers.ssrn.com/abstract=3518482>>. Citado na página 35.

FRANCISCO, P. *Mensagem do Santo Padre Francisco para a Celebração do Dia Mundial da Paz 2024*. 2024. Disponível em: <<https://www.vatican.va/content/francesco/pt/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html>>. Citado na página 24.

GANDOLFI, M. et al. explainable ai allows predicting upper limb rehabilitation outcomes in sub-acute stroke patients. *IEEE Journal of Biomedical and Health Informatics*, v. 27, p. 263 – 273, 2023. Cited by: 15. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141606409&doi=10.1109%2fJBHI.2022.3220179&partnerID=40&md5=47174bb0378ea3178e20d18701ad7f96>>. Citado 2 vezes nas páginas 52 e 129.

GILLATH, O. et al. Attachment and trust in artificial intelligence. *Computers in Human Behavior*, v. 115, p. 106607, 2021. ISSN 0747-5632. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S074756322030354X>>. Citado na página 20.

GOASDUFF, L. *2018 Will Mark the Beginning of AI Democratization*. 2017. Disponível em: <<https://www.gartner.com/smarterwithgartner/2018-will-mark-the-beginning-of-ai-democratization/>>. Citado na página 24.

GOODMAN, B.; FLAXMAN, S. European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”. *AI Magazine*, AI Access Foundation, v. 38, n. 3, p. 50–57, oct 2017. ISSN 2371-9621. Disponível em: <<https://ojs.aaai.org/index.php/aimagazine/article/view/2741>>. Citado na página 29.

GUNNING, D. et al. XAI-Explainable artificial intelligence. *Science Robotics*, American Association for the Advancement of Science, v. 4, n. 37, dec 2019. ISSN 24709476. Citado na página 30.

GUNNING, D. et al. Darpa’s explainable ai (xai) program: A retrospective. *Applied AI Letters*, John Wiley & Sons, Ltd, v. 2, p. e61, 12 2021. ISSN 2689-5595. Disponível em: <<https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61https://onlinelibrary.wiley.com/doi/10.1002/ail2.61>>. Citado na página 21.

- HAIBE-KAINS, B. et al. Transparency and reproducibility in artificial intelligence. *Nature*, v. 586, n. 7829, p. E14–E16, Oct 2020. ISSN 1476-4687. Disponível em: <https://doi.org/10.1038/s41586-020-2766-y>. Citado na página 20.
- HEALTH, N. I. of. *Stroke* / *NHLBI, NIH*. 2020. Disponível em: <https://www.nhlbi.nih.gov/health-topics/stroke>. Citado na página 37.
- HESSEL, A. M. D. G.; LEMES, D. de O. Criatividade da inteligência artificial generativa. *TECCOGS: Revista Digital de Tecnologias Cognitivas*, p. 119–130, 6 2024. ISSN 1984-3585. Disponível em: <https://revistas.pucsp.br/index.php/teccogs/article/view/67075>. Citado na página 20.
- HINTON, G. E. et al. Improving neural networks by preventing co-adaptation of feature detectors. jul 2012. Disponível em: <https://arxiv.org/abs/1207.0580>. Citado na página 84.
- HOFFMAN, R. R. et al. Metrics for Explainable AI: Challenges and Prospects. dec 2018. ISSN 2331-8422. Disponível em: <http://arxiv.org/abs/1812.04608>. Citado 6 vezes nas páginas 33, 74, 75, 76, 77 e 78.
- HOUSE, T. W. *Biden-Harris Administration Announces Key AI Actions 180 Days Following President Biden's Landmark Executive Order*. 2024. Disponível em: <https://www.whitehouse.gov/briefing-room/statements-releases/2024/04/29/biden-harris-administration-announces-key-ai-actions-180-days-following-president-bidens-landmark-e>>. Citado na página 24.
- JOHN, N. M.; SANGEETHA, R. Stroke rehabilitation using virtual reality. In: *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. [S.l.: s.n.], 2024. p. 1033–1040. Citado na página 38.
- JON, M. *GitHub - zetane/viewer: ML models and internal tensors 3D visualizer*. 2021. Disponível em: <https://github.com/zetane/viewer>. Citado 2 vezes nas páginas 65 e 66.
- JUNIOR, E. A.; NAKAYA, H.; RIZZO, L. V. Inteligência artificial em saúde. *Revista USP*, n. 141, p. 41–50, jun. 2024. Disponível em: <https://www.revistas.usp.br/revusp/article/view/225206>. Citado na página 36.
- KANTAYYA, S. *Coded Bias*. Netflix, 2020. Disponível em: <https://www.netflix.com/br/title/81328723>. Citado na página 23.
- KELLEY, D.; KELLEY, T. *Creative confidence : unleashing the creative potential within us all*. [S.l.]: Crown Business, 2013. 288 p. ISBN 9780385349369. Citado na página 21.
- KELLIHER, A. et al. Designing modular rehabilitation objects for interactive therapy in the home. In: . Association for Computing Machinery, 2019. p. 251–257. ISBN 9781450361965. Disponível em: <https://doi.org/10.1145/3294109.3300983>. Citado na página 132.
- KIE-TEC. *O que Inteligência artificial explicável tem a ver com Tableau? / Kie-tec*. 2019. Disponível em: <https://www.kietec.com.br/o-que-e-inteligencia-artificial-explicavel-e-o-que-ela-tem-a-ver-com-o-tableau/>>. Citado na página 28.

- KINGERY, K. *Duke Professor Wins \$1 Million Artificial Intelligence Prize, A 'New Nobel' / Duke Pratt School of Engineering*. 2021. Disponível em: <<https://pratt.duke.edu/about/news/rudin-squirrel-award>>. Citado na página 24.
- KLEIN, G.; HOFFMAN, R. R. Macrocognition, mental models, and cognitive task analysis methodology. *Naturalistic Decision Making and Macrocognition*, p. 57–80, 2008. Citado na página 76.
- KOKHLIKYAN, N. et al. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. Citado 2 vezes nas páginas 62 e 63.
- KOLARIK, M. et al. Explainability of deep learning models in medical video analysis: a survey. *PeerJ Computer Science*, PeerJ Inc., v. 9, p. 1–39, 2023. ISSN 23765992. Tabela. Citado na página 37.
- KOROBOV, M. *Overview — ELI5 0.11.0 documentation*. 2017. Disponível em: <<https://eli5.readthedocs.io/en/latest/overview.html>>. Citado 2 vezes nas páginas 59 e 60.
- LAURIERE, J. L. *Problem-solving and artificial intelligence*. [S.l.]: Prentice Hall, 1990. 506 p. ISBN 0137117485. Citado na página 27.
- LEINO, K. *A Hands-on Introduction to Explaining Neural Networks with TruLens / by Klas Leino / TruLens / Medium*. 2021. Disponível em: <<https://medium.com/trulens/a-hands-on-introduction-to-explaining-neural-networks-with-trulens-504bfab1a578>>. Citado na página 64.
- LEINO, K. et al. *truera/trulens: TruLens*. sep 2021. Disponível em: <<https://doi.org/10.5281/zenodo.5503946#.YW9fB4C-PNA.mendeley>>. Citado na página 63.
- LENT, M. van; FISHER, W.; MANCUSO, M. An explainable artificial intelligence system for small-unit tactical behavior. In: *Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence*. [S.l.]: AAAI Press, 2004. (IAAI'04), p. 900–907. ISBN 0262511835. Citado 2 vezes nas páginas 77 e 82.
- LEO, M. et al. Video-based automatic baby motion analysis for early neurological disorder diagnosis: State of the art and future directions. *Sensors*, v. 22, 2022. ISSN 1424-8220. Disponível em: <<https://www.mdpi.com/1424-8220/22/3/866>>. Citado na página 37.
- LIN, P.-J. et al. A transferable deep learning prognosis model for predicting stroke patients' recovery in different rehabilitation trainings. *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, v. 26, p. 6003–6011, 12 2022. ISSN 2168-2194. Citado 2 vezes nas páginas 51 e 128.
- LIUZZI, P. et al. Eeg fractal dimensions predict high-level behavioral responses in minimally conscious patients. *Journal of Neural Engineering*, v. 20, 2023. Cited by: 2. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168805730&doi=10.1088%2F1741-2552%2Ffaceaac&partnerID=40&md5=3846ff0a7b13880b123ce538d3f68bff>>. Citado 2 vezes nas páginas 51 e 129.
- LIUZZI, P. et al. Predicting outcome of patients with prolonged disorders of consciousness using machine learning models based on medical complexity. *SCIENTIFIC REPORTS*, v. 12, 8 2022. ISSN 2045-2322. Citado 2 vezes nas páginas 52 e 131.

- LOPES, P. et al. Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, v. 12, n. 19, 2022. ISSN 2076-3417. Disponível em: <<https://www.mdpi.com/2076-3417/12/19/9423>>. Citado 3 vezes nas páginas 25, 74 e 78.
- LUGARESI, C. et al. MediaPipe: A Framework for Building Perception Pipelines. jun 2019. Disponível em: <<http://arxiv.org/abs/1906.08172>>. Citado na página 83.
- LUNDBERG, S. *GitHub - slundberg/shap: A game theoretic approach to explain the output of any machine learning model*. 2021. Disponível em: <<https://github.com/slundberg/shap>>. Citado 2 vezes nas páginas 56 e 57.
- LUNDBERG, S. M.; LEE, S. I. A unified approach to interpreting model predictions. *arXiv*, n. Section 2, p. 1–10, 2017. ISSN 23318422. Citado 2 vezes nas páginas 47 e 57.
- MARINO, D. L.; WICKRAMASINGHE, C. S.; MANIC, M. An adversarial approach for explainable ai in intrusion detection systems. In: *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2018. p. 3237–3243. ISBN 978-1-5090-6684-1. Disponível em: <<https://ieeexplore.ieee.org/document/8591457/>>. Citado na página 34.
- MASÍS, S. *Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples*. Packt, 2021. ISBN 1800206577. Disponível em: <https://books.google.com/books/about/Interpretable_Machine_Learning_with_Pyth.html?hl=pt-BR&id=iIwmEAAQBAJ>. Citado 3 vezes nas páginas 29, 34 e 36.
- MCCARTHY, J. *WHAT IS ARTIFICIAL INTELLIGENCE?* [S.l.], 2007. Disponível em: <<http://www-formal.stanford.edu/jmc/>>. Citado na página 27.
- MCCAY, K. D. et al. Towards explainable abnormal infant movements identification: A body-part based prediction and visualisation framework. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. ISBN 9781665403580. Citado na página 37.
- MEHRABI, N. et al. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 6, jul 2021. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/3457607>>. Citado na página 22.
- MILLER, K. *Should AI Models Be Explainable? That depends*. 2021. Disponível em: <<https://hai.stanford.edu/news/should-ai-models-be-explainable-depends>>. Citado 2 vezes nas páginas 31 e 32.
- MOHSENI, S.; ZAREI, N.; RAGAN, E. D. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. nov 2018. Disponível em: <<https://arxiv.org/abs/1811.11839>>. Citado 3 vezes nas páginas 74, 76 e 77.
- NAYEBI, A. et al. An empirical comparison of explainable artificial intelligence methods for clinical data: A case study on traumatic brain injury. 8 2022. Disponível em: <<https://arxiv.org/abs/2208.06717>>. Citado na página 82.

NAYEBI, A. et al. Windowshap: An efficient framework for explaining time-series classifiers based on shapley values. *Journal of Biomedical Informatics*, v. 144, p. 104438, 2023. ISSN 1532-0464. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1532046423001594>>. Citado na página 92.

NEIVA, F.; SILVA, R. Revisão sistemática da literatura em ciência da computação - um guia prático. 06 2016. Citado 2 vezes nas páginas 41 e 42.

NETO, M. G. *Explainable AI: extraíndo explicações e aumentando a confiança dos modelos de ML* / by Milton Gama Neto / Data Hackers / Apr, 2021 / Medium. 2021. Disponível em: <<https://medium.com/data-hackers/explainable-ai-extraíndo-explicações-e-aumentando-a-confiança-dos-modelos-de-ml-3a89b7b5a58>>. Citado na página 29.

NETRON. *GitHub - lutzroeder/netron: Visualizer for neural network, deep learning, and machine learning models*. 2021. Disponível em: <<https://github.com/lutzroeder/Netron>>. Citado na página 65.

NIGAR, N. et al. A deep learning approach based on explainable artificial intelligence for skin lesion classification. *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 10, p. 113715–113725, 2022. ISSN 21693536. Citado na página 56.

OPENAI et al. *GPT-4 Technical Report*. 2024. Disponível em: <<https://arxiv.org/abs/2303.08774>>. Citado na página 28.

OZGUR, A. G. et al. Iterative design of an upper limb rehabilitation game with tangible robots. In: . Association for Computing Machinery, 2018. p. 241–250. ISBN 9781450349536. Disponível em: <<https://doi.org/10.1145/3171221.3171262>>. Citado na página 132.

PANDIT, S. et al. Exercisecheck: a scalable platform for remote physical therapy deployed as a hybrid desktop and web application. In: . Association for Computing Machinery, 2019. p. 101–109. ISBN 9781450362320. Disponível em: <<https://doi.org/10.1145/3316782.3321537>>. Citado na página 132.

PARSIFAL. *About · Parsifal*. 2018. Disponível em: <<https://parsif.al/about/>>. Citado na página 44.

PAVLLO, D. et al. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.]: IEEE, 2019. p. 7745–7754. ISBN 978-1-7281-3293-8. Citado na página 79.

POLLO, L. *Cidade-berço do MIT bane reconhecimento facial; decisão tem peso simbólico - 16/01/2020 - UOL TAB*. 2020. Disponível em: <<https://tab.uol.com.br/noticias/redacao/2020/01/16/cambridge-bane-reconhecimento-facial-entenda-por-que-temos-de-pensar-nisso.htm>>. Citado na página 23.

POURSABZI-SANGDEH, F.; GOLDSTEIN, D. G.; HOFMAN, J. M. Manipulating and measuring model interpretability. *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, v. 67, feb 2021. Disponível em: <<https://arxiv.org/abs/1802.07810v5>>. Citado na página 33.

- RAMESH, A. et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. Disponível em: <<https://arxiv.org/abs/2204.06125>>. Citado na página 28.
- REGO, P.; MOREIRA, P. M.; REIS, L. P. Serious games for rehabilitation: A survey and a classification towards a taxonomy. In: *5th Iberian Conference on Information Systems and Technologies*. [S.l.: s.n.], 2010. p. 1–6. Citado na página 38.
- RIBEIRO, E. H. et al. Biomechanics sensor nodes for body tracking: a development solution for virtual reality interaction. In: *Anais do XXIII Simpósio de Realidade Virtual e Aumentada*. Porto Alegre, RS, Brasil: SBC, 2021. p. 119–125. Disponível em: <<https://sol.sbc.org.br/index.php/svr/article/view/17526>>. Citado na página 83.
- RIBEIRO, M. T.; SINGH, S.; GUESTIN, C. "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, v. 13-17-Augu, p. 1135–1144, 2016. Citado 3 vezes nas páginas 47, 54 e 55.
- RIBEIRO, M. T. C. *GitHub - marcotcr/lime: Lime: Explaining the predictions of any machine learning classifier*. 2021. Disponível em: <<https://github.com/marcotcr/lime>>. Citado na página 58.
- RICH, E.; KNIGHT, K. *Inteligencia artificial*. [S.l.]: McGraw-Hill, 1994. 703 p. ISBN 8448118588. Citado na página 27.
- RIGHTS, U. N. H. *OHCHR / Artificial intelligence risks to privacy demand urgent action – Bachelet*. 2021. Disponível em: <<https://www.ohchr.org/EN/NewsEvents/Pages/media.aspx?IsMediaPage=true>>. Citado na página 23.
- RIKAKIS, T. et al. Semi-automated home-based therapy for the upper extremity of stroke survivors. In: . Association for Computing Machinery, 2018. p. 249–256. ISBN 9781450363907. Disponível em: <<https://doi.org/10.1145/3197768.3197777>>. Citado na página 132.
- RODRIGUES, L. G. S. et al. Classification of human movements with motion capture data in a motor rehabilitation context. In: *Anais do XXIII Simpósio de Realidade Virtual e Aumentada*. Porto Alegre, RS, Brasil: SBC, 2021. p. 55–62. Disponível em: <<https://sol.sbc.org.br/index.php/svr/article/view/17519>>. Citado 3 vezes nas páginas 10, 83 e 85.
- RODRIGUES, L. G. S. et al. Upper Limb Motion Tracking and Classification: A Smartphone Approach. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: Association for Computing Machinery, 2021. (WebMedia '21), p. 61–64. ISBN 9781450386098. Disponível em: <<https://doi.org/10.1145/3470482.3479618>>. Citado na página 83.
- RODRIGUES, L. G. S. et al. Supervised classification of motor-rehabilitation body movements with rgb cameras and pose tracking data. *Journal on Interactive Systems*, Sociedade Brasileira de Computacao - SB, v. 13, p. 221–231, 9 2022. Citado 3 vezes nas páginas 24, 38 e 83.
- RODRIGUES, L. G. S. et al. Upper limb motion tracking and classification: A smartphone approach. In: . Association for Computing Machinery, 2021. p. 61–64. ISBN

9781450386098. Disponível em: <<https://doi.org/10.1145/3470482.3479618>>. Citado 2 vezes nas páginas 52 e 131.

ROTH, A. E. The Shapley Value. *The Shapley Value*, Cambridge University Press, oct 1988. Disponível em: <<https://www.cambridge.org/core/books/shapley-value/D3829B63B5C3108EFB62C4009E2B966E>>. Citado na página 54.

RUDIN, C. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Research, 2019. 206–215 p. Disponível em: <<https://www.nature.com/articles/s42256-019-0048-x>>. Citado 2 vezes nas páginas 34 e 43.

RUSSEL, S.; NORVIG, P. *Artificial intelligence—a modern approach 3rd Edition*. Prentice Hall, 2012. v. 11. 78 p. ISSN 0269-8889. ISBN 0136042597. Disponível em: <http://www.journals.cambridge.org/abstract{_}S0269888900007>. Citado na página 27.

SAHARIA, C. et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. Disponível em: <<https://arxiv.org/abs/2205.11487>>. Citado na página 28.

SAKKOS, D. et al. Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy. *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 9, p. 94281–94292, 2021. ISSN 21693536. Neste artigo, propomos uma nova estrutura de aprendizado profundo para essa tarefa de classificação. Também propomos uma estrutura de visualização que identifica as partes do corpo com maior contribuição para uma decisão de classificação. Achar código Solução própria, sem lime ou shap. Citado na página 37.

SARAE, E. et al. Exercisecheck: data analytics for a remote monitoring and evaluation platform for home-based physical therapy. In: . Association for Computing Machinery, 2019. p. 110–118. ISBN 9781450362320. Disponível em: <<https://doi.org/10.1145/3316782.3316790>>. Citado na página 132.

SARKAR, T. *Google’s new ‘Explainable AI’ (xAI) service - Towards Data Science*. 2019. Disponível em: <<https://towardsdatascience.com/googles-new-explainable-ai-xai-service-83a7bc823773>>. Citado 2 vezes nas páginas 20 e 21.

SCHNEEBERGER, D. et al. The tower of babel in explainable artificial intelligence (xai). In: HOLZINGER, A. et al. (Ed.). *Machine Learning and Knowledge Extraction*. Cham: Springer Nature Switzerland, 2023. p. 65–81. ISBN 978-3-031-40837-3. Citado na página 30.

SIEAN, A.-I.; VATAVU, R.-D. Wearable interactions for users with motor impairments: Systematic review, inventory, and research implications. In: . Association for Computing Machinery, 2021. ISBN 9781450383066. Disponível em: <<https://doi.org/10.1145/3441852.3471212>>. Citado na página 50.

SISK, M. et al. Analyzing xai metrics: Summary of the literature review. Institute of Electrical and Electronics Engineers (IEEE), out. 2022. Disponível em: <<http://dx.doi.org/10.36227/techrxiv.21262041.v1>>. Citado 2 vezes nas páginas 25 e 78.

SKATER. *Model Interpretation with Skater — skater 0 documentation*. 2017. Disponível em: <<https://oracle.github.io/Skater/index.html>>. Citado 2 vezes nas páginas 61 e 62.

SUNAK, R.; DONELAN, M.; CLEVERLY, J. *Countries agree to safe and responsible development of frontier AI in landmark Bletchley Declaration*. 2023. Disponível em: <<https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration>>. Citado na página 24.

TEAM, E. *Detecting bad posture with machine learning*. 2023. Disponível em: <<https://towardsai.net/p/machine-learning/detecting-bad-posture-with-machine-learning>>. Citado na página 80.

TJOA, E.; GUAN, C. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, Institute of Electrical and Electronics Engineers (IEEE), v. 14, n. 8, p. 1–21, jul 2019. Disponível em: <<https://arxiv.org/abs/1907.07374v5>>. Citado na página 29.

TOUVRON, H. et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. Disponível em: <<https://arxiv.org/abs/2302.13971>>. Citado na página 28.

TROTTA, T. et al. Rebase: data acquisition and management system for neuromotor rehabilitation supported by virtual and augmented reality. In: *Anais do XXIII Simpósio de Realidade Virtual e Aumentada*. Porto Alegre, RS, Brasil: SBC, 2021. p. 181–185. Disponível em: <<https://sol.sbc.org.br/index.php/svr/article/view/17535>>. Citado 3 vezes nas páginas 82, 84 e 98.

TSUR, E. E.; ELKANA, O. Intelligent robotics in pediatric cooperative neurorehabilitation: A review. *ROBOTICS*, v. 13, 3 2024. Citado na página 50.

TUREK, M. *Explainable Artificial Intelligence*. 2018. Disponível em: <<https://www.darpa.mil/program/explainable-artificial-intelligence>>. Citado na página 21.

VAKANSKI, A. et al. A data set of human body movements for physical rehabilitation exercises. *Data*, v. 3, n. 1, 2018. ISSN 2306-5729. Disponível em: <<https://www.mdpi.com/2306-5729/3/1/2>>. Citado na página 97.

VALENTE, F. R. et al. A framework for neuromotor and neurofunctional rehabilitation using multi-agent systems. In: . Association for Computing Machinery, 2024. p. 142–150. ISBN 9798400700026. Disponível em: <<https://doi.org/10.1145/3604479.3604524>>. Citado na página 132.

VALENTE, J. *Europa lança diretrizes éticas para o uso da inteligência artificial | Agência Brasil*. 2019. Disponível em: <<http://agenciabrasil.ebc.com.br/internacional/noticia/2019-04/europa-lanca-diretrizes-eticas-para-o-uso-da-inteligencia-artificial>>. Citado na página 24.

VISUALDL. *GitHub - PaddlePaddle/VisualDL: Deep Learning Visualization Toolkit*. 2021. Disponível em: <<https://github.com/PaddlePaddle/VisualDL>>. Citado na página 64.

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, Oxford Academic, v. 7, n. 2, p. 76–99, may 2017. ISSN

20444001. Disponível em: <<https://academic.oup.com/idpl/article/7/2/76/3860948>>. Citado na página 32.
- WARD, M.; GRINSTEIN, G.; KEIM, D. *Interactive Data Visualization: Foundations, Techniques, and Applications*. [S.l.: s.n.], 2010. ISBN 9780429108433. Citado na página 78.
- WU, Y. et al. *Detectron2*. 2019. Citado na página 79.
- XKCD. *xkcd: Machine Learning*. 2017. Disponível em: <<https://xkcd.com/1838/>>. Citado 2 vezes nas páginas 22 e 23.
- YANG, S. et al. Transpose: Keypoint localization via transformer. *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., p. 11782–11792, 12 2020. Disponível em: <<http://arxiv.org/abs/2012.14214>>. Citado na página 79.
- ZHOU, J. et al. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics 2021, Vol. 10, Page 593*, Multidisciplinary Digital Publishing Institute, v. 10, n. 5, p. 593, mar 2021. ISSN 20799292. Disponível em: <<https://www.mdpi.com/2079-9292/10/5/593/htmhttps://www.mdpi.com/2079-9292/10/5/593>>. Citado na página 72.
- ZHU, M. et al. Interpreting deep learning based cerebral palsy prediction with channel attention. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. ISBN 9781665403580. Citado na página 37.
- ŠIMIĆ, I.; SABOL, V.; VEAS, E. Xai methods for neural time series classification: A brief review. 8 2021. Disponível em: <<http://arxiv.org/abs/2108.08009>>. Citado 2 vezes nas páginas 25 e 37.

Apêndices

APÊNDICE A – Dados dos estudos incluídos

Neste Capítulo são apresentados os dados extraídos dos estudos incluídos na revisão, através da Tabela [10](#).

Tabela 10 – Estudos incluídos na revisão.

Título	Autor (Ano)	Nome da técnica ou método	Qual o objetivo?	Categoria de objetivo	Cita direta- mente uso de XAI?	Cita uso de alguma biblio- teca de XAI? Qual?	Técnica de IA desenvolvida com explicabi- lidade ou explicação posterior?	Qual a fonte de dados utilizada?	Faz uso de visualização da informação para obter ex- plicabilidade? Como?	Houve estudo de avaliação ou validação com usuários sobre o método de XAI?
A Transferable Deep Learning Prognosis Model for Predicting StrokePatients' Recovery in Different Rehabilitation Trainings	(LIN et al., 2022)	Não informado	Avaliar prognósticos antes e depois de tratamentos	Avaliação da evolução, prognóstico e desfecho clínico dos pacientes	Sim	Sim, SHAP	Explicação posterior	Dados coletados, clínicos e fisiológicos de 15 pacientes sobreviventes de AVC	Sim, através dos gráficos do SHAP	Não
An Interpretable Deep Learning Optimized Wearable Daily Detection System for Parkinson's Disease	(CHEN et al., 2023)	Não informado	Detectar e avaliar caminhada do paciente com doença de Parkinson	Avaliação das sessões de reabilitação, considerando dados de movimento	Sim	Sim, Grad-CAM	Explicação posterior	Dados coletados de 100 indivíduos	Sim, através dos gráficos do Grad-CAM	Não
Cross-validation of predictive models for functional recovery after post-stroke rehabilitation	(CAM-PAGNINI et al., 2022)	Não informado	Analisando a contribuição de cada fator prognóstico na alta após reabilitação intensiva pós-AVC	Avaliação da evolução, prognóstico e desfecho clínico dos pacientes	Sim	Sim, SHAP	Explicação posterior	Dados coletados de 278 pacientes pós-AVC	Sim, através dos gráficos do SHAP	Não

(continuação)

Título	Autor (Ano)	Nome da técnica ou método	Qual o objetivo?	Categoria de objetivo	Cita direta- mente uso de XAI?	Cita uso de alguma biblio- teca de XAI? Qual?	Técnica de IA desenvolvida com explicabi- lidade ou explicação posterior?	Qual a fonte de dados utilizada?	Faz uso de visualização da informação para obter ex- plicabilidade? Como?	Houve estudo de avaliação ou validação com usuários sobre o método de XAI?
EEG fractal dimensions predict high-level behavioral responses in minimally conscious patients	(LIUZZI et al., 2023)	Não informado	Analisar registros de eletroencefalografia para classificação de estado do paciente com consciência mínima	Avaliação da evolução, prognóstico e desfecho clínico dos pacientes	Sim	Sim, SHAP	Explicação posterior	Dados coletados, 68 pacientes com estado de consciência mínimo	Sim, através dos gráficos do SHAP	Não
eXplainable AI Allows Predicting Upper Limb Rehabilitation Outcomes in Sub-Acute Stroke Patients	(GAN-DOLFI et al., 2023)	Não informado	Prever diretamente os escores finais de recuperação funcional do paciente	Avaliação da evolução, prognóstico e desfecho clínico dos pacientes	Sim	Sim, PFI, RFI, SHAP e LIME	Explicação posterior	Dados coletados, pacientes internados na Unidade de Neuroreabilitação. Eram 192 indivíduos, após exclusão ficaram 95 registros	Sim, através de gráficos de barras, demonstrando a importância dos recursos globais	Não

(continuação)

Título	Autor (Ano)	Nome da técnica ou método	Qual o objetivo?	Categoria de objetivo	Cita direta- mente uso de XAI?	Cita uso de alguma biblio- teca de XAI? Qual?	Técnica de IA desenvolvida com explicabi- lidade ou explicação posterior?	Qual a fonte de dados utilizada?	Faz uso de visualização da informação para obter ex- plicabilidade? Como?	Houve estudo de avaliação ou validação com usuários sobre o método de XAI?
Factors influencing neuromuscular responses to gait training with arobotic ankle exoskeleton in cerebral palsy	(CON- NER et al., 2023)	Não informado	Identificar fatores que influenciam as respostas neu- romusculares a um exoesqueleto resistivo de tornozelo em indivíduos com paralisia cerebral	Avaliação das sessões de reabilitação, considerando dados de movimento	Sim	Sim, efeitos locais acumula- dos (ALE)	Explicação posterior	Dados coletados	Sim, através dos gráficos de efeitos locais acumulados (ALE)	Não
Implementation of eXplainable Artificial Intelligence Case Study on the Assessment of Movements to Support Neuromotor Rehabilitation	(CA- MARGO; DIAS; BREGA, 2023)	Não informado	Aplicar XAI em dados de classificação de movimentos, para apoio à reabilitação neuromotora	Avaliação das sessões de reabilitação, considerando dados de movimento	Sim	Não	Explicação posterior	Dados coletados	Sim, pretende usar um avatar para demonstração das coordenadas e dos movimentos	Não

(continuação)

Título	Autor (Ano)	Nome da técnica ou método	Qual o objetivo?	Categoria de objetivo	Cita direta- mente uso de XAI?	Cita uso de alguma biblio- teca de XAI? Qual?	Técnica de IA desenvolvida com explicabi- lidade ou explicação posterior?	Qual a fonte de dados utilizada?	Faz uso de visualização da informação para obter ex- plicabilidade? Como?	Houve estudo de avaliação ou validação com usuários sobre o método de XAI?
Predicting outcome of patients with prolonged disorders of consciousness using machine learning models based on medical complexity	(LIUZZI et al., 2022)	Não informado	Investigar o impacto das complicações médicas na predição do desfecho clínico	Avaliação da evolução, prognóstico e desfecho clínico dos pacientes	Sim	Sim, SHAP	Explicação posterior	Dados coletados de um grande grupo de pacientes inscritos em um projeto multicêntrico, observacional e longitudinal, 176 pacientes	Sim, através dos gráficos do SHAP	Não
Upper Limb Motion Tracking and Classification: A Smartphone Approach	(RODRIGUES et al., 2021)	Não informado	Capturar, transformar e classificar dados de movimentos com aprendizado de máquina	Avaliação das sessões de reabilitação, considerando dados de movimento	Não	Não	Não se aplica	Dados coletados, 232 instâncias utilizadas no treinamento	Não	Não

Fonte - Produzida pelo autor.

APÊNDICE B – Estudos excluídos

Os seguintes estudos foram excluídos durante a etapa de extração:

- A framework for neuromotor and neurofunctional rehabilitation using multi-agent systems (VALENTE et al., 2024) Motivo da exclusão: **Falta de afinidade do estudo com os temas desejados** - Foca no desenvolvimento, engenharia de software apoiando aplicações de reabilitação, sem uso de IA;
- Designing Modular Rehabilitation Objects for Interactive Therapy in the Home (KELLIHER et al., 2019) Motivo da exclusão: **Falta de afinidade do estudo com os temas desejados** - Cita aprendizado e visão computacional, mas não detalha, focando em outros temas;
- Engineering Rehabilitation: Blending Two Tool-supported Approaches to Close the Loop from Tasks-based Rehabilitation to Exercises and Back Again (AXEL et al., 2023) Motivo da exclusão: **Estudo não possui acesso liberado para a instituição** - Download não disponível para a Unesp na ACM;
- ExerciseCheck: a scalable platform for remote physical therapy deployed as a hybrid desktop and web application (PANDIT et al., 2019) Motivo da exclusão: **Falta de afinidade do estudo com os temas desejados** - Foca em reabilitação doméstica, a distância, mas sem uso de IA;
- ExerciseCheck: data analytics for a remote monitoring and evaluation platform for home-based physical therapy (SARAEI et al., 2019) Motivo da exclusão: **Falta de afinidade do estudo com os temas desejados** - Foca em reabilitação doméstica, a distância, mas sem uso de IA;
- Iterative Design of an Upper Limb Rehabilitation Game with Tangible Robots (OZGUR et al., 2018) Motivo da exclusão: **Falta de afinidade do estudo com os temas desejados** - Utiliza jogo físico, avaliação manual dos movimentos, sem uso de IA.
- Semi-automated home-based therapy for the upper extremity of stroke survivors (RIKAKIS et al., 2018) Motivo da exclusão: **Falta de afinidade do estudo com os temas desejados** - Foca em reabilitação doméstica, a distância, mas sem uso de IA;