LUCAS HENNA SALLABERRY

APRENDIZADO DE MÁQUINA NA AVALIAÇÃO DE DESEMPENHO DE APRENDIZES EM SIMULADORES DE REALIDADE VIRTUAL COM DISPOSITIVO HÁPTICO

São Paulo
2023

LUCAS HENNA SALLABERRY

APRENDIZADO DE MÁQUINA NA AVALIAÇÃO DE DESEMPENHO DE APRENDIZES EM SIMULADORES DE REALIDADE VIRTUAL COM DISPOSITIVO HÁPTICO

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências.

São Paulo
2023

LUCAS HENNA SALLABERRY

APRENDIZADO DE MÁQUINA NA AVALIAÇÃO DE DESEMPENHO DE APRENDIZES EM SIMULADORES DE REALIDADE VIRTUAL COM DISPOSITIVO HÁPTICO

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Ciências.

Área de Concentração:

Engenharia de Computação

Orientadora:

Fátima de Lourdes dos Santos Nunes Marques

Coorientador:

Romero Tori

São Paulo
2023

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 06 de fevereiro de 2023

Assinatura do autor: Lucas Sallaberry

Assinatura do orientador: Fátima Nunes

Catalogação-na-publicação

Sallaberry, Lucas
Aprendizado de máquina na avaliação de desempenho de aprendizes em simuladores de realidade virtual com dispositivo háptico / L. Sallaberry, R. Tori, F. Nunes -- versão corr. -- São Paulo, 2023.
103 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.AVALIAÇÃO DE DESEMPENHO 2.SIMULAÇÃO (APRENDIZAGEM) 3.APRENDIZADO COMPUTACIONAL 4.REALIDADE VIRTUAL 5.ANESTESIA ODONTOLÓGICA I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t. III.Tori, Romero IV.Nunes, Fátima

Dedico este trabalho a Diva Moraes Sallaberry *(in memoriam)*.

AGRADECIMENTOS

Este trabalho leva poucos nomes em sua capa, porém muitas pessoas participaram dele, ainda que indiretamente, para torná-lo possível durante os últimos anos.

Primeiramente, agradeço à minha esposa Bianca Sallaberry, pelo apoio nos momentos mais difíceis, pelo encorajamento, compreensão e carinho. Foi uma jornada cansativa, mas sua presença trouxe o conforto e direcionamento que precisei em muitos momentos. Esta conquista não é apenas minha.

Aos meus pais, sem os quais nada disso seria possível, pelo apoio e esforço durante todos esses anos, pelos ensinamentos e valores que carrego. Os caminhos que trilhei são reflexo das referências que tive.

À professora Fátima Nunes, pelo aprendizado e profissionalismo, mas acima de tudo, pelo constante acolhimento e dedicação, um exemplo de profissional e pessoa. Também agradeço ao professor Romero Tori, que me orientou desde o início da minha trajetória acadêmica. Com eles, aprendi o que é fazer pesquisa com o rigor científico necessário.

Aos colegas de laboratório, pelos momentos e risadas, certamente fizeram falta durante o período de pandemia da COVID-19. Agradeço especialmente à colega Elen Collaço, com quem trabalhei por um longo período no simulador VIDA Odonto e em seus experimentos, pelas longas conversas. Também agradeço a todos os professores que lutam diariamente pelo crescimento e divulgação da ciência no Brasil.

Por fim, agradeço às agências de financiamento que possibilitaram a realização desta pesquisa: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Processo 887.388306/2019-00, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Processos 315180/2018-8 e 309030/2019-6, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – Processo 2016.26290-3, Instituto Nacional de Ciência e Tecnologia em Medicina Assistida por Computação Científica (INCT-MACC) – Processo 2014/50889-7.

RESUMO

Simuladores de Realidade Virtual (RV) voltados para a saúde possibilitam o treinamento de estudantes antes do primeiro contato com o paciente, de maneira segura e sem gasto de material. Dados coletados durante a execução do treinamento nestes simuladores favorecem a implementação de sistemas de avaliação automática, que podem tanto fornecer retorno de desempenho para o aluno durante o treinamento quanto auxiliar o professor durante avaliações de estudantes. Resultados obtidos em uma revisão sistemática da literatura apontaram que estudos que utilizam aprendizado de máquina para realizar a avaliação de aprendizes em simuladores de RV com retorno háptico ainda são escassos. Este trabalho comparou as técnicas de aprendizado de máquina *Naive Bayes* (NB), *Random Forest* (RF), *Support Vector Machine* (SVM), *Multi Layer Perceptron* (MLP) e *Extreme Gradient Boosting* (XGB), combinadas com técnicas de seleção e fusão de características para definir um método de avaliação automática a fim de mensurar o desempenho de estudantes a partir do uso de simuladores de realidade virtual para treinamento médico. A trajetória da seringa em um simulador odontológico usando realidade virtual foi coletada e segmentada, possibilitando a extração de 98 características relacionadas ao desempenho dos participantes durante o procedimento. Foram realizadas classificações binárias com dois tipos diferentes de rotulação: com base no nível de experiência que os participantes possuíam no procedimento simulado e com base em uma avaliação da trajetória realizada por uma especialista. Foi encontrado que, em ambos os casos, o classificador SVM apresentou o melhor resultado final, com acurácia de 0,77, especificidade de 0,60 e sensibilidade de 0,94 para a rotulação realizada com base na avaliação da especialista. Os resultados obtidos neste trabalho se mostraram promissores e o modelo definido pode ser aplicado a simuladores hápticos usando RV que coletam dados de trajetórias.

Palavras-Chave – Avaliação de aprendiz, simulador odontológico, retorno háptico, aprendizado de máquina, realidade virtual.

ABSTRACT

Virtual Reality (VR) simulators for health procedures allow students to train in a safe manner and without loss of material before they have the first contact with a real patient. Data collected during training sessions using these simulators favor the implementation of automatic assessment systems, which could both provide a performance feedback to the trainee, without the need of an instructor present, and assist instructors during the evaluation of students. Results obtained through a systematic review of the literature indicated that studies using machine learning to assess apprentices in VR simulators with haptic feedback are still scarce. This work compared the machine learning techniques Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Multi Layer Perceptron (MLP) and Extreme Gradient Boosting (XGB), combined with techniques for feature fusion and selection to define an automatic assessment method to measure the student's performance while using VR simulators for medical training. The syringe trajectory in a VR dental simulator was collected and segmented, allowing the extraction of 98 features related to the participants' performance during the procedure. Binary classifications were performed with two different types of labeling: based on the level of experience the participants had in the simulated procedure and based on an expert's evaluation of the trajectory. It was found that in both cases, the SVM classifier presented the best final result, with an accuracy of 0.77, specificity of 0.60, and sensitivity of 0.94 for the labeling based on the expert's evaluation. The results obtained in this work are promising and the model can be applied to VR haptic simulators that collect trajectory data.

Keywords – Apprentice assessment, dental simulator, haptic feedback, machine learning, virtual reality.

LISTA DE FIGURAS

LISTA DE TABELAS

SIGLAS

2D bidimensional. 60, 63

3D tridimensional. 20, 52

AE Avaliação pelo Especialista. 62, 72, 73, 74, 76, 77, 78, 79, 80, 81, 82, 85, 89, 92, 93

ANN *Artificial Neural Network*. 16, 28, 86, 87

BNAI Bloqueio do Nervo Alveolar Inferior. 59, 61, 88, 91

CPs Componentes Principais. 32, 80

DT Diferentes tipos de treinamento. 45, 46

DTW *Distance Time Warping*. 70, 71, 85, 86

DV Diferentes versões do procedimento. 45

EN Especialistas *versus* Novatos. 45, 61, 72, 73, 74, 76, 77, 78, 79, 80, 81, 82, 85, 89, 92, 93

FN Falso Negativo. 29

FP Falso Positivo. 29

GA *Genetic Algorithm*. 33, 64, 73, 76, 77, 79, 80, 85

HCRF *Hidden-state Conditional Random Fields*. 86, 88

HMD *Head Mounted Display*. 20, 21, 35, 38, 52, 59, 60, 63

HMM *Hidden Markov Model*. 16, 47, 86, 88, 91

Hz Hertz. 53, 60, 62

LDA *Linear Discriminant Analysis*. 87

ML *Machine Learning*. 16, 24, 25, 30, 49, 62, 64, 65, 88, 93

MLP *Multilayer perceptron*. 28, 53, 64, 73, 76, 77, 79, 80, 87, 91, 92

MS Aperfeiçoamento ao longo de múltiplas sessões. 45, 46

NB *Naive Bayes*. 28, 53, 64, 73, 76, 77, 79, 80, 82, 85, 86, 87, 91

PCA *Principal Component Analysis*. 32, 53, 64, 73, 76, 77, 79, 80

RA Realidade Aumentada. 19, 20, 40, 41, 42, 45, 47, 48, 49, 54, 66

ResNet *Residual Neural Network*. 87, 88

RF *Random Forests*. 27, 28, 53, 64, 73, 76, 77, 78, 79, 80, 91

RFF Com e sem Retorno de Força. 45

RV Realidade Virtual. 15, 16, 18, 19, 20, 21, 22, 33, 34, 35, 38, 39, 40, 41, 42, 45, 47, 48, 49, 50, 51, 54, 55, 56, 58, 66, 75, 76, 86, 87, 89, 92

SMOTE *Synthetic Minority Over-sampling Technique*. 30

SVM *Support Vector Machine*. 16, 26, 53, 64, 73, 76, 77, 79, 80, 81, 82, 85, 86, 87, 88, 91, 92

TN treinado *versus* não treinado. 45, 46

VN Verdadeiro Negativo. 29

VP Verdadeiro Positivo. 29

XGB *Extreme Gradient Boosting*. 28, 53, 64, 73, 76, 77, 78, 79, 80, 81, 91

SUMÁRIO

1 INTRODUÇÃO

A Realidade Virtual (RV) vem adquirindo grande popularidade como forma de diversão, mas suas aplicações vão muito além do ramo do entretenimento (JERALD, 2016). Destacam-se os sistemas voltados a ensino e treinamento, principalmente na área da saúde, uma vez que possibilitam expor aprendizes a situações realistas, sem os riscos de cometerem erros em pacientes reais (BURDEA; COIFFET, 2003).

Simuladores de RV são muitas vezes apresentados na forma de "*serious games*", definidos como jogos voltados para um objetivo primário diferente do entretenimento, geralmente visando ao ensino ou ao treinamento (BREUER; BENTE, 2010). Esses sistemas têm sido utilizados para o treinamento de neurocirurgias, laparoscopias, procedimentos odontológicos, dentre várias outras aplicações(CHAN et al., 2013; PRASAD et al., 2018; RHIENMORA et al., 2011).

O aprendizado de habilidades sensório-motoras para procedimentos médicos requer muitas horas dedicadas ao treinamento (CHIKWE; de Souza; PEPPER, 2004). Esta atividade deve ser supervisionada por um instrutor capaz de indicar os erros realizados pelo aluno, mas os profissionais capacitados para tal dispõem de tempo limitado. Simuladores de RV com retorno de força possibilitam a coleta de diversos dados objetivos, geralmente obtidos por meio de dispositivo háptico, relacionados às ações do aprendiz, enquanto o estudante treina um procedimento. Estes dados podem ser utilizados para a implementação de sistemas de avaliação de habilidades, capazes de fornecer retorno de desempenho do treinamento para o aluno, automaticamente. Isso possibilita que o aprendiz identifique os pontos da execução do procedimento que devem ser melhorados, possibilitando que o aluno possa tecer uma análise crítica de seu desempenho, conferindo maior autonomia ao seu aprendizado e qualificando a ação docente durante todo o treinamento.

Sistemas de treinamento usando RV também podem beneficiar o professor, auxiliando-o na avaliação de estudantes ao proporcionar dados do desempenho dos alunos. Tais informações podem ser difíceis de serem examinadas com exatidão por

observação visual, tal como a velocidade, a força e a precisão dos movimentos. Técnicas de avaliação automática também podem proporcionar critérios de avaliação mais padronizados e objetivos como ferramenta para os professores.

Existem diversos simuladores de RV em fase de pesquisa, voltados para treinamento de procedimentos relacionados à saúde, tal como o *haptic technology-enhanced learning* (hapTEL) (RIA et al., 2018), simulador de RV para treinamento do preparo cavitário e remoção de cáries, e o *The Western myringotomy simulator* (HUANG et al., 2018), voltado para o procedimento de miringotomia. Esses simuladores variam quanto aos procedimentos simulados e tecnologias empregadas, entre outros fatores. Alguns deles fornecem métricas calculadas a partir do desempenho do usuário (TAI et al., 2017; Wing-Yin Chan et al., 2012; TAI et al., 2018), sendo que alguns ainda utilizam essas métricas para calcular uma nota final de desempenho (ESEN et al., 2007; RIA et al., 2018; MIRGHANI et al., 2018).

Um levantamento da literatura referente aos métodos de avaliação de desempenho em simuladores de RV, apresentado no Capítulo 3 deste trabalho, indicou que, entre as técnicas utilizadas para avaliação de participantes, algoritmos de aprendizado de máquina, em inglês *Machine Learning* (ML), foram pouco explorados. Tais algoritmos podem proporcionar uma abordagem precisa, ao mesmo tempo em que oferecem flexibilidade, favorecendo seu uso em diferentes procedimentos, quando comparados com outros métodos empregados na literatura. Outras revisões, com finalidade similar (DIAS; GUPTA; YULE, 2019; Winkler-Schwartz et al., 2019), mostraram que técnicas de ML, como Modelos Ocultos de Markov, em inglês *Hidden Markov Model* (HMM), Máquinas de Vetores de Suporte, em inglês *Support Vector Machine* (SVM), e Redes Neurais Artificiais, em inglês *Artificial Neural Network* (ANN), já mostraram resultados para avaliação de desempenho em procedimentos médicos.

Em uma parceria do Laboratório de Tecnologias Interativas (Interlab - EPUSP), Laboratório de Aplicações de Informática em Saúde (LApIS - EACH/USP) e Laboratório de Simulação e Treinamento (LaSiT - FOB/USP), foram desenvolvidos diversos simuladores para aquisição de habilidade sensório-motora no contexto da saúde. Entre eles está o VIDA Odonto (TORI et al., 2018), um simulador de procedimentos odontológicos de RV e com retorno de força. O foco atual do VIDA Odonto é o treinamento da aplicação de anestesias de maneira realista e confiável, especificamente para o bloqueio do nervo alveolar inferior. Esse simulador, no entanto, ainda não disponibiliza um método automático de avaliação dos aprendizes, sendo que este constitui o foco do presente trabalho de Mestrado.

1.1 Objetivos

O objetivo deste trabalho é definir, implementar e validar um modelo de avaliação automática de desempenho em simuladores de RV, com retorno háptico, utilizando conceitos de ML. O modelo resultante deve ser capaz de utilizar os dados coletados pelo simulador para classificar os aprendizes de acordo com a avaliação de um especialista. Como estudo de caso, o modelo foi definido com base nos dados coletados a partir do uso do simulador VIDA Odonto.

Para alcançar o objetivo geral, foram estabelecidos os seguintes objetivos específicos:

- analisar e definir métricas de desempenho que possam ser extraídas de procedimentos;
- comparar diferentes algoritmos de ML com as métricas de desempenho obtidas;
- definir e validar um modelo de avaliação automática de desempenho;

1.2 Organização do documento

Este documento está organizado da seguinte forma: o Capítulo 2 apresenta os principais conceitos utilizados para a realização do projeto; o Capítulo 3 exibe os resultados obtidos durante uma revisão sistemática da literatura, com o objetivo de identificar metodologias e técnicas empregadas para avaliação automática de desempenho; o Capítulo 4 define os materiais e métodos deste trabalho; o Capítulo 5 apresenta o modelo definido; o Capítulo 6 expõe os resultados obtidos, juntamente com a discussão e o Capítulo 7 apresenta as considerações finais.

2 CONCEITOS FUNDAMENTAIS

Considerando o caráter interdisciplinar da presente pesquisa, este capítulo apresenta os conceitos fundamentais envolvidos no estudo. A Seção 2.1 expõe os conceitos básicos sobre realidade virtual e aumentada, além dos dispositivos que podem ser empregados neste tipo de sistema. A Seção 2.2 apresenta conceitos relacionados às habilidades sensório-motoras e demonstra como são adquiridas. A Seção 2.3 mostra conceitos fundamentais e algoritmos clássicos de aprendizado de máquina. A Seção 2.4 caracteriza simuladores de RV e demonstra como podem ser utilizados para avaliação. A Seção 2.5 apresenta o simulador VIDA Odonto, suas características e tecnologias, e a Seção 2.6 apresenta as considerações finais deste capítulo.

2.1 Realidade virtual e aumentada

Segundo Jerald (2016), RV consiste em ambientes gerados por computador com os quais se pode interagir como se fossem reais. Jerald (2016) ainda aponta que um sistema ideal possibilitaria que o usuário fosse capaz de navegá-lo e interagir com objetos, tal como faria em um ambiente real. Ambientes de RV podem variar em diversos aspectos, tal como no nível de interação e imersão. De acordo com Burdea e Coiffet (2003), um aspecto importante da RV é o fato de que a simulação e a interação devem ocorrer em tempo real, ou seja, o sistema deve ser capaz de reconhecer as entradas do usuário e realizar as modificações no ambiente virtual instantaneamente. Dessa forma, um dos maiores objetivos dos sistemas de RV que buscam realismo é tornar o intermédio entre as ações do usuário e a resposta do sistema o mais imperceptível possível, fazendo o usuário se sentir no controle das interações com o ambiente. Burdea e Coiffet (2003) também definem a imaginação como essencial para a tecnologia de RV, pois ela é necessária tanto para o usuário, na compreensão do mundo virtual no qual ele se encontra, quanto para o desenvolvedor, na criação do ambiente virtual e no desenvolvimento do sistema.

Milgram et al. (1995) definiram o contínuo de virtualidade (Figura 1), o qual nos mostra que ambientes podem variar em um espectro de possibilidades, entre o real e o virtual. Neste contexto, **ambientes reais** são aqueles nos quais todas as entidades presentes são reais, ou seja, não existe nenhum objeto virtual inserido nele. Assim, o maior exemplo de ambiente real é o mundo real no qual vivemos. De forma análoga, **ambientes virtuais** são compostos exclusivamente por entidades virtuais. A combinação do real com o virtual resulta em ambientes de **realidade mista**, os quais podem ser predominantemente reais, chamados ambientes de **Realidade Aumentada (RA)**, ou predominantemente virtuais, os ambientes de **virtualidade aumentada**. Ambientes de RA sobrepõem o mundo real com objetos virtuais, idealmente de forma fluida, sem que que o usuário seja capaz de diferenciar o que é real e o que é virtual. Já ambientes de virtualidade aumentada são capazes de capturar objetos do mundo real e "transportá-los" para o mundo virtual.

Figura 1 – Contínuo de virtualidade. Fonte:(MILGRAM et al., 1995)(p. 283) adaptado pelo autor, tradução nossa.

Uma das características mais importantes de sistemas de RV e RA é a imersão. Diferentes definições para imersão em ambientes virtuais são apresentadas na literatura. Neste trabalho, consideramos a definição elaborada por Jerald (2016), na qual imersão é um fator objetivo referente ao nível de estímulo que o sistema projeta no usuário. Diversos aspectos contribuem para o nível de imersão atingido em uma aplicação, tal como os sentidos envolvidos (visão, audição, tato, etc), a qualidade técnica (resolução da imagem, qualidade do áudio) e o nível de interação com o ambiente. Desta forma, imersão é um fator relativo à tecnologia usada por um sistema, porém não reflete, necessariamente, a experiência de um usuário (JERALD, 2016).

Presença, por outro lado, é um aspecto referente à experiência que o usuário tem no ambiente. Presença é a sensação que um indivíduo tem de estar em um local, sem que necessariamente esteja fisicamente nele (JERALD, 2016). Quando o usuário se sente presente em um ambiente, ele não se atenta mais à tecnologia que realiza o intermédio entre ele e o ambiente virtual, mas passa a interpretar suas ações como

interações diretas com o ambiente. A partir do momento no qual o usuário se sente presente em um ambiente, o foco do sistema é de que ele permaneça nesse estado, porém sinais externos ao ambiente no qual a pessoa se encontra, como a fala de pessoas que não se encontram nessa realidade, podem causar uma quebra da ilusão (JERALD, 2016). O grau de imersão de um sistema, portanto, pode limitar o nível de presença sentido. Desta forma, diversos tipos de equipamentos foram desenvolvidos com o objetivo de melhorar a imersão de sistemas de RV e RA. A seguir, serão apresentados os *Head Mounted Display* e dispositivos hápticos:

- ***Head Mounted Display* (HMD):** HMDs (Figura 2a) são equipamentos acoplados à cabeça e que reproduzem o ambiente virtual por meio de uma tela posicionada a alguns centímetros dos olhos do usuário. Com a utilização de lentes localizadas no interior do HMD, a imagem da tela é ampliada de forma a tomar todo (ou quase todo) o campo de visão. Atualmente, modelos comerciais desses equipamentos possuem imagem estereoscópica e rastreio de movimento, de forma que são capazes de identificar a orientação (direção e sentido) da cabeça do usuário. O rastreio de movimento é de grande importância para que o sistema consiga alterar adequadamente a imagem mostrada, de forma que ela concorde com a orientação da cabeça do usuário no ambiente real, impedindo a quebra na sensação de presença (BURDEA; COIFFET, 2003). Alguns modelos também conseguem detectar a posição do usuário dentro de uma área pré-definida, de forma que ele possa se movimentar pelo ambiente real e ter sua locomoção reproduzida no ambiente virtual, proporcionando uma navegação natural, ainda que em um espaço limitado.

  HMDs modernos também proporcionam saídas de áudio acopladas à cabeça do usuário com som tridimensional (3D), proporcionando uma maior imersão em relação a alto-falantes fixos no ambiente real, uma vez que são capazes de reproduzir sons originados de qualquer local do ambiente virtual. Equipamentos com sons 3D conseguem fornecer estímulos capazes de convencer o usuário de que a fonte do som está situada em alguma posição ao seu redor (BURDEA; COIFFET, 2003). Por exemplo, caso o usuário se encontrasse em um ambiente virtual aberto (tal como um parque) e ouvisse um cachorro latir, o usuário seria capaz de identificar a direção do som e até mesmo estimar a sua distância dependendo da intensidade com que se ouve o latido. Caso o usuário decida virar a cabeça, o som emitido pela saída de áudio também é alterado, de forma que o usuário tenha a percepção de que a fonte emissora, no caso o cachorro,

permaneceu fixa no espaço virtual.

HMDs voltados para RV são capazes de isolar, em grande parte, a visão que o usuário tem do ambiente real. Tal habilidade, somada à liberdade proporcionada ao usuário para movimentar sua cabeça (poder olhar para qualquer local do ambiente virtual, além de poder se movimentar, ainda que em um espaço físico limitado), tornam os HMDs ferramentas com grande poder de imersão.

Figura 2 - a) HMD Oculus Rift CV1 Fonte: Página do Oculus Rift CV1 na Wikipedia[1]. b) Dispositivo háptico Touch Fonte: Página do *Touch* no site da 3D *Systems*[2].

(a) (b)

- **Dispositivo háptico:** dispositivos hápticos são equipamentos capazes de simular interações de força com objetos do ambiente virtual. Interações hápticas são geralmente divididas em duas categorias: retorno de força e retorno tátil.

  O retorno de força corresponde a interações envolvendo atividade muscular e movimento de membros do corpo (JERALD, 2016). Por exemplo, ao tentar empurrar uma caixa pesada, sentimos a resistência proporcionada pelo atrito do objeto com o chão, o qual se torna proporcional ao esforço muscular que devemos realizar. Se, então, tentarmos abrir essa caixa com uma faca, sentiremos a resistência do material que procura impedir o rompimento da superfície, o qual varia de acordo com o tipo de material.

  Já o retorno tátil é vinculado às sensações proporcionadas ao tocar uma superfície. Tais sensações podem ser relacionadas à temperatura e rugosidade dessa superfície (BURDEA; COIFFET, 2003).

  Os dispositivos hápticos também podem ser categorizados entre ativos e passivos. O retorno de força nos dispositivos ativos, tal como o da Figura 2b, é

[1]Disponível em: <https://en.wikipedia.org/wiki/Oculus_Rift_CV1>. Acesso em: 22 jan. 2021.
[2]Disponível em: <https://br.3dsystems.com/haptics-devices/touch>. Acesso em: 22 jan. 2021.

modelado e controlado dinamicamente por um computador, enquanto dispositivos passivos consistem de objetos reais similares aos modelos virtuais que se pretende fornecer retorno háptico, e que são posicionados no mundo real em um local equivalente ao que se encontraria no mundo virtual (JERALD, 2016). Por exemplo, para fornecer a sensação háptica durante o treinamento de um procedimento, o simulador poderia utilizar um boneco real compatível com o paciente virtual. Dessa forma, quando o aprendiz interagisse com o paciente virtual, seja com as mãos ou com um instrumento, ele sentiria o retorno háptico proporcionado pelo boneco virtual, porém sentiria que ele era fruto das interações com o paciente virtual. Enquanto dispositivos com retorno háptico ativo são capazes de criar interações hápticas a partir de modelos computacionais, sem a necessidade de criar novos objetos reais equivalentes, modelar interações capazes de rivalizar com a sensação natural de tocar um objeto real é uma tarefa bastante desafiadora.

2.2 Aquisição de habilidades sensório-motoras

Durante a revisão da literatura conduzida neste trabalho (Capítulo 3), foi observado que o principal objetivo de simuladores médicos em RV é o treinamento de aprendizes para aquisição das habilidades sensório-motoras necessárias para realizar o procedimento treinado.

Habilidades motoras são aquelas empregadas para realizar tarefas que requerem a utilização voluntária dos membros e articulações de um indivíduo, a fim de alcançar um objetivo (MAGILL; ANDERSON, 2017). Por exemplo, a simples ação de caminhar exige habilidade motora, uma vez que são utilizadas principalmente as pernas, embora todo o conjunto do corpo possa estar envolvido, com o objetivo de deslocar o indivíduo de um local a outro, de forma voluntária.

Algumas habilidades motoras ainda necessitam de outros estímulos sensoriais para serem realizadas. O ato de pegar um objeto sobre uma mesa, por exemplo, exige conhecimento da posição do objeto e das mãos no espaço e esse conhecimento é geralmente obtido pela visão. Sensações táteis e de propriocepção (percepção dos movimentos e localização dos membros, corpo e cabeça no espaço) possibilitam que o indivíduo seja capaz de segurar o objeto, levantá-lo e colocá-lo em outro local (MAGILL; ANDERSON, 2017).

Dessa forma, habilidades sensório-motoras podem depender fortemente de estímulos sensoriais para que sejam executadas corretamente, alcançando o objetivo desejado (WOLPERT; DIEDRICHSEN; FLANAGAN, 2011). Em tarefas que exigem grande sincronia entre movimentos motores e estímulos sensoriais, ocorre o acoplamento percepção-ação, no qual os movimentos do corpo e a percepção sensorial se comportam de maneira coordenada (MAGILL; ANDERSON, 2017). Por exemplo, um jogador de futebol, ao realizar uma cabeceada, deve examinar a trajetória e velocidade da bola, quando esta vem em sua direção. Em seguida, deve coordenar o salto e o movimento do cabeceio, de forma a interceptar a trajetória da bola. Tal ação mostra grande dependência, tanto da visão para prever o trajeto da bola, quanto de habilidades motoras para possibilitar o contato com ela durante o salto.

O aprendizado de habilidades motoras consiste, geralmente, no treinamento por repetição, que, por sua vez, aumenta a capacidade de um indivíduo ter um desempenho bom, e relativamente permanente, em uma determinada tarefa (SCHMIDT; LEE, 2014). Durante o treinamento, é necessário que o aprendiz possa experimentar, alterando a sua maneira de executar o movimento a ser aprendido, a fim de encontrar a forma que seja a mais efetiva (SCHMIDT; LEE, 2014). Treinamentos que exigem um desempenho ótimo a cada tentativa podem impedir que se faça essa experimentação, uma vez que ela pode acarretar uma queda momentânea de desempenho, apesar de apresentarem melhores resultados a longo prazo(SCHMIDT; LEE, 2014). Esse fato pode ser especialmente problemático em treinamentos de tarefas com alto risco, tal como pilotagem de aviões, condução de veículos e procedimentos cirúrgicos, em que um desempenho ruim pode resultar em consequências severas (SCHMIDT; LEE, 2014). Para tais atividades, o uso de simuladores, por exemplo, pode ser de grande ajuda, na medida que criam ambientes de treinamento livres de tais riscos.

Além de aumentar a capacidade do indivíduo ter um bom desempenho em uma tarefa, o treinamento de habilidades motoras também reduz a atenção necessária para a realização da atividade e torna o treinando mais eficaz na identificação dos erros que foram cometidos (SCHMIDT; LEE, 2014). Durante o aprendizado, porém, é de importância crucial que seja fornecido *feedback* sobre o desempenho na atividade. O *feedback* durante o treinamento pode pertencer a uma de duas categorias, sendo elas *feedback* intrínseco ou *feedback* aumentado (SCHMIDT; LEE, 2014). O *feedback* intrínseco é considerado como as consequências naturais da tarefa sendo realizada, por exemplo, ver e ouvir uma bola de basquete batendo no aro ao realizar um arremesso errado. *Feedbacks* aumentados, por sua vez, são aqueles fornecidos artificialmente

para o aprendiz, tal como um instrutor avaliando a qualidade de salto ornamental. Em treinamentos nos quais não é possível identificar se a tarefa foi realizada ou não com sucesso por *feedback* intrínseco, o fornecimento de *feedback* aumentado se torna crucial para que ocorra o aprendizado corretamente (SCHMIDT; LEE, 2014).

Não é possível mensurar o aprendizado de habilidades motoras diretamente, porém ele pode ser inferido por meio de métricas de desempenho relacionadas à tarefa. Um método frequentemente utilizado para tal são as curvas de desempenho (muitas vezes chamadas curvas de aprendizado). Tal técnica consiste na construção de uma curva, com base no valor obtido para uma determinada métrica, ao longo de múltiplas sessões (MAGILL; ANDERSON, 2017). Dessa forma, espera-se que, com o aumento do número de repetições em uma tarefa, o desempenho seja incrementado. Por exemplo, um piloto de corrida, ao treinar um percurso, pode realizar múltiplas voltas neste percurso. A cada volta, o tempo transcorrido é anotado e colocado como um novo ponto na curva de desempenho. Conforme o piloto realiza novas voltas, espera-se que o tempo transcorrido por volta diminua, porém é possível que, em algumas das voltas, o tempo aumente por conta de alterações que o piloto possa fazer na forma de navegar pela pista, a fim de buscar modos mais eficientes de executar a volta. Eventualmente, espera-se que seja alcançado um *plateau*, no qual o tempo pouco se altere a cada volta. Um bom desempenho obtido em uma única tentativa não significa, necessariamente, que tenha ocorrido aprendizado, porém a consistência de bom desempenho ao longo de múltiplas sessões pode ser um indicativo do aprendizado.

2.3 Aprendizado de máquina

Segundo Murphy (2012, p. 1, tradução nossa), ML é “um conjunto de métodos capazes de detectar padrões em dados automaticamente e então usar estes padrões descobertos para predizer dados futuros ou realizar outros tipos de decisão diante da incerteza"(MURPHY, 2012).

De maneira similar, Bishop (2006, p. 1, tradução nossa) indica que o interesse da área de reconhecimento de padrões é “com a descoberta automática de regularidades nos dados através de algoritmos computacionais para então utilizar essas regularidades para tomada de ações, tal como a classificação de dados em categorias"(BISHOP, 2006). Dessa forma, o que difere algoritmos de ML de outros algoritmos é a capacidade de aprender, ou seja, detectar padrões em dados automaticamente e utilizar

esses padrões para futuras tomadas de decisões.

Mitchell (1997, p. 2, tradussão nossa) define que “um programa computacional aprende por uma experiência E em relação a uma tarefa T e uma métrica de desempenho P, se o seu desempenho em T, medido por P, melhora com experiência E"(MITCHELL, 1997). Assim, podemos considerar que um algoritmo aprendeu com dados quando, a partir de um conjunto de dados de treinamento, o algoritmo consegue melhorar o seu desempenho em uma determinada tarefa.

Devido à capacidade de identificar padrões em dados automaticamente , algoritmos de ML podem ser particularmente efetivos para solucionar problemas que exigem muitas regras e ajustes, ou que não apresentam um algoritmo de solução conhecido (GÉRON, 2019). Um exemplo comum de problema é a leitura de textos manuscritos. Apesar de as formas de letras e números serem conhecidas, um método à base de regras resultaria em uma extensa quantidade de regras, que deveriam reconhecer as mais variadas caligrafias, o que geralmente resultaria em uma solução ruim (BISHOP, 2006).

Muitos métodos de ML também são interpretáveis, ou seja, após o algoritmo realizar uma predição, é possível compreender como essa tarefa foi realizada (GÉRON, 2019). A interpretabilidade é especialmente importante em casos nos quais existe interesse em saber como o problema foi resolvido. Por exemplo, um algoritmo que utiliza dados de pacientes que deram entrada em um hospital poderia ser utilizado para identificar rapidamente a chance de futuros pacientes apresentarem determinada doença. Seriam coletadas informações, tais como o histórico de saúde, os sintomas iniciais e o diagnóstico final para a doença, e o algoritmo indicaria a probabilidade de um novo paciente vir a ter a mesma doença antes mesmo de realizar exames, o que poderia resultar em uma resposta médica mais rápida e específica. Neste exemplo, cada paciente representaria uma instância do conjunto de dados, os atributos relacionados ao paciente (como idade, peso, sintomas) são chamados características do conjunto de dados e o diagnóstico final é a solução. Dependendo do nível de interpretabilidade do algoritmo, seria possível extrair dele quais foram as particularidades dos pacientes que parecem estar relacionadas à doença, descobrindo, possivelmente, novos sintomas ou grupos de risco a serem considerados.

Uma das classificações mais utilizadas de algoritmos de aprendizado os divide em duas categorias: supervisionado e não supervisionado. As técnicas de aprendizado supervisionado utilizam dados de treinamento no qual cada exemplo é identificado

com a sua respectiva solução (GÉRON, 2019). Em problemas de classificação supervisionada, o algoritmo treinado deve ser capaz de identificar a qual classe uma nova instância pertence, podendo existir duas ou mais classes pré-determinadas. Neste caso, o conjunto de treinamento fornecido ao algoritmo contém a identificação da classe à qual cada uma das instâncias de treinamento pertence. Por exemplo, um algoritmo de classificação supervisionada aplicado ao problema descrito no parágrafo anterior utilizaria, como conjunto de treinamento, os dados de pacientes anteriormente diagnosticados como positivos ou negativos para a doença a ser estudada. Após treinado, ele poderia receber as informações de um novo paciente, ainda não diagnosticado, e ser capaz de classificar este novo paciente como positivo ou negativo para uma determinada doença, utilizando como base todos os casos de pacientes anteriores fornecidos.

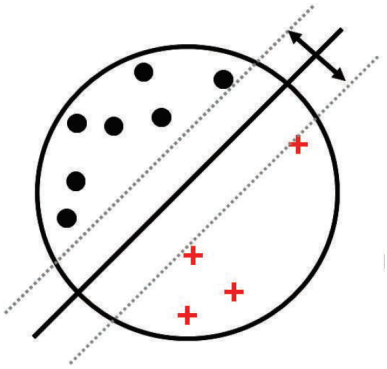
Para técnicas de classificação supervisionada, é comum utilizar parte do conjunto de dados para realizar o treinamento do classificador e o restante dos dados para testá-lo (GÉRON, 2019). Após treinado, o algoritmo classifica cada uma das instâncias do conjunto de teste e a solução obtida é, então, comparada com a solução real para verificar se a classificação foi realizada corretamente. Por meio deste processo, é possível estimar como seria o desempenho desse classificador ao utilizar novos dados coletados, uma vez que ele não tinha conhecimento prévio dos dados de teste (GÉRON, 2019).

2.3.1 Algoritmos de classificação

Para classificação binária com algoritmos de aprendizado supervisionado, alguns dos algoritmos mais conhecidos serão descritos a seguir.

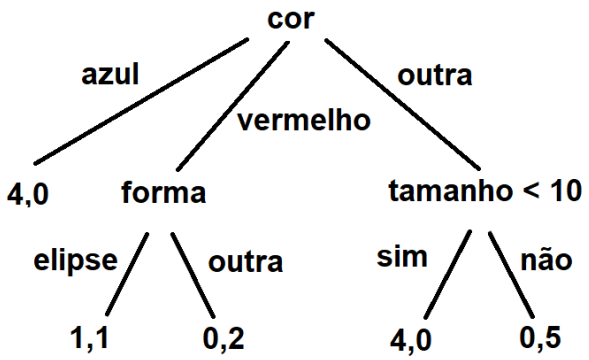
- **Máquinas de Vetores de Suporte:** a SVM é capaz de realizar tarefas de classificação linear ou não-linear e regressão. O objetivo da SVM para classificação é encontrar uma fronteira de decisão que divide o espaço amostral entre as instâncias de cada classe. Para tal, o algoritmo busca um hiperplano que separe as classes e no qual a menor distância entre ele e uma instância qualquer seja a maior possível (Figura 3). Em casos em que não é possível ou é muito difícil delimitar uma fronteira que separe as classes perfeitamente, é possível utilizar uma SVM de margem *soft*, na qual é permitido que algumas das instâncias de uma classe se apresentem na região delimitada para a outra classe (GÉRON, 2019).

Figura 3 – Ilustração de uma SVM. O algoritmo encontra a fronteira de decisão que divide as duas classes e que apresenta a maior distância mínima para as instâncias de cada uma delas. Fonte: Murphy (2012, p. 500) (MURPHY, 2012).



- ***Árvores de decisão:*** são capazes de realizar tarefas tanto de classificação como de regressão (GÉRON, 2019). A Figura 4 mostra um exemplo de uma árvore de decisão. Neste exemplo, a partir de um nó inicial, é verificada a cor da instância que será classificada e a aresta a ser seguida é definida a partir deste parâmetro. De forma semelhante, o nó seguinte apresenta uma nova condição a ser analisada para determinar a aresta a ser seguida. Este processo se repete até que se alcance um nó terminal, que corresponde à predição realizada. Este tipo de algoritmo é capaz de trabalhar com dados tanto contínuos quanto discretos, é de fácil interpretação e é capaz de selecionar as variáveis que serão utilizadas automaticamente. No entanto ele apresenta pior acurácia na predição quando comparado com outros modelos(MURPHY, 2012).

Figura 4 – Exemplo de árvore de decisão. Fonte: Murphy (2012, p. 545) (MURPHY, 2012) adaptado pelo autor, tradução nossa.



- ***Random Forests*** **(RF):** são comitês de árvores de decisão no qual cada árvore é gerada a partir de um conjunto aleatório diferente de instâncias de treinamento. Cada árvore é também gerada a partir de um subconjunto aleatório de carac-

terísticas do conjunto de dados, resultando em uma baixa correlação entre as árvores, tornando o classificador mais robusto. Para problemas de classificação, a RF atribui a uma instância a classe mais votada pelas árvores de decisão. (MURPHY, 2012)

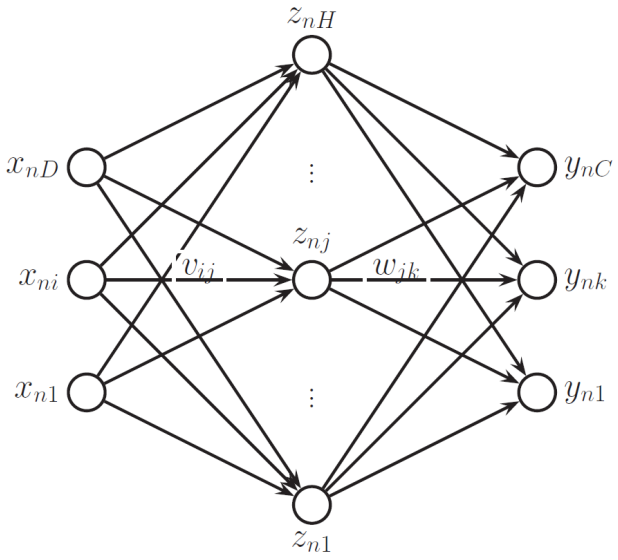
- ***Naive Bayes* (NB):** é um tipo de classificador probabilístico, que assume que as características do conjunto de dados são condicionalmente independentes umas das outras. A denominação "*naive*" (do inglês, ingênuo) deve-se ao fato de não se esperar que as características sejam realmente independentes, mas esta abordagem nos permite definir a densidade condicional de classe como um produto de densidades unidimensionais (MURPHY, 2012).

- **Redes Neurais Artificiais:** são algoritmos capazes de modelar funções matemáticas para tarefas de regressão ou de classificação. As ANNs (Figura 5) têm semelhança com as redes neurais naturais pela analogia criada entre os nós das ANNs com neurônios e as funções de ativação com sinapses. A *Multilayer perceptron* (MLP) é um tipo de ANN em que os neurônios de cada camada "ativam" os neurônios da camada seguinte, e que por sua vez ativarão os neurônios da camada subsequente até se obter um ou mais valores de saída na camada final. ANNs multicamadas, como a MLP, são capazes de modelar funções matemáticas complexas, porém o algoritmo costuma apresentar pouca interpretabilidade (GÉRON, 2019).

- ***Extreme Gradient Boosting* (XGB):** O algoritmo XGB para classificação tem como objetivo adicionar múltiplos classificadores sequencialmente a um comitê, no qual cada novo classificador tenta corrigir o seu predecessor com base no erro residual anteriormente cometido (GÉRON, 2019). O XGB é uma versão otimizada do sistema de *boosting* com árvores, que foi criado para ser um modelo altamente escalável (CHEN; GUESTRIN, 2016)

Já no aprendizado não supervisionado, os dados de treinamento não estão identificados com uma solução(GÉRON, 2019). Nesse caso, o objetivo é encontrar estruturas (tais como agrupamentos) que possam fornecer informações sobre o conjunto de dados estudado (MURPHY, 2012). Este tipo de técnica também pode ser aplicada de forma mais abrangente, uma vez que não requer um especialista para identificar, manualmente, cada exemplo no conjunto de treinamento (MURPHY, 2012), algo que pode ser custoso em termos de tempo e de dinheiro (GÉRON, 2019).

Figura 5 – Estrutura de uma ANN. Os neurônios da primeira camada (à esquerda) ativam os neurônios da camada seguinde e assim sucessivamente, até que os neurônios da última camada (à direita) sejam ativados. Fonte: Murphy (2012, p. 564) (MURPHY, 2012).



2.3.2 Avaliação de classificadores

Para o caso de classificação binária (apenas duas classes), definimos uma das classes como positiva e a outra como negativa. Uma instância positiva do conjunto de teste corretamente classificada pelo algoritmo é chamada Verdadeiro Positivo (VP). De forma análoga, uma instância negativa corretamente classificada é denominada Verdadeiro Negativo (VN). Instâncias positivas e negativas erroneamente classificadas são chamadas Falso Negativo (FN) e Falso Positivo (FP), respectivamente (MURPHY, 2012). Diversas métricas de estimativa de erro do classificador podem ser calculadas a partir deste processo, tal como a acurácia (Equação 2.1), a revocação (Equação 2.2) e a precisão (Equação 2.4) (GÉRON, 2019).

$$\textit{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.1}$$

$$\textit{Sensibilidade} = \frac{VP}{VP + FN} \tag{2.2}$$

$$\textit{Especificidade} = \frac{VN}{VN + FN} \tag{2.3}$$

$$\textit{Precisão} = \frac{VP}{VP + FP} \tag{2.4}$$

$$\textit{Acurácia Balanceada} = \frac{Sensibilidade + Especificidade}{2} \tag{2.5}$$

$$\textit{F1-Score} = 2 * \frac{Sensibilidade * Precisão}{Sensibilidade + Precisão} \tag{2.6}$$

Comitês de classificadores podem também ser utilizados para combinar vários classificadores com o objetivo de gerar uma predição mais robusta (BISHOP, 2006). Para uma dada instância a ser classificada, cada classificador do comitê deve predizer à qual classe ela pertence e a decisão final pode ser formada por voto majoritário entre todos os classificadores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Em alguns casos, o conjunto de dados pode também apresentar desbalanceamentos em relação à proporção de instâncias que pertencem a cada classe. Este problema pode resultar no favorecimento pela classe predominante quando o algoritmo classificar novas instâncias (MOHAMMED; RAWASHDEH; ABDULLAH, 2020). Para contornar esta situação, podem ser utilizados algoritmos de *Undersampling*, no qual o número de instâncias da classe majoritária é reduzido para igualar a minoritária, ou algoritmos de *Oversampling*, que criam novas instâncias da classe minoritária a partir das instâncias originais, até que igualem às da classe majoritária (MOHAMMED; RAWASHDEH; ABDULLAH, 2020). Um algoritmo capaz de combinar ambas as técnicas é o *SMOTE Tomek*, no qual é inicialmente aplicada a técnica *Synthetic Minority Over-sampling Technique* (SMOTE) para criar novas instâncias sinteticamente, a partir das instâncias da classe minoritária, e então usa-se a técnica de *Undersampling Tomek links* para reduzir o número de instâncias ruidosas que foram geradas pelo SMOTE (BATISTA et al., 2003).

2.3.3 Calibração

Algoritmos de ML podem apresentar múltiplos hiper parâmetros que alteram como o aprendizado ocorre e devem ser calibrados para que a melhor configuração seja obtida. Os valores ideais de cada hiper parâmetro que resultariam no melhor modelo poderiam ser testados manualmente, porém essa tarefa é trabalhosa e demorada, dependendo da quantidade de hiper parâmetros que devem ser calibrados.

Para realizar a calibração de maneira automática, uma das possíveis técnicas utilizadas é o *Grid Search*, no qual, para um determinado hiper parâmetro, são definidos os valores mínimo, máximo e o incremento realizado para cada iteração do hiper parâmetro. Dessa forma, o algoritmo testa, inicialmente, o algoritmo de classificação com o valor mínimo para o hiper parâmetro a ser calibrado. A cada nova iteração, o algoritmo é testado, porém é acrescido o valor do incremento ao hiper parâmetro, até que seja atingido o valor máximo definido. A combinação que obtiver o melhor resultado baseado em métricas de estimativa de erro (Seção 2.3.2) é escolhida para ser utilizada no modelo final. É possível escolher qual métrica deve ser otimizada durante a calibração. Dessa forma, o processo será direcionado para obter o melhor valor para tal métrica, como, por exemplo, acurácia ou *F1-Score*. Esta técnica, porém, pode ser muito custosa quando são calibrados múltiplos hiper parâmetros, uma vez que o *Grid Search* testa todas as combinações possíveis de valores (Scikit Learn, 2022a).

Outra técnica também muito utilizada é o *Random Search*, no qual não é utilizado um incremento, uma vez que, para cada iteração, valores são escolhidos aleatoriamente para cada hiper parâmetro, respeitando os valores máximo e mínimo definidos (Scikit Learn, 2022b). O *Random Search* não realiza uma calibração tão precisa quanto o *Grid Search*, porém o seu custo pode ser melhor controlado, uma vez que é determinado pelo número de iterações totais que serão realizados, não dependendo do número de hiper parâmetros (GÉRON, 2019). Também é possível que o *Random Search* teste um número maior de valores para um determinado hiper parâmetro, uma vez que é utilizado um valor aleatório a cada iteração, enquanto o *Grid Search* apenas testa os valores pré-definidos com base no incremento (por exemplo, a execução do *Random Search* com 500 iterações resultará em 500 valores diferentes para cada hiper parâmetro, enquanto o *Grid Search* utilizará apenas os valores pré definidos) (GÉRON, 2019).

Para estimar o erro de cada configuração de parâmetros testados durante a calibração, a técnica de validação cruzada *k-fold* estratificado pode ser utilizada, na qual a amostra de treinamento é dividida em k partições, cada uma delas mantendo as proporções de classes originais. Uma dessas partições é então separada como amostra de teste, e as outras k-1 partições são utilizadas para treinamento. O algoritmo é então treinado sobre as partições de treinamento. O algoritmo realiza a predição de cada instância da partição anteriormente separada para teste e as métricas de erro são calculadas, tal como acurácia, precisão e revocação. Em seguida, uma das k-1 partições restantes é escolhida para teste e o processo se repete até que todas as

partições tenham sido utilizadas para teste. Ao final do processo, pode-se calcular a média das métricas de erro. A configuração dos parâmetros do algoritmo é então alterada e o processo realizado novamente (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Após testar diversas combinações de parâmetros, pode-se escolher a configuração que tiver apresentado os melhores valores médios das métricas de erro, por exemplo.

2.3.4 Redução de dimensionalidade

Um conjunto de dados utilizado para treinamento de métodos de ML pode apresentar uma quantidade de características elevada, resultando em um conjunto de alta dimensionalidade. Como consequência, além de tornar o treinamento do algoritmo mais lento, conjuntos de dados com grande quantidade de características podem ser bastante esparsos uma vez que espaços de dimensões maiores tendem a tornar as instâncias mais distantes entre si, tornando as predições realizadas pelos algoritmos de ML menos confiáveis (GÉRON, 2019).

Uma das técnicas frequentemente empregadas para realizar a redução de dimensionalidade de um conjunto de dados são os métodos de fusão de características. Estes algoritmos combinam características já existentes no conjunto de dados para criar novas características mais apropriadas para resolver o problema (GÉRON, 2019). A Análise de Componentes Principais, em inglês *Principal Component Analysis* (PCA), é um dos algoritmos mais utilizados para tal. Para uma base de dados de *n* características, esta técnica encontra os *n* Componentes Principais (CPs) que configurem uma base ortonormal(MURPHY, 2012). Estes *n* CPs são ordenados de acordo com o nível de variabilidade do conjunto de dados que elas representam, de forma que o primeiro CP não apenas apresenta a maior variabilidade entre todos os CPs, mas também representa o eixo de maior variabilidade possível para este conjunto de dados (GÉRON, 2019). Em seguida, pode-se escolher os *d* CPs de maior variabilidade como novas características do conjunto de dados, sendo $d \leq n$.

Além de algoritmos de fusão de características, técnicas de seleção de características também são utilizadas para reduzir a dimensionalidade de um conjunto de dados. Estes modelos buscam selecionar as características mais relevantes entre aquelas presentes no conjunto de dados (GÉRON, 2019).

Um dos algoritmos utilizados para tal é o *ReliefF*, no qual são selecionadas *m* instâncias do conjunto de dados aleatoriamente e são atribuídos níveis de relevância para cada característica com base na sua capacidade de distinguir ou não cada uma

das *m* instâncias selecionadas em relação às *k* instâncias mais próximas à ela, tanto da mesma classe que ela (chamadas *NearHit*), como de classe distinta (*NearMiss*) (KONONENKO; ŠIMEC; Robnik-Šikonja, 1997).

Outro algoritmo de seleção também utilizado é o Algoritmo Genético, em inglês *Genetic Algorithm* (GA). Este algoritmo treina um classificador específico utilizando subconjuntos aleatórios de características, chamados de geração zero. Os subconjuntos que apresentarem os melhores resultados com base em uma função de aptidão são mais susceptíveis a serem utilizados para cruzamento (*crossover*, uma combinação de parte das características selecionadas num subconjunto com parte das características de outro subconjunto) para criar novos subconjuntos para a geração seguinte. Os subconjuntos com melhores resultados de uma geração são também transferidos para a geração seguinte, de modo a manter as melhores soluções obtidas até o momento, e os subconjuntos de uma geração podem também sofrer mutação (as características podem ser selecionadas aleatoriamente para serem incluídas/excluídas no subconjunto) (MITCHELL, 1996).

2.4 Avaliação em simuladores RV

Simuladores de RV são sistemas computacionais capazes de criar ambientes virtuais para a simulação de situações reais. Estes tipos de simuladores são especialmente interessantes quando a habilidade que se pretende aprender é cara, perigosa ou inviável de se treinar no contexto real, tal como pilotagem de avião ou determinados procedimentos médicos (SCHMIDT; LEE, 2014).

Com o recente avanço da tecnologia, simuladores de RV voltados para treinamento de aprendizes na área da saúde têm ganhado popularidade. Uma das vantagens desse tipo de sistema é o treinamento realista de aprendizes, sem o risco de se cometer erros em paciente reais(BURDEA; COIFFET, 2003). Atualmente, diversos simuladores de RV para treinamento de procedimentos da área da saúde podem ser encontrados na literatura voltados para diversos tipos de procedimentos, tal como laparoscopia (PRASAD et al., 2018), cirurgias neurológicas(CHAN et al., 2013), aplicação de anestesias(CORREA; TORI; NUNES, 2013), procedimentos guiados por ultrassom (Wing-Yin Chan et al., 2012), preparo da coroa dentária (RHIENMORA et al., 2011), entre muitas outras.

Muitos desses simuladores possuem algum tipo de retorno de força, seja ele ativo

ou passivo. Por exemplo, o simulador LapSim (Surgical Science, Göteborg, Suécia) (Surgicalscience, 2020) é um simulador de laparoscopia de RV que utiliza dois dispositivos hápticos ativos para simular os instrumentos utilizados durante procedimentos de laparoscopia. O rastreamento dos instrumentos é feito pelos dispositivos hápticos, possibilitando a coleta de diversas métricas de desempenho objetivas, tais como o comprimento total da trajetória percorrida pelos instrumentos e os pontos de contato em locais errados entre o instrumento e as estruturas internas do paciente. Já o ProMIS (Haptica, Dublin, Irlanda) (SICKLE et al., 2005) é também um simulador de laparoscopia, porém com retorno de força passivo. Neste sistema, modelos físicos são utilizados para simular o retorno háptico das estruturas internas do paciente. Um conjunto de câmeras realizam o rastreamento dos instrumentos utilizados na simulação e métricas de desempenho, como o tempo transcorrido para a realização do procedimento, distância percorrida pelo instrumento e suavidade nos movimentos, são registradas pelo sistema.

Alternativamente, podem ser realizadas avaliações automáticas, que ocorrem em tempo real, durante o treinamento. Avaliação automática será considerada neste trabalho como qualquer tipo de avaliação que utilize métricas objetivas coletadas pelo simulador com o objetivo de medir o desempenho de habilidades sensório-motoras em um procedimento. Métricas de desempenho objetivas, por sua vez, são aquelas calculadas a partir dos dados coletados pelo simulador e que não utilizam nenhuma informação obtida por meio de análises de especialistas ou por qualquer outra pessoa envolvida. Exemplos de métricas objetivas são a velocidade média do instrumento utilizado durante a simulação de um procedimento médico e a distância total que ele percorreu.

A capacidade de embutir avaliações automáticas de desempenho que simuladores de RV podem apresentar possibilita que esses sistemas possam proporcionar *feedback* aos aprendizes durante o treinamento, possibilitando uma maior flexibilização do aprendizado, sem a necessidade da supervisão de um instrutor durante parte do processo. Muitos métodos são utilizados para realizar este tipo de avaliação. Neste trabalho, o foco será na aplicação de técnicas de ML, mas outros métodos serão apresentados no Capítulo 3.

Diversas validações também são citadas na literatura para que se possa inferir a qualidade de um simulador para treinamento médico. A validade de constructo, em particular, verifica se o simulador é capaz de diferenciar entre usuários de diferentes níveis de habilidades (MCDOUGALL, 2007). Usualmente, são calculadas métricas

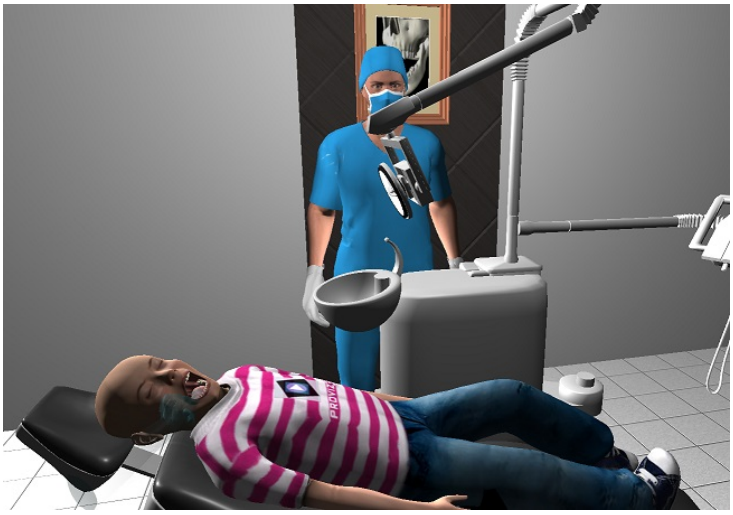
de desempenho a partir de dados coletados durante a simulação de procedimentos com participantes experientes e inexperientes (e outros possíveis graus intermediários) e, então, são realizados testes para verificar se as métricas apresentam diferenças significativas entre os grupos. Considerando que participantes que possuem mais tempo de treinamento no procedimento real provavelmente obteriam resultados superiores aos dos participantes com pouca ou nenhuma experiência, espera-se que essa discrepância no desempenho também se reflita no simulador, mostrando que as habilidades desenvolvidas no mundo real foram transferidas para o ambiente virtual (MCDOUGALL, 2007).

2.5 Simulador VIDA Odonto

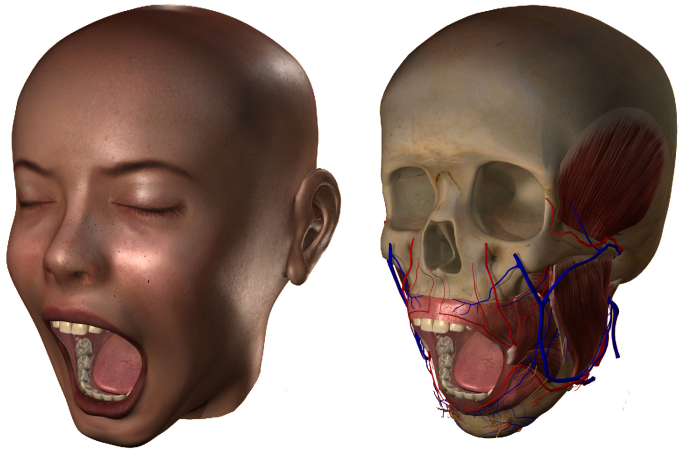
O VIDA Odonto (TORI et al., 2018) é um simulador de procedimentos odontológicos em RV, desenvolvido em uma parceria entre o Laboratório de Tecnologias Interativas (Interlab - EPUSP), o Laboratório de Aplicações de Informática em Saúde (LApIS - EACH/USP) e o Laboratório de Simulação e Treinamento (LaSiT - FOB/USP). O sistema, criado usando a *engine* de jogos *Unity* (Unity Technologies, 2020), possibilita que o usuário interaja com um paciente virtual, dentro de um consultório odontológico (Figura 6a). O modelo do paciente é uma criança de 10 anos com as partes mais relevantes dos músculos, nervos, vasos sanguíneos e ossos (Figura 6b). O aluno pode usar um HMD que proporciona imersão total e utiliza uma seringa carpule real, acoplada a um dispositivo háptico para realizar o procedimento de anestesia (Figura 7).

Atualmente, o VIDA Odonto é utilizado para a prática da técnica direta de anestesia para bloqueio do nervo alveolar inferior. O sistema possui um módulo de ensino, no qual o aluno se encontra dentro do consultório odontológico e assiste à explicação do procedimento, feita por uma professora, dentro da realidade virtual. Durante a explicação, são usadas marcações para auxiliar o aluno a visualizar as etapas do procedimento (Figura 8). Para executar corretamente a tarefa, a seringa deve seguir uma trajetória a partir do dente canino do lado oposto ao da região que se deve anestesiar, inserindo a agulha 2 cm acima do último dente molar. Essa trajetória garante uma melhor precisão no ponto de inserção e no ângulo de punção e a prática incorreta deste procedimento pode resultar em múltiplas complicações, tal como a quebra da agulha no ponto da injeção, paralisia facial, hematomas causados por danos a vasos sanguíneos, entre outras(KHALIL, 2014). Após ser dada a explicação, o aluno assiste

Figura 6 – a) Consultório odontológico virtual. b) Modelo da cabeça do paciente virtual.



(a)



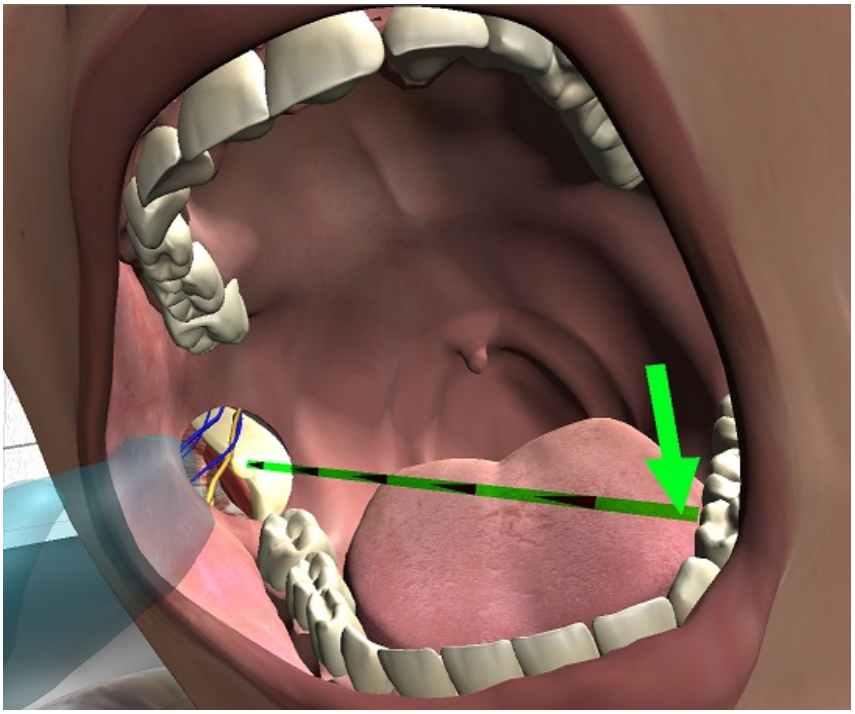
(b)

à professora realizando o procedimento no ambiente virtual.

No módulo de treinamento, o aluno pode realizar o procedimento de anestesia no paciente virtual. Durante a execução da tarefa, o retorno háptico pode ser percebido quando a ponta da agulha penetra a mucosa da boca e a resistência do tecido é simulada. As forças que definem as interações do dispositivo háptico foram modeladas com base no *feedback* de especialistas. O *plugin* do dispositivo háptico para Unity (School of Simulation & Visualization - Glasgow School of Art, 2016) possibilita que seja configurado o valor da resistência de tecidos moles, tal como a mucosa da boca, até que eles sejam rompidos pela agulha da seringa. É possível também definir parâmetros para a resistência apresentada na agulha enquanto ela desliza dentro do tecido após penetrá-lo. Ossos da mandíbula e dentes foram definidos como objetos impenetráveis, ou seja, este tipo de objeto não pode ser rompido pela agulha.

Figura 7 – Aluno executando o treinamento no simulador VIDA Odonto

Figura 8 – Marcações mostradas durante a explicação do procedimento de anestesia do nervo alveolar inferior.



Durante a realização do procedimento, o simulador é capaz de coletar a posição e orientação da seringa ao longo do tempo, assim como a posição e rotação da cabeça do aluno. A posição de um objeto, no ambiente virtual, refere-se a um vetor tridi-

mensional, que tem sua extremidade de origem em um ponto definido como a origem da cena, de coordenadas $(0, 0, 0)$, e a outra extremidade no centro do objeto. Dessa forma, o vetor pode ser representado apenas pelas coordenadas (x, y, z) do objeto, uma vez que as coordenadas do ponto de origem são fixas. Já a orientação é determinada pela sequência de rotações que devem ser realizadas no objeto, ao redor dos eixos cartesianos do ambiente virtual, a partir de uma orientação inicial. Essas informações são coletadas 60 vezes por segundo, de forma que todos os movimentos da seringa e da cabeça do paciente sejam coletados. Esses dados são utilizados no módulo de reprodução, no qual toda a trajetória da seringa é reproduzida, podendo ser utilizada tanto para que alunos possam visualizar o procedimento de um especialista, durante o aprendizado, quanto para que o professor possa rever o desempenho dos estudantes, quando for avaliá-los. Essa reprodução pode ser realizada em um monitor, de forma não imersiva, no qual o simulador reproduz a visão do usuário no momento que realizou o procedimento, ou com um HMD, de forma imersiva , possibilitando que o espectador tenha liberdade de movimento dentro do ambiente virtual.

2.6 Considerações Finais

Este capítulo apresentou conceitos importantes em relação a diferentes áreas do conhecimento, as quais terão papel central no desenvolvimento de um método de avaliação automática de habilidades sensório-motoras em simuladores de RV com retorno háptico. No próximo capítulo será apresentado um levantamento bibliográfico sobre as técnicas e metodologias encontradas na literatura para realizar avaliação de aprendizes em simuladores de RV voltados para a saúde.

3 REVISÃO BIBLIOGRÁFICA

Neste capítulo é apresentada uma revisão sistemática da literatura abordando as técnicas e metodologias empregadas na avaliação automática de desempenho em simuladores de RV com retorno háptico.

Este capítulo está dividido em subseções referentes às etapas do processo de revisão da literatura. A Seção 3.1 apresenta os métodos empregados; a Seção 3.2 exibe os resultados obtidos; a Seção 3.3 discute os resultados; enquanto a Seção 3.4 expõe as considerações finais.

3.1 Método

A revisão conduzida buscou responder à seguinte questão: “Quais as metodologias e técnicas utilizadas para avaliar desempenho utilizando métricas objetivas em simuladores de RV hápticos voltados para procedimentos médicos?”. As questões específicas foram definidas como: “Quais procedimentos voltados para a saúde foram simulados nos estudos?”, “Quais as métricas de desempenho calculadas?”, “Quais tipos de experimentos foram conduzidos?” e “Quais técnicas foram aplicadas para avaliar desempenho?”.

A *string* de pesquisa utilizada foi: “(simulation OR simulator OR “virtual reality”) AND (evaluation OR analysis OR measurement OR assessment) AND (surgery OR “surgical skills” OR “medical procedure” OR “medical training” OR dentistry) AND (haptic* OR “force feedback”)”. A *string* foi utilizada para realizar buscas nas plataformas Scopus (*Elsevier*) (ELSEVIER, 2021) , PubMed (*NCBI*) (NCBI, 2021), IEEE *Xplore* (*Institute of Electrical and Electronics Engineers*) (IEEE, 2021) e ACM *Digital Library* (*Association of Computing Machinery*) (ACM, 2021). A pesquisa foi realizada no dia 14 de setembro de 2019 e limitada à publicações nas áreas de medicina, odontologia, ciência da computação e engenharia. A pesquisa foi então repetida no dia 25 de janeiro de 2021 para incluir estudos publicados desde que a primeira busca foi conduzida.

Foram incluídos apenas estudos originais em inglês que utilizaram métricas objetivas para avaliar desempenho em simuladores de RV ou RA com retorno háptico e voltados para procedimentos médicos. Para que esses estudos pudessem ser incluídos, eles deveriam ter conduzido e apresentado, detalhadamente, experimentos realizados nos respectivos simuladores. Como critérios de exclusão, foram retirados os estudos com menos de cinco páginas, os estudos que utilizaram simuladores sem elementos virtuais, os estudos que não explicitaram os métodos e técnicas empregados no processo de avaliação, os estudos que investigaram procedimentos não médicos e os estudos que realizaram avaliação de aspectos não relacionados a habilidades sensório-motoras.

O fluxograma que detalha o processo de revisão é apresentado na Figura 9. Dos 1471 artigos encontrados inicialmente, 496 foram identificados como duplicados. Foram excluídos todos os estudos que atenderam, pelo menos, um critério de exclusão e, dos 118 artigos restantes, foram calculadas notas de qualidade a partir de uma média ponderada dos itens da tabela 1 e foram removidos todos os artigos com uma nota abaixo de 4.00. Este critério foi definido a partir de uma avaliação empírica dos estudos, a fim de garantir que os trabalhos incluídos fossem relevantes para este projeto.

Figura 9 – Fluxograma do processo de revisão sistemática da literatura.

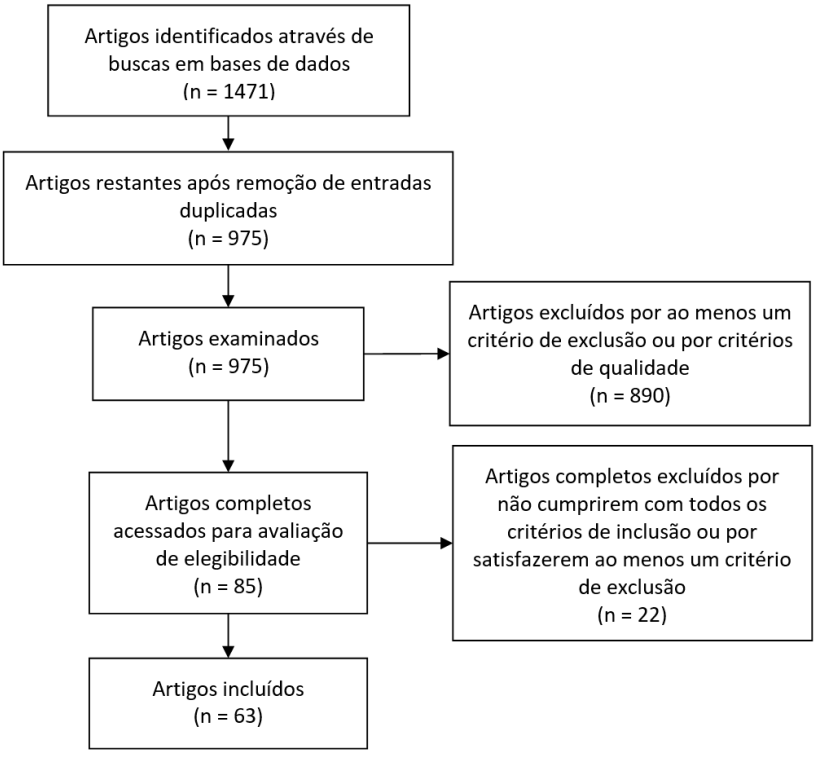


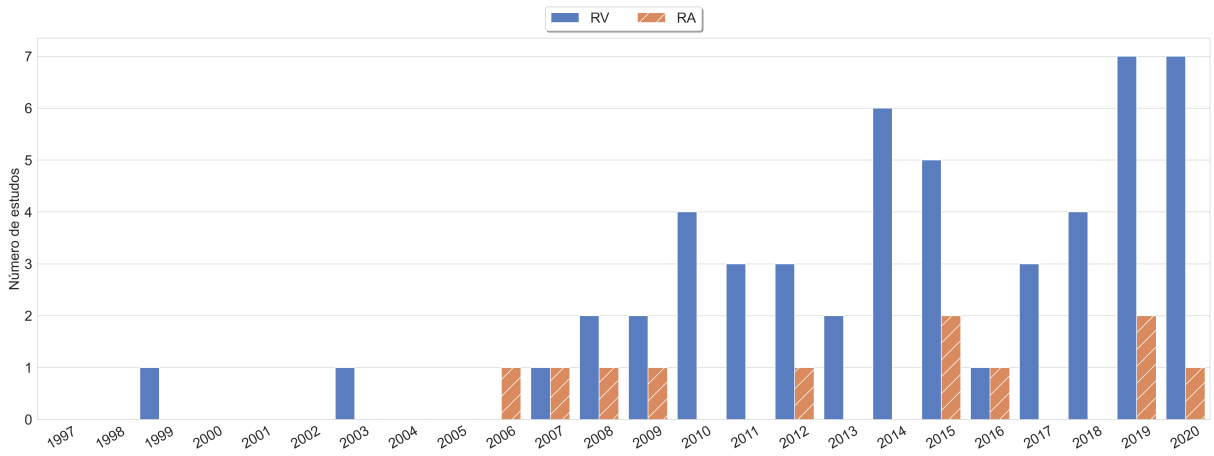
Tabela 1 – Critérios de qualidade para os estudos.

Critério	Peso	Valor	Descrição
Índice h5 normalizado	1	Contínuo (0 a 1)	A normalização foi separada entre artigos de conferência e artigos de revista.
Estudos relacionados	0,5	Binário (0 ou 1)	O artigo apresentou trabalhos relacionados na introdução ou em uma seção separada.
Tabela comparativa	0,5	Binário (0 ou 1)	Os autores apresentaram uma tabela comparando o seu estudo com trabalhos similares.
Figuras	0,5	0; 0.5 ou 1	1- Figuras que apresentam tanto a parte física como a parte virtual do simulador. 0,5- Figuras que apresentam apenas a parte física ou a parte virtual do simulador.
Diagrama do simulador	1	Binário (0 ou 1)	O artigo apresenta um diagrama claro que descreve as funcionalidades do simulador.
Procedimento	1	0; 0,5 ou 1	1- O procedimento foi descrito em detalhes. 0,5- Apenas o nome e uma explicação superficial do procedimento foram fornecidos.
Usuário final	1	Binário (0 ou 1)	O experimento apresentado no artigo foi realizado com o usuário final.
Tabela com resultados	1	Binário (0 ou 1)	O artigo utilizou tabelas claras para apresentar os resultados finais.
Limitações	0,75	0; 0,5 ou 1	1- As limitações do estudo foram apresentadas e discutidas. 0,5- As limitações do estudo foram apresentadas, mas não foram discutidas.
Passos futuros	0,5	0; 0,5 ou 1	1- Os passos futuros do estudo foram apresentadas e discutidos. 0,5- Os passos futuros foram apresentadas, mas não foram discutidos.

3.2 Resultados

A Figura 10 mostra a disposição dos estudos incluídos nesta revisão sistemática, com base no ano de publicação, e divididos em simuladores de RV e RA. Dos 63 estudos incluídos na revisão, 52 utilizaram simulador de RV, enquanto que apenas 11 utilizaram simulador de RA.

Figura 10 – Número de estudos incluídos por ano e divididos entre simuladores de RV e RA.



3.2.1 Procedimentos

Os artigos incluídos foram classificados com base no tipo de procedimento simulado no estudo. Ao todo, foram definidas quatro categorias, sendo elas Cirurgia, Navegação, Osso e Perfuração. Um procedimento pode ser classificado em mais de uma classe.

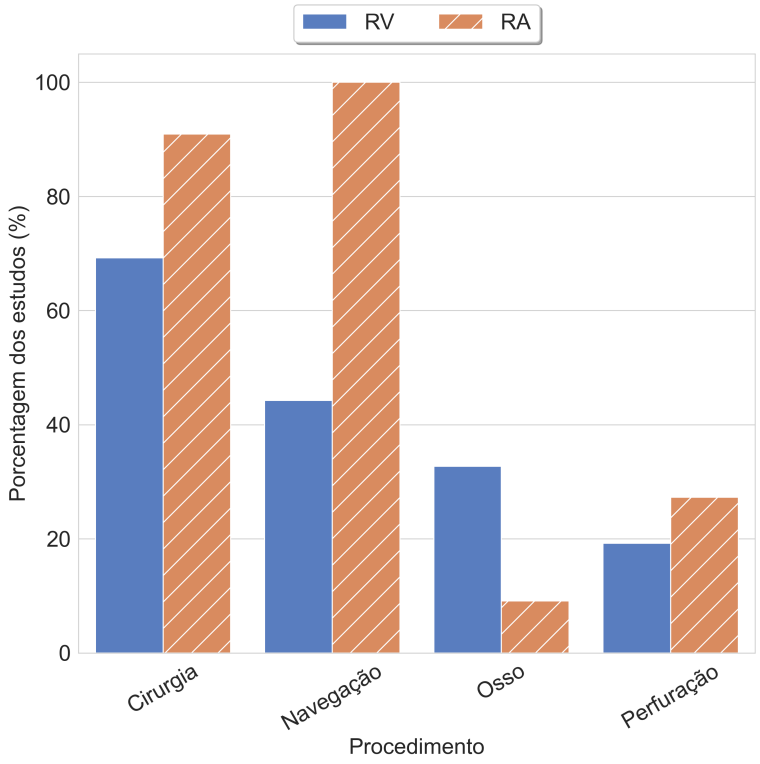
A categoria Cirurgia engloba todos os procedimentos cirúrgicos e de suturas, tais como laparoscopias e remoção de tumores, tendo sido responsável por 73% do número total de estudos. Procedimentos classificados como Navegação são aqueles em que é necessário localizar uma estrutura ou um alvo dentro do corpo do paciente, podendo ser citados como exemplo a palpação e os procedimentos guiados por imagem. No total, 54% dos estudos foram classificados como Navegação. Na categoria Osso, encontram-se todos os procedimentos em relação aos quais um osso ou um grupo de ossos são o foco. Esta classe corresponde a 29% dos estudos e nela foram incluídos procedimentos como o de corte de ossos e alguns procedimentos odontológicos. Por fim, na categoria Perfuração (21% dos estudos), foram inseridos todos os estudos, em relação aos quais uma parte importante do procedimento requer a utilização de uma agulha ou de um instrumento similar para perfurar o paciente, por exemplo, o posicionamento de catéter.

A Figura 11a mostra a porcentagem dos trabalhos que foram classificados em cada categoria de procedimento, separados em RV e RA. É possível perceber que a maioria dos estudos foram categorizados como Cirurgia e/ou Navegação, tanto para RV como RA.

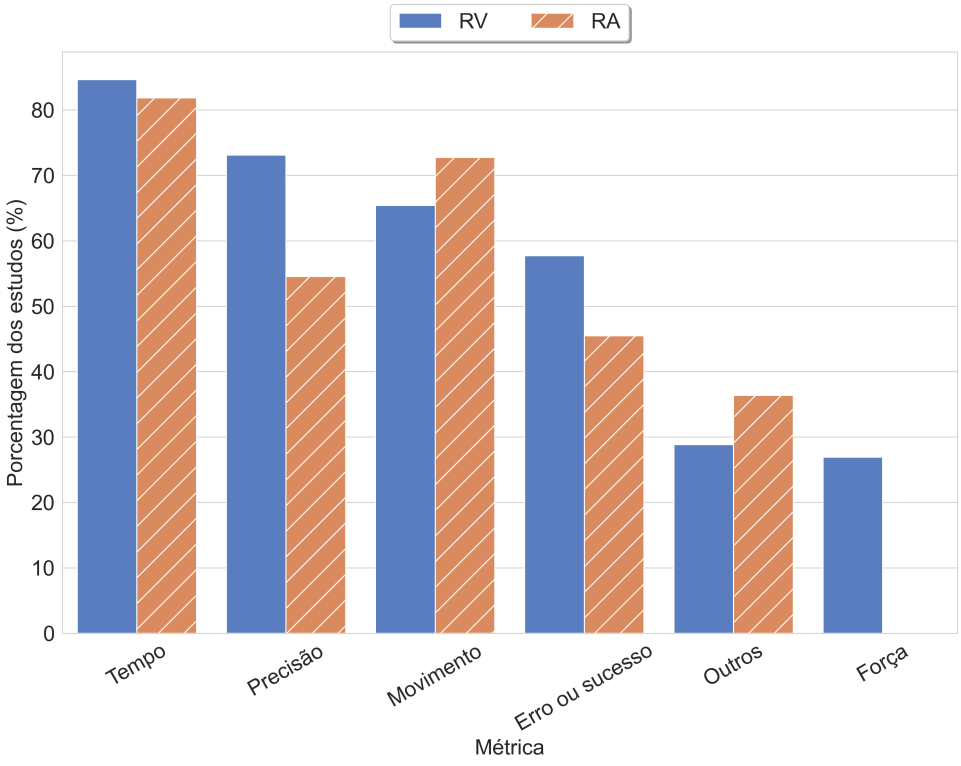
3.2.2 Métricas

As métricas utilizadas nos estudos foram combinadas com base no tipo de informação que mediram. Essa abordagem possibilitou o agrupamento de métricas que foram utilizadas em diferentes procedimentos. Por exemplo, Chen et al. (2019) (p. 5, tradução nossa) calculou a “distância entre a posição final da ponta da agulha para o centro da veia” em um cateterismo venoso central e Roitberg et al. (2015) (p. 1166, tradução nossa) mediu a “distância euclidiana entre a ponta da broca e o ponto alvo” em um simulador de neurocirurgia . Embora as métricas tenham sido aplicadas em procedimentos diferentes, ambas tinham como objetivo medir a distância entre o instrumento utilizado e um alvo pré-definido; portanto ambas foram agrupadas na métrica

Figura 11 – a) Porcentagem dos estudos que foram classificados em cada categoria de procedimento, separados em RV e RA. b) Porcentagem dos estudos que utilizaram cada uma das métricas, separados em RV e RA.



(a)



(b)

combinada “Distância do instrumento ao alvo”.

A Tabela 2 mostra as 36 métricas criadas a partir da combinação das métricas dos estudos. Elas foram então classificadas de acordo com o aspecto do procedimento em que focaram, sendo que algumas delas foram atribuídas a duas classes. As seis classes de métricas definidas foram: Tempo, Precisão, Movimento, Força, Erro ou Sucesso e Outro.

Tabela 2 – Lista das métricas de desempenho combinadas encontradas nos estudos.

Métrica	**Classe**	**Porcentagem (%)**
Tempo transcorrido durante o experimento por completo	Tempo	63
Comprimento do trajeto do instrumento	Movimento	30
Tempo transcorrido em uma etapa chave do procedimento	Tempo	22
Toques errôneos no tecido	Erro ou Sucesso / Precisão	16
Lesões em órgãos ou estruturas	Erro ou Sucesso / Precisão	16
Número de repetições em uma parte do procedimento	Precisão	16
Distância do instrumento ao alvo	Movimento / Precisão	14
Quantidade de estrutura alvo removida	Precisão	13
Técnica específica corretamente/incorretamente aplicada	Erro ou Sucesso / Outro	11
Quantidade de tecido saudável removido	Erro ou Sucesso / Precisão	10
Número de etapas do procedimento realizadas com sucesso/fracassadas	Erro ou Sucesso	10
Força média aplicada	Força / Tempo	8
Economia de movimento (desvio em relação à trajetória ideal)	Movimento / Precisão	8
Posição/orientação do instrumento através do tempo	Movimento / Tempo	8
Métricas específicas do procedimento	Outro	8
Visibilidade de estruturas chave	Outro	8
Ângulo entre o instrumento e o alvo	Movimento / Precisão	6
Velocidade média	Movimento / Tempo	6
Quantidade de sangue perdido pelo paciente	Erro ou Sucesso	6
Profundidade de incisão/inserção do instrumento	Movimento / Precisão	6
Força máxima aplicada	Força	6
Número de erros totais	Erro ou Sucesso	6
Número de vezes que a força aplicada estava acima de um limiar	Força	5
Acúmulo do ângulo do instrumento	Movimento	5
Sistema progressivo de escore	Outro	5
Firmeza (Movimentos repentinos)	Movimento / Tempo	5
Aceleração máxima	Movimento / Tempo	3
Modelagem de estrutura	Precisão	3
Visibilidade do instrumento	Outro	3
Eficiência (precisão/tempo)	Precisão / Tempo	2
Variações na força	Força / Tempo	2
Comprimento da incisão	Movimento / Precisão	2
Similaridade entre trajetórias	Movimento / Precisão	2
Número de vezes que a velocidade estava acima de um limiar	Movimento / Tempo	2

Métricas de Tempo, utilizadas em 84% dos estudos incluídos, são aquelas referentes ao tempo transcorrido durante o experimento ou métricas que medem variações no

tempo, por exemplo a velocidade e aceleração do instrumento. Métricas de Precisão são aquelas relacionadas à sutileza ou à exatidão dos movimentos realizados e foram empregadas em 70% dos artigos. Métricas de Movimento são calculadas a partir da posição e orientação do instrumento utilizado durante a simulação, sendo que 42% dos estudos aplicaram métricas desta classe. Métricas de Força, empregadas em 22% dos estudos, são relacionadas à força aplicada pelo participante no instrumento ou à força gerada pelo dispositivo háptico. Métricas de Erro ou Sucesso se referem aos erros cometidos pelo participante durante o procedimento ou às etapas realizadas de forma correta, sendo utilizadas em 35% dos artigos. Por fim, as métricas que possuem foco em algum aspecto do procedimento não abordado por outras classes ou que focam em um aspecto específico de determinado procedimento pertencem à classe de métricas Outro, sendo utilizadas em 30% dos estudos incluídos.

Foi observado que a classe de métricas Tempo foi a mais utilizada nas categorias de procedimentos Cirurgia, Osso e Navegação e a segunda mais utilizada na categoria Perfuração, juntamente com as métricas de Movimento e atrás apenas de métricas de Precisão. Métricas de Força foram pouco utilizadas em todas as categorias de procedimentos, com exceção da categoria Osso.

A Figura 11b mostra a porcentagem de estudos em simuladores de RV e RA que utilizaram cada uma das métricas. Pode-se notar que uma menor porcentagem de estudos com simuladores de RA utilizaram métricas de Precisão, Erro ou Sucesso e Força, sendo que a última não foi utilizada em nenhum desses estudos.

3.2.3 Tipos de experimentos

Foram identificadas seis categorias de experimento: Especialistas *versus* Novatos (EN), Diferentes versões do procedimento (DV), Aperfeiçoamento ao longo de múltiplas sessões (MS), Com e sem Retorno de Força (RFF), Diferentes tipos de treinamento (DT) e treinado *versus* não treinado (TN). Cada estudo poderia ser classificado em mais de uma categoria.

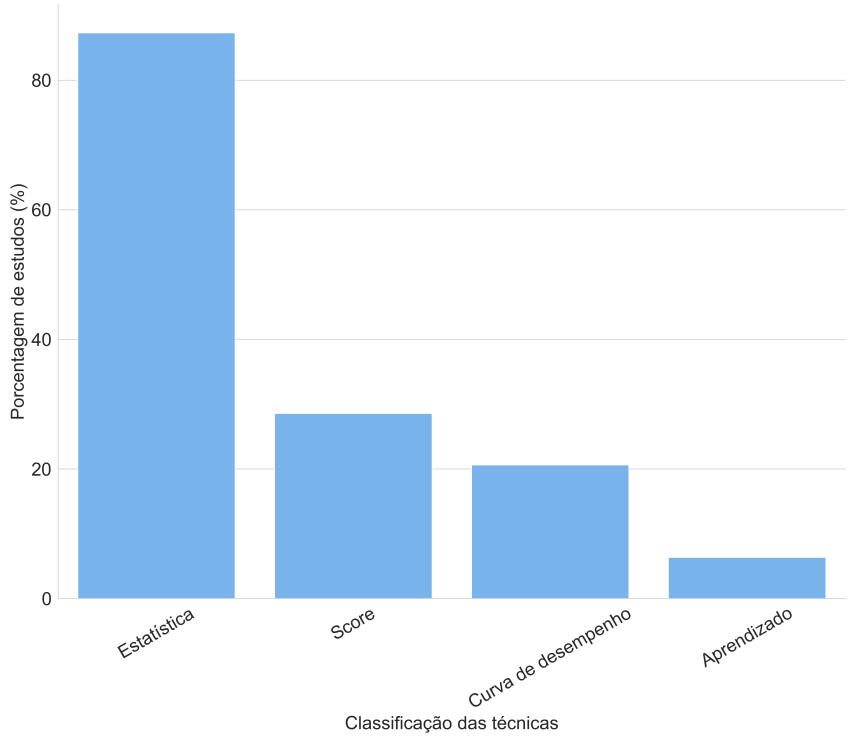
Experimentos EN comparam o desempenho entre participantes dotados de diferentes níveis de experiência no procedimento simulado e foi a categoria mais frequente nos estudos (73%). Experimentos DV (24%) comparam participantes que realizaram diferentes versões do procedimento no simulador ou que utilizaram diferentes recursos. Experimentos RFF (14%) são similares aos DV, porém a diferença entre as versões dos procedimentos realizados se dá em relação à existência ou não de retorno

háptico. Experimentos MS (29%) estudam a melhora no desempenho do participante durante múltiplas sessões no mesmo simulador, podendo ocorrer ou não algum tipo de treinamento entre as sessões. Por fim, experimentos DT (10%) comparam participantes com um mesmo nível de experiência em geral, que passaram por diferentes tipos de treinamentos antes de realizar o procedimento no simulador. Foram também classificados como TN (6%) os estudos no qual existe a comparação entre participantes que realizaram treinamento e participantes sem treinamento.

3.2.4 Técnicas de avaliação

As técnicas de avaliação de habilidades sensório-motoras encontradas nos estudos foram classificadas em 4 categorias: Estatística, Curva de Desempenho, Escore e Aprendizado. Alguns dos estudos utilizaram mais de um tipo de técnica durante a avaliação.

Figura 12 – Porcentagem dos estudos incluídos que utilizaram cada categoria de técnica de classificação.



As técnicas Estatísticas foram utilizadas em 87% dos estudos, sendo a categoria mais citada (Figura 12). Nesses estudos, foram aplicados testes de significância para verificar se, dada uma métrica coletada no simulador, seria possível distinguir diferentes grupos de participantes (por exemplo, novatos e experientes) utilizando a métrica

em questão. Este tipo de abordagem é comumente utilizado para verificar a validade de constructo de um simulador (descrita na Seção 2.4).

As Curvas de Desempenho foram descritas na Seção 2.2 e utilizadas em 21% dos estudos. Este tipo de técnica pode ser utilizado tanto para avaliar se um determinado participante teve melhora na métrica em estudo, como para comparar o progresso de participantes de diferentes grupos em múltiplas seções, no simulador. Neste caso, espera-se que participantes mais experientes no procedimento apresentem um menor índice de melhora entre as seções de treinamento, enquanto que novatos apresentem grande melhora, uma vez que ainda estão aprendendo o procedimento.

Técnicas do tipo Escore (29%) geralmente combinam múltiplas métricas em uma função matemática definida pelos pesquisadores para gerar um único valor final (chamado escore). Esta abordagem deve representar o desempenho do participante de forma direta, partindo do pressuposto de que a função definida é capaz de refletir, com precisão, as habilidades do participante.

Técnicas classificadas como Aprendizado utilizaram algoritmos de ML (definido na Seção 2.3) para realizar a avaliação dos participantes. Foram incluídos apenas quatro estudos que utilizaram técnicas de Aprendizado (6%), dois deles do mesmo grupo, que fizeram uso da técnica HMM em um procedimento de preparo para coroa dentária (RHIENMORA et al., 2011; RHIENMORA et al., 2010).

3.3 Discussão

A Figura 10 mostra que, entre os estudos incluídos nesta revisão, ocorreu uma clara predominância de artigos publicados nos últimos 15 anos e, em especial, nos últimos 7 anos. Esta tendência pode ser resultante dos recentes avanços nas tecnologias de RA e RV, que tornaram os simuladores mais viáveis para treinamento e educação (JERALD, 2016).

Métricas. A predominância de métricas de Tempo pode ser decorrente da facilidade de coleta e interpretação desse tipo de dado. Métricas relacionadas ao tempo transcorrido durante o procedimento estiveram presentes em 75% dos estudos, mas a execução de um procedimento de forma mais rápida não indica, necessariamente, que o aprendiz é mais habilidoso, uma vez que é possível realizar um procedimento rapidamente, porém com resultados imprecisos (FUCENTESE et al., 2015; LEBLANC et al., 2013), enquanto outro aprendiz pode realizar o procedimento com mais cuidado

e, portanto, levar mais tempo para finalizá-lo. Neste contexto, a utilização combinada do tempo transcorrido durante o procedimento com métricas de Precisão e de Erro ou Sucesso pode refletir se a atuação dos participantes que realizaram o procedimento mais rapidamente resultou em mais erros ou menor acurácia. Um exemplo disso é a métrica “Eficiência”, definida em (EVE et al., 2014) como a razão entre a precisão e o tempo transcorrido durante um procedimento de remoção de cáries. Foi também observado que a falta de familiaridade de um indivíduo com um simulador pode resultar em uma execução mais lenta do procedimento, ainda que esse participante seja experiente no procedimento real ou em outro simulador, tal como pode ser observado em (LEBLANC et al., 2013), em que alguns participantes que estavam no ápice do treinamento em um simulador físico realizaram o procedimento mais lentamente em um simulador de RV, quando comparados com participantes com menos treinamento. Apesar da importância de métricas relacionadas com o tempo transcorrido no experimento, elas não devem ser utilizadas isoladamente para validação (FUCENTESE et al., 2015), pois deve-se verificar se a qualidade da execução foi mantida. Portanto, ao combinar métricas de tempo com outras métricas, é necessário cuidado ao atribuir a importância de cada uma delas na sua composição, uma vez que esta definição pode ser altamente dependente do procedimento em questão.

Apesar de todos os estudos incluídos utilizarem simuladores com *feedback* de força, seja passivo em simuladores de RA ou ativo em simuladores de RV, métricas de Força foram as menos utilizadas. Entre os possíveis motivos para isso, temos: (a) a predominância de dispositivos hápticos comerciais que podem não disponibilizar funções de coleta de dados de Força, (b) a maior dificuldade de interpretação de dados de Força quando comparada às outras classes de métricas e (c) a maior dificuldade em simuladores de RA coletarem este tipo de dado por utilizarem háptico passivo, como pode ser visto na Figura 11b. Por fim, enquanto pesquisadores que estudaram procedimentos da categoria Osso podem ter considerado que a força utilizada pelo participante durante o procedimento é de grande importância, pesquisadores das outras categorias de procedimentos podem ter considerado outras métricas mais importantes.

Apesar de métricas da classe Tempo serem utilizadas em quase 84% dos estudos incluídos, apenas 41% utilizaram mais de três classes de métricas, indicando que as características das interações hápticas foram pouco exploradas nos estudos. Esse fato pode ser decorrente da utilização de simuladores comerciais em parte dos estudos incluídos, como já mencionado. Adicionalmente, pode ser resultado do processo

utilizado por pesquisadores para criar novas métricas, no qual um especialista é consultado sobre os aspectos do procedimento que considera como mais importantes, o que pode resultar na exclusão de métricas que esse especialista não teria acesso em um procedimento comum, tal como a força utilizada durante o procedimento.

Técnicas. As técnicas Estatísticas encontradas nos estudos foram majoritariamente aplicadas à cada uma das métricas individualmente, o que pode limitar consideravelmente a capacidade do simulador de avaliar habilidades, diferentemente de técnicas de Escore ou Aprendizado, que utilizam várias métricas como entrada.

Foram retornados na nossa pesquisa apenas cinco artigos que utilizaram técnicas de Aprendizado para avaliação de desempenho em simuladores de RV ou RA para procedimentos médicos, tendo um deles sido excluído por critérios de qualidade(MORRIS et al., 2006), porém foram encontrados outros trabalhos relacionados na literatura (Winkler-Schwartz et al., 2019; DIAS; GUPTA; YULE, 2019). Uma análise dos artigos retornados na pesquisa, mas que não atenderam aos critérios de inclusão e exclusão, nos mostrou que técnicas de ML foram aplicadas nos estudos para a avaliação de desempenho, porém em simuladores físicos (Rafii-Tari et al., 2017), modelos animais (ROSEN et al., 2006; ROSEN et al., 2000; ROSEN et al., 2001) e simuladores de RV não voltados para a saúde (P; MENON; RAO, 2018). Também foram encontradas aplicações não voltadas para avaliação de habilidades, tais como deformação de tecidos moles e interação háptica (MA et al., 2004; DEO; DE, 2010; DE et al., 2011) e auxílio ao praticante durante o procedimento (ENGELHARDT et al., 2014). Dias, Gupta e Yule (2019) sugerem que o número de estudos utilizando ML para avaliar competência em procedimentos médicos tem crescido nos últimos anos, porém a maioria deles ainda está em estágio inicial e necessita de uma validação mais rigorosa. Os resultados dessa revisão mostraram que técnicas de ML ainda foram pouco exploradas no contexto apresentado, mas este cenário pode mudar nos próximos anos, com a crescente popularização do ML.

Diversas formas de utilizar técnicas de ML podem ser exploradas na avaliação de desempenho em simuladores de RA ou RV. Algoritmos de classificação poderiam ser treinados com os dados de especialistas, novatos e possíveis níveis intermediários de experiência para classificar novos participantes de acordo seu nível de habilidade. Essa abordagem assumiria que praticantes mais experientes no procedimento real seriam mais habilidosos que aqueles menos experientes e essas habilidades foram transferidas para o ambiente virtual. Dessa forma, um futuro participante que seja classificado como experiente seria mais habilidoso do que um participante classifi-

cado como novato. Outra possível abordagem requisitaria de um especialista que pudesse avaliar todas as execuções do procedimentos realizadas pelos participantes e classificá-los com base no desempenho observado em aprovados ou não aprovados (e possíveis classificações intermediárias) ou criar notas numéricas que poderiam ser utilizadas por algoritmos de regressão. Esta solução, porém, possui maior grau de subjetividade e pode ser sensível a diversos fatores, como, por exemplo, o cansaço do especialista após uma longa sessão de avaliação. Nesta abordagem, o algoritmo tentaria replicar o método de avaliação do especialista.

Fatores limitantes. Muitos estudos reportaram que o número pequeno de participantes pode ter limitado o poder estatístico das técnicas utilizadas (RHIENMORA et al., 2010; SUEBNUKARN et al., 2010; HUANG et al., 2018; REILINK et al., 2011; BITTNER et al., 2010; FIARD et al., 2014; RHIENMORA et al., 2011; O'TOOLE et al., 1999; SIKDER et al., 2015; BOUAICHA et al., 2020). Alguns autores também acreditam que procedimentos mais simples (Gélinas-Phaneuf et al., 2014; WANG et al., 2015) e tarefas de curta duração (AZARNOUSH et al., 2017; AKHTAR et al., 2015) possuem uma capacidade menor de diferenciar o nível de experiência dos participantes, além de que partes específicas do procedimento podem requerer mais habilidade e experiência, tornado-as mais relevantes do que outras(AKHTAR et al., 2015).

Foram também reportadas dificuldades na utilização de dispositivos hápticos comerciais(REILINK et al., 2011; FIARD et al., 2014; VÅPENSTAD et al., 2013; CHELLALI et al., 2015), decorrentes das necessidades particulares de cada procedimento, como, por exemplo, necessidade de um alcance maior do braço robótico ou uma maior força gerada. Alguns autores na literatura também realizaram adaptações no dispositivo para fornecer uma maior ergonomia ao usuário, tal como a substituição da caneta háptica por uma seringa, em (CORRÊA et al., 2017), e uma extensão de acrílico para possibilitar melhor aderência durante uma simulação de palpação mamária em (RIBEIRO; NUNES; ELIAS, 2016).

3.4 Considerações Finais

Neste capítulo foram apresentados a metodologia, os resultados e a discussão de uma revisão sistemática, abordando as técnicas e metodologias empregadas na avaliação automática de desempenho em simuladores de RV com retorno háptico.

Foram levantadas diversas métricas de desempenho utilizadas na literatura para

medir o desempenho de aprendizes em simuladores de RV, as quais poderão ser aplicadas neste trabalho. Foi também observado que técnicas de ML ainda foram pouco utilizadas neste contexto de pesquisa. O artigo com os dados completos do estudo já foi aceito para publicação (SALLABERRY; TORI; NUNES, 2022).

O próximo capítulo apresentará os materiais que foram necessários para a realização deste projeto e os métodos que aplicados.

4 MATERIAIS E MÉTODOS

Neste capítulo serão apresentados os materiais e métodos que foram empregados na realização deste trabalho. A Seção 4.1 apresenta os materiais que foram utilizados durante o andamento do projeto e na Seção 4.2 são discutidos os métodos empregados.

4.1 Materiais

O estudo de caso desta pesquisa utilizou como base o simulador VIDA Odonto, descrito na Seção 2.5. Este simulador foi desenvolvido na plataforma de jogos *Unity* (versão 5.5.2f1) (Unity Technologies, 2020) devido à facilidade que ela proporciona para integração de dispositivos externos, tal como o HMD e dispositivo háptico. A sala dentária virtual e o paciente virtual foram modelados no software *Autodesk Maya* (Autodesk, 2022).

Nos experimentos, o simulador executava com taxa de a 90 quadros por segundo em um *Oculus Rift* CV1. Este HMD possibilita rastreamento da cabeça e campo de visão de 110 graus, com resolução de 1080 x 1200 *pixels* por olho. A configuração do computador utilizado consistiu em um processador *Intel* i5-9400F, 16 GB de RAM e uma placa gráfica *NVIDIA GeForce RTX* 2070.

O dispositivo háptico utilizado foi um *Geomatic Touch* (*3D Systems Corporation*),que possibilita 6 graus de liberdade de movimento e também pode fornecer *feedbacks* de força de até 3,3 Newtons. Um acessório criado com impressora 3D também foi modelado de modo a acoplar a empunhadura de uma seringa real com o dispositivo háptico, resultando em uma melhor ergonomia na manipulação do dispositivo pelo participante (Figura 13).

Durante experimentos realizados no simulador, descritos na Seção 5.2, foram coletados dados sobre o desempenho de especialistas (professores, dentistas e alunos

Figura 13 – Dispositivo háptico com seringa acoplada.

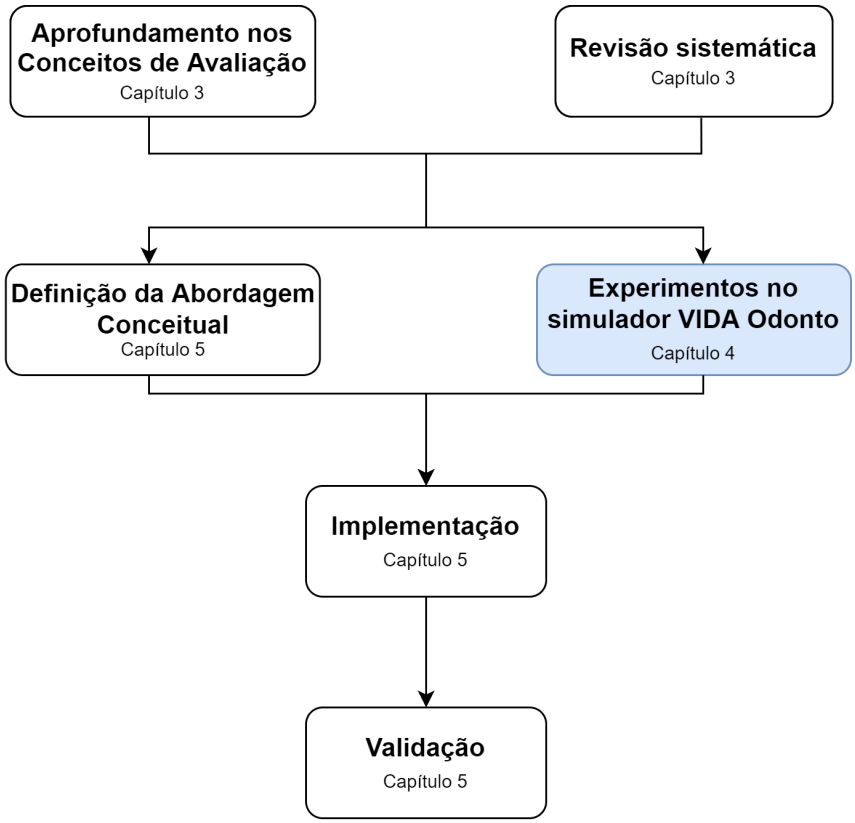
de pós-graduação), assim como de aprendizes (alunos considerados experientes e inexperientes, definido na Seção 5.2) no procedimento de anestesia para bloqueio do nervo alveolar inferior (COLLAÇO et al., 2020). As informações coletadas consistem na posição e orientação da seringa ao longo do procedimento, com frequência de 60 Hertz (Hz). O conjunto de dados também possui uma trajetória referência denominada “padrão ouro", a qual foi validada por uma especialista na área e utilizada para a explanação do procedimento durante o experimento. Em um trabalho anterior (COLLAÇO et al., 2020), os procedimentos realizados por participantes durante os experimentos foram reproduzidos no simulador para que pudessem ser avaliados como “Aprovado"ou “Reprovado"pela especialista. No presente trabalho, os dados coletados foram utilizados para treinar diferentes modelos de ML visando a discriminar participantes com base em suas habilidades. Também foi utilizada a classificação nas categorias “Aprovado"ou “Reprovado"realizada pela especialista (definido na Seção 5.2).

Foram utilizadas bibliotecas na linguagem *python*, como o *scikit-learn* para os classificadores NB, SVM e RF (PEDREGOSA et al., 2011), *Keras API* para o classificador MLP (CHOLLET et al., 2015) e *XGBoost Python Package* (DMLC, 2022) para o classificador XGB. Foi também utilizada a implementação do PCA da biblioteca *scikit-learn* para fusão de características. A seleção de características foi realizada com o método *ReliefF* da biblioteca *scikit-rebate* (URBANOWICZ et al., 2018) e o algoritmo genético foi implementado com a biblioteca *Sklearn-Genetic* (Manuel Calzolari, 2022). Para balanceamento do conjunto de dados, foi utilizada a técnica *Smote-Tomek links* da biblioteca *imbalanced-learn* (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017). A calibração com a técnica *Random Search* foi realizada com a biblioteca *scikit-learn*.

4.2 Métodos

A Figura 14 apresenta a visão geral das etapas do presente projeto, descritas nas subseções a seguir. A etapa “Experimentos no simulador VIDA Odonto”, em azul na figura, não foi realizada durante o período desse trabalho de mestrado.

Figura 14 – Diagrama das etapas cumpridas durante o desenvolvimento do projeto.



4.2.1 Revisão sistemática

Foi realizado um levantamento bibliográfico da literatura (SALLABERRY; TORI; NUNES, 2022), que já foi aceito para publicação na forma de artigo, o qual foi abordado no Capítulo 3. Esta revisão sistemática teve como objetivo identificar as principais técnicas e metodologias aplicadas para avaliação automática de aprendizes em simuladores de RV e RA com retorno háptico voltados para a saúde. Além das técnicas, foram identificados outros aspectos dos trabalhos analisados, tais como métricas de desempenho e métodos utilizados para avaliação automática de desempenho.

Esta etapa foi de grande importância no contexto deste trabalho, pois, a partir das informações coletadas e analisadas, foi possível definir o escopo deste trabalho. Tam-

bém foram definidas métricas de desempenho utilizadas na literatura e que poderiam ser aplicadas neste contexto.

4.2.2 Aprofundamento nos Conceitos de Avaliação

Foram estudados os conceitos envolvidos na avaliação de desempenho em simuladores médicos, sendo eles a aquisição de habilidades sensório-motoras, RV, aprendizado de máquina e avaliação em simuladores em RV, como descrito no Capítulo 2. Esta etapa teve grande importância na fundamentação deste trabalho, proporcionando uma base de conhecimento em áreas importantes para o desenvolvimento do modelo proposto.

Os estudos sobre a aquisição de habilidades motoras possibilitou a melhor compreensão de como ocorre o aprendizado de técnicas aplicadas em procedimentos médicos por aprendizes, indicando pontos que facilitariam a distinção entre participantes sem experiência e participantes que já dominam a técnica.

O estudo de RV buscou esclarecer as particularidades da tecnologia de RV que a difere de outros tipos de mídia e métodos de simulação, para que uma abordagem mais específica para este tipo de tecnologia pudesse ser desenvolvida. Também foram estudados os conceitos de avaliação em simuladores em RV, quais são as abordagens mais utilizadas e quais caminhos poderiam ser percorridos. Esta etapa deixou clara as vantagens que uma avaliação automática poderia oferecer para um simulador de treinamento.

Por fim, com base nos resultados obtidos durante a revisão sistemática da literatura, descrita no Capítulo 3, foram estudados diversos conceitos de ML, especialmente aqueles voltados para algoritmos de classificação, seleção e fusão de características.

4.2.3 Experimentos no simulador VIDA Odonto

Este trabalho utilizou dados coletados pelo simulador VIDA Odonto, descrito no Capítulo 2. Os experimentos realizados não fizeram parte deste trabalho de mestrado, porém tiveram participação do aluno enquanto estudante de iniciação científica.

Os experimentos foram realizados com alunos, profissionais e professores de Odontologia, e estão descritos no Capítulo 5.

Os dados obtidos nesses experimentos foram utilizados para o treinamento e vali-

dação dos modelos de avaliação de desempenho em simuladores médicos.

4.2.4 Definição da Abordagem Conceitual

A partir dos conceitos estudados e dos resultados da revisão sistemática, foi definida a abordagem proposta para um modelo de avaliação automática de desempenho para simuladores médicos em RV.

A abordagem busca utilizar dados coletados anteriormente no simulador VIDA Odonto de participantes experientes e inexperientes, com o intuito de diferenciar esses participantes. Também foi realizado um segundo modelo em que uma especialista no procedimento rotulou cada uma das trajetórias com uma nota. O modelo então classificou os participantes com base na avaliação dessa especialista. A abordagem e a arquitetura do modelo estão descritas no Capítulo 5.

4.2.5 Implementação

A implementação teve o objetivo de transformar a abordagem conceitual em módulos para que fosse possível ter um modelo executável que pudesse ser utilizado para realização do estudo de caso, com dados coletados no simulador VIDA Odonto. No total, foram definidos quatro módulos para a implementação (descritos na Seção 5.3): 1- Coleta de dados; 2- Extração e seleção das características; 3- Composição do modelo; 4- Avaliação do aprendiz.

Durante a implementação foram extraídas diversas características do conjunto de dados, definidas com base em outros estudos encontrados na revisão sistemática e nas observações realizadas por uma especialista no procedimento.

Foram aplicadas diferentes técnicas de ML para seleção e fusão de características, com o intuito de se comparar qual deles obteria os melhores resultados para o conjunto de dados. Foram também treinados diferentes algoritmos de classificação, tanto para diferenciar participantes com base em sua experiência na área, como para classificar os participantes com base na avaliação da especialista.

As combinações de algoritmo de seleção/fusão de características com os algoritmos de classificação foram calibradas, a fim de identificar as configurações que obtiveram o melhor resultado.

4.2.6 Validação do modelo

Inicialmente planejou-se realizar a validação do modelo a partir da coleta de dados de novos alunos e especialistas. Devido ao isolamento social imposto pela pandemia da COVID-19 (GUO et al., 2020), a realização de novos experimentos dentro do prazo previsto para esta pesquisa se tornou inviável. Por esse motivo, parte do conjunto de dados foi separado como conjunto de teste para validar o modelo final.

Utilizando o conjunto de testes, os algoritmos foram aplicados novamente com as configurações que obtiveram os melhores resultados na etapa anterior, e foram comparadas entre si.

Para fins de comparação, foram utilizadas as métricas acurácia, sensibilidade, especificidade, precisão, acurácia balanceada e *F1-score*, descritas no Capítulo 2. A calibração foi realizada com o método *Random Search*, descrito na Seção 2.3.3. A estimativa de erro foi conduzida com a técnica de validação cruzada estratificada K *folds*, sendo sua implementação descrita na Seção 5.3.7.

4.3 Considerações Finais

Neste capítulo foram apresentados os materiais e os métodos empregados neste projeto. No capítulo seguinte, será apresentado o modelo conceitual definido, assim como a descrição dos experimentos realizados e a implementação do modelo com os dados coletados pelo simulador VIDA Odonto, como estudo de caso.

5 MODELO PARA AVALIAÇÃO AUTOMÁTICA DE DESEMPENHO DE APRENDIZES EM SIMULADORES DE RV COM DISPOSITIVO HÁPTICO

Neste capítulo será apresentado o modelo para avaliação automática de desempenho de aprendizes em simuladores de RV com dispositivo háptico proposto neste trabalho. A Seção 5.3 define o modelo conceitual e detalha os módulos que compõem a sua arquitetura, a Seção 5.2 descreve experimentos anteriormente realizados com o simulador VIDA Odonto, que possibilitaram a coleta de dados de participantes. As seções seguintes detalham cada etapa da criação do modelo, desde a extração de características até a validação do mesmo.

5.1 Modelo conceitual

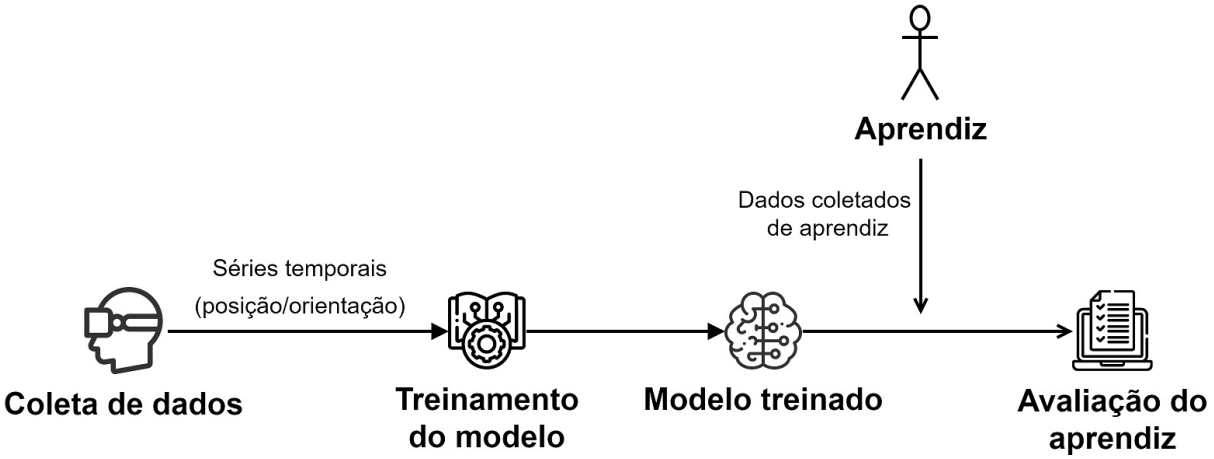
O modelo conceitual, apresentado na Figura 15, foi definido com a proposta de avaliar o desempenho de aprendizes com base nos dados coletados durante a utilização de simuladores médicos em RV com dispositivo háptico.

O modelo utiliza dados de treinamento de usuários com diferentes níveis de experiência no procedimento médico em questão, coletados pelo simulador. Os dados são coletados na forma de séries temporais, considerando a posição e a orientação da seringa a cada instante de tempo, durante a execução do procedimento simulado.

Os dados de treinamento (séries temporais) são utilizados para treinar um modelo de avaliação de desempenho utilizando algoritmos de aprendizado de máquina com o objetivo de encontrar padrões capazes de diferenciar usuários de classes distintas. Assim, o modelo deve ser capaz de realizar a classificação de novos usuários na classe apropriada, ao utilizarem o simulador.

Uma vez que o modelo foi treinado, um novo aprendiz que utiliza o simulador

Figura 15 – Modelo conceitual utilizado para avaliação do aprendiz.



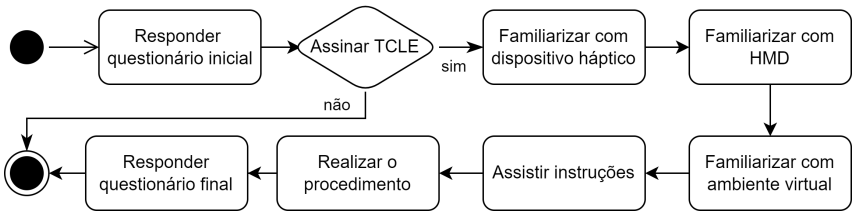
pode ter seus dados coletados durante o procedimento, processados e fornecidos ao modelo. O modelo realiza a avaliação de desempenho classificando se o desempenho do aprendiz foi mais próximo de usuários experientes ou de usuários inexperientes.

5.2 Experimentos

Foram realizados dois experimentos no simulador VIDA Odonto. O primeiro experimento ocorreu na Faculdade de Odontologia de Bauru da USP (FOB), em novembro de 2018, e teve como participantes 16 especialistas (alunos de pós-graduação, professores e dentistas), 28 alunos experientes (já haviam aprendido e realizado o procedimento em questão) e 42 alunos inexperientes no procedimento de anestesia do nervo alveolar inferior. O segundo experimento ocorreu na Faculdade de Odontologia da USP (FOUSP), em maio de 2019, e teve como participantes 159 alunos experientes (já haviam realizado a disciplina de anestesia e realizado o procedimento real) e 12 alunos inexperientes (experimento detalhado em (COLLAÇO et al., 2020)). Porém, como o experimento fez parte de um projeto mais abrangente, foram apenas considerados os participantes que realizaram o técnica de Bloqueio do Nervo Alveolar Inferior (BNAI) no simulador de forma totalmente imersiva (utilizando o HMD e dispositivo háptico), para um total de 155 participantes. A execução de experimentos está relacionada ao módulo de coleta de dados do modelo proposto (Figura 18) e as atividades conduzidas pelos participantes durante os dois experimentos (Figura 16) estão descritas a seguir.

Ambos os experimentos foram iniciados com a assinatura de um termo de con-

Figura 16 – Atividades conduzidas no experimento



sentimento livre e esclarecido, aprovado pelo Comitê de Ética em Pesquisa em Seres Humanos da CEPH-IPUSP (CAAE - 79294617.9.0000.5561), seguido pelo preenchimento de um pré-questionário, abordando o grau de familiaridade do participante com o procedimento e o seu nível de experiência na área de Odontologia. Em seguida, o participante executava três etapas, durante as quais deveria utilizar a seringa virtual para perfurar alvos predeterminados, sendo que um novo mecanismo do simulador era introduzido a cada uma delas. A primeira etapa consistia em utilizar a seringa háptica e o monitor bidimensional (2D) de 40 polegadas para perfurar quatro bolinhas no ambiente virtual (Figura 17a). Na segunda etapa, as quatro bolinhas deveriam ser perfuradas novamente, porém o participante vestia um HMD para realizar o procedimento, ao invés de utilizar o monitor 2D (Figura 17b). Na terceira etapa, o participante era colocado dentro do consultório virtual, ao lado do paciente, e deveria acertar dois pontos diferentes dentro de sua boca. Essas etapas foram criadas para proporcionar uma familiarização gradativa com o simulador e com as tecnologias empregadas.

Na etapa seguinte, o participante assistia, ainda dentro do consultório virtual, a uma explicação do procedimento de anestesia do nervo alveolar inferior gravado por uma professora (as marcações visuais utilizadas durante esta etapa foram apresentadas na Figura 8 da Seção 2.5). Na sequência, era solicitado ao participante que realizasse o procedimento completo no paciente virtual (Figura 7 da Seção 2.5). Antes de realizar o procedimento de anestesia do nervo alveolar inferior, o participante deveria se posicionar de forma adequada em relação ao paciente virtual, segurar a seringa háptica e se declarar pronto para iniciar o procedimento. A partir deste momento, o simulador iniciava a coleta da posição e orientação da seringa virtual e do HMD, a cada instante de tempo (60 Hz). O simulador era capaz de identificar o momento no qual o participante acertava a região próxima ao ponto de anestesia e a coleta terminava quando a seringa era removida de dentro da mucosa da boca do paciente.

Como último procedimento, o participante deveria preencher um pós-questionário referente às impressões obtidas em relação ao simulador. Todos os áudios com instru-

ções de cada etapa e com a explicação do procedimento foram previamente gravados, de forma que houvesse interação mínima entre os pesquisadores e o participante.

Figura 17 – a) Primeira etapa do experimento: o participante deveria perfurar cada uma das quatro bolinhas em ordem numérica crescente utilizando um monitor 2D para tal. b) Segunda etapa do experimento: as quatro bolinhas deveriam ser perfuradas novamente, porém o participante utilizava um HMD.

(a)

(b)

Dois modelos foram construídos neste estudo. Para construir o primeiro modelo, foram definidos os rótulos dos participantes do conjunto de dados como novatos e experientes, em que novatos eram estudantes de Odontologia que não sabiam como realizar a técnica BNAI e os experientes eram estudantes de graduação que já sabiam como realizar o procedimento, assim como estudantes de pós-graduação, professores e dentistas profissionais (todos já sabiam como realizar o procedimento de BNAI). Esta rotulação foi chamada de EN (Experientes e Novatos).

Para o segundo modelo, um instrutor em Odontologia avaliou cada trajetória e

atribuiu pontuação a três quesitos: distância até o alvo ideal de inserção, angulação durante a inserção e trajetória da seringa. A pontuação atribuída foi de 1 (reprovada) ou 2 (aprovada). Para cada participante, as três pontuações foram somadas e reescalonadas para um intervalo de zero a 10, considerada a pontuação final do participante (o intervalo foi escolhido apenas como uma forma de facilitar a interpretação da pontuação final atribuída a cada trajetória, uma vez que este intervalo de notas é adotado em muitos ambientes de ensino). Os participantes com uma pontuação final de 5 ou mais foram rotulados como aprovados, enquanto os participantes com uma pontuação final inferior a 5 foram rotulados como não aprovados. Esta rotulação foi chamada de Avaliação pelo Especialista (AE).

5.3 Implementação

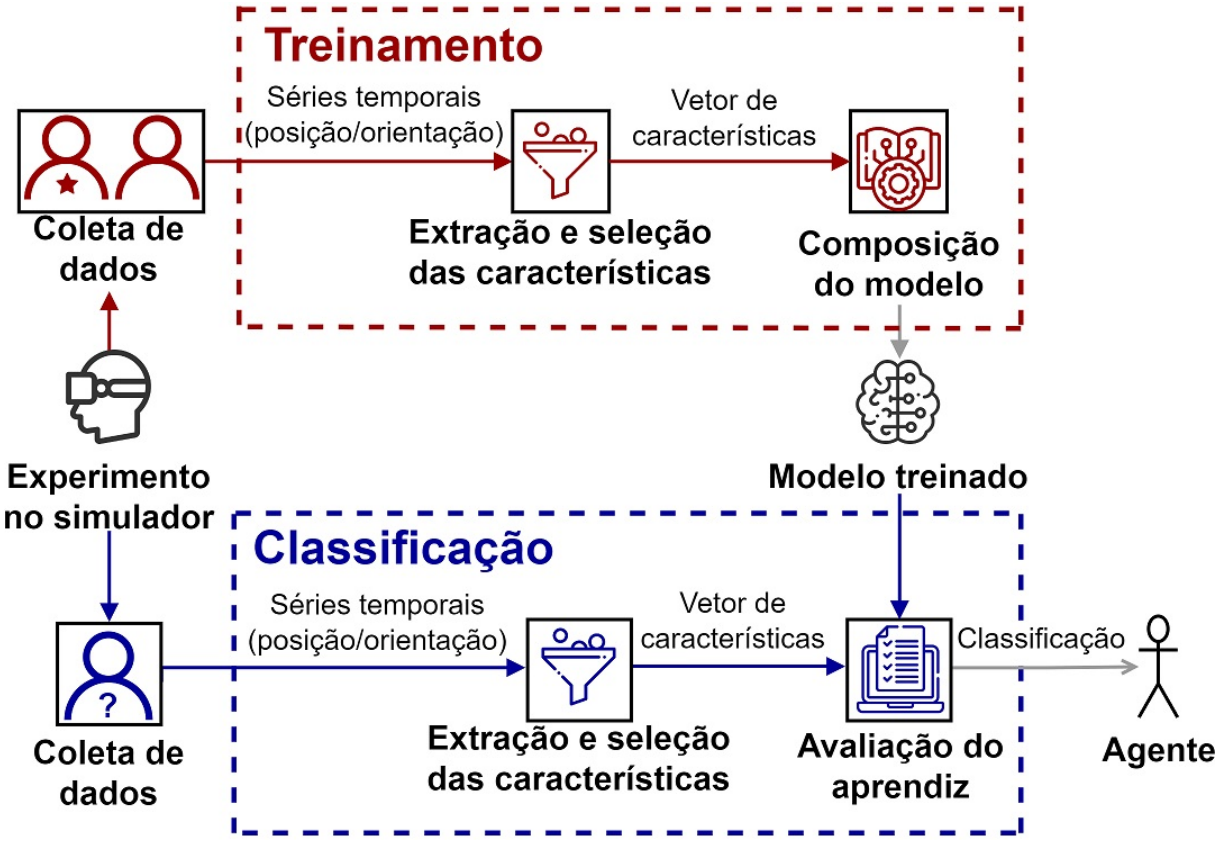
A Figura 18 ilustra a arquitetura do modelo desenvolvido, que consiste de duas etapas. A **etapa de Treinamento** engloba desde a coleta de dados de participantes de diferentes níveis de experiência, até a manipulação e a utilização desses dados para treinar um algoritmo de ML capaz de classificar aprendizes. A **etapa de Classificação** consiste na utilização do modelo já treinado para avaliar estudantes durante o aprendizado do procedimento odontológico.

Dentro das duas etapas que constituem a arquitetura, são definidos quatro módulos: (a) **Módulo de coleta de dados**, que realiza a coleta das informações sobre o desempenho dos participantes; (b) **Módulo de extração e seleção das características**, que gera o vetor de características que alimenta o algoritmo; (c) **Módulo de composição do modelo**, que treina o algoritmo de ML com os dados de participantes de vários níveis de experiência e (d) **Módulo de avaliação do aprendiz**, que utiliza o algoritmo para classificar um aprendiz com base no seu desempenho. Cada um dos módulos está detalhado a seguir.

5.3.1 Módulo de coleta de dados

O módulo de coleta de dados é responsável por adquirir e armazenar os dados referentes às ações executadas pelo participante, durante o procedimento. Atualmente, o simulador VIDA Odonto é capaz de coletar dados sobre o tempo transcorrido no procedimento, a posição e a orientação da seringa virtual, em uma frequência de 60 Hz. Em experimentos realizados anteriormente, os dados foram coletados a partir do

Figura 18 – Arquitetura do modelo desenvolvido.



momento em que o participante se declarava pronto para iniciar o procedimento. O participante deveria penetrar a região alvo com a agulha da seringa e, ao retirá-la de dentro da mucosa da boca, a coleta era finalizada.

Os dados armazenados pelo simulador podem ser acessados por um professor que queira visualizar o procedimento realizado por um dos aprendizes, de forma imersiva, na qual os movimentos que o aluno realizou com a seringa são reproduzidos no ambiente virtual, enquanto o professor utiliza o HMD para visualizar o procedimento, como se estivesse dentro do consultório. A reprodução também pode ser apresentada de forma não imersiva, em um monitor 2D, de forma que a visão representada corresponde à visão do aluno durante o procedimento. Os dados coletados podem também ser exportados para serem lidos e analisados por sistemas computacionais externos.

Para o estudo de caso, na etapa de Treinamento, o módulo de coleta de dados foi utilizado durante experimentos para coletar os dados de participantes de diferentes níveis de experiência no procedimento de anestesia do nervo alveolar inferior, enquanto realizavam este procedimento no simulador VIDA Odonto (descrito na Seção 5.2). Já na etapa de Classificação, o módulo é utilizado para coletar os dados de novos aprendizes enquanto usam o simulador.

5.3.2 Etapa de Treinamento

A etapa de Treinamento é composta pelo módulo de extração e seleção das características e pelo módulo de composição do modelo. Nesta etapa, são utilizados dados coletados de usuários de diferentes níveis de experiência no procedimento simulado para treinar um modelo de ML capaz de avaliar aprendizes.

5.3.2.1 Módulo de extração e seleção das características

Um levantamento bibliográfico (Capítulo 3) mostrou métricas de desempenho encontradas na literatura e que puderam ser utilizadas como características do conjunto de dados.

Durante a etapa de Treinamento, o módulo é utilizado para realizar a extração, fusão e/ou seleção de características utilizando os algoritmos PCA, *ReliefF* e o GA (descritos na Seção 2.3). Este procedimento é muito importante para reduzir a dimensionalidade do conjunto de dados, uma vez que utilizar um grande número de características pode comprometer o poder de classificação de um algoritmo, especialmente quando existem poucas instâncias de treinamento.

Este módulo fornece, como saída, um vetor de características utilizado como entrada para os algoritmos de ML.

5.3.2.2 Módulo de composição do modelo

Na etapa de Treinamento, a partir do vetor de características obtido no módulo anterior, são treinados algoritmos de ML para avaliar aprendizes com base nos seus desempenhos no simulador, utilizando dados objetivos coletados durante a utilização do sistema. São treinados algoritmos clássicos para tal, como o NB, SVM, RF, MLP e XGB (apresentados na Seção 2.3), escolhidos por serem algoritmos bastante conhecidos e estudados na literatura, tendo mostrado bons resultados para diferentes situações.

Para realizar a calibração do modelo, a técnica de validação cruzada *K-Fold* estratificada (Seção 2.3) é utilizada, com $K = 10$ (12 a 13 instâncias por *fold*), de forma que a melhor configuração de hiper parâmetros dos classificadores é considerada como aquela que apresentou o melhor resultado, em média, entre os *folds* de calibração.

5.3.3 Etapa de Classificação

A etapa de Classificação é composta pelo módulo de extração e seleção das características e pelo módulo de avaliação do aprendiz. Esta etapa utiliza os dados coletados por um aprendiz utilizando o simulador e, utilizando o modelo de ML já treinado, avalia este aprendiz com base em seu desempenho.

Durante a etapa de Classificação, o processo realizado pelo módulo de extração e seleção das características (descrito na Seção 5.3.2.1) é reproduzido, de forma que as mesmas manipulações de características realizadas durante a etapa de Treinamento são repetidas.

Assim como na etapa de Treinamento, este módulo fornece, como saída, um vetor de características utilizado como entrada para os algoritmos de ML.

5.3.3.1 Módulo de avaliação do aprendiz

Após treinar os algoritmos, os dados separados para validação são utilizados para estimar o erro do modelo. São utilizadas métricas de erro clássicas, tal como acurácia, sensibilidade e precisão (Seção 2.3).

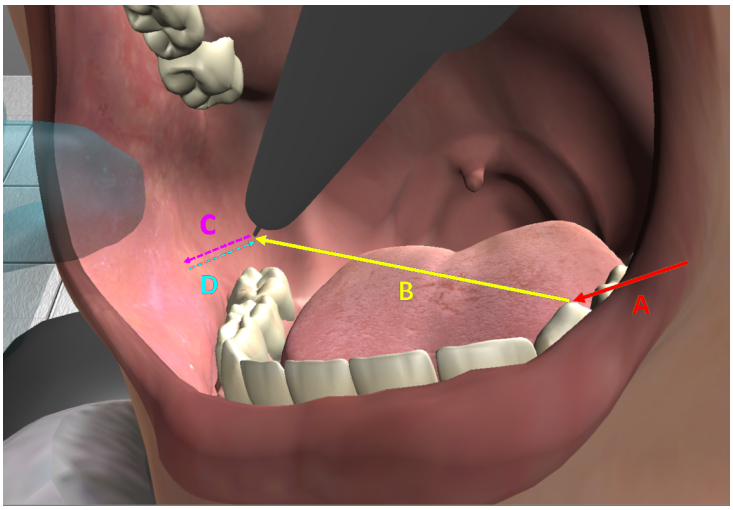
Para o estudo de caso, o algoritmo que apresentou o melhor resultado foi apresentado no Capítulo 6 e o modelo poderá ser futuramente implementado no simulador VIDA Odonto. Dessa forma, uma vez que novas coletas de dados forem realizadas em experimentos ou na sala de aula, o simulador seria capaz de gerar, em tempo real, a avaliação do desempenho dos aprendizes.

5.3.4 Extração de características

Utilizando os dados de ambos os experimentos, foram calculadas características com base nas trajetórias das seringas dos participantes, enquanto realizavam o procedimento no simulador. O cálculo de características a partir dos dados coletado está associado ao módulo de extração e seleção de características da arquitetura proposta (Figura 18). Os dados foram separados entre dados de treinamento e dados de teste, em uma proporção de 80% (124 instâncias) e 20% (31 instâncias), respectivamente. Todas as etapas de pré processamento dos dados, extração, seleção e fusão de características aplicadas ao conjunto de treinamento foram replicadas no conjunto de teste.

A trajetória completa do experimento foi dividida em 4 segmentos (Figura 19). O segmento A refere-se ao início do procedimento, momento no qual a agulha da seringa se encontra fora da boca do paciente virtual. O segmento B remete ao trajeto da agulha da seringa dentro da boca do paciente virtual, do momento que esta entra na boca, até o momento em que toca a mucosa na região de aplicação da anestesia. O segmento C é definido pela parte da trajetória na qual a agulha da seringa penetra na mucosa da boca e D a trajetória da agulha sendo retirada da mucosa. Um segmento E foi também definido como sendo a combinação de todos os outros quatro segmentos, ou seja, a trajetória completa da seringa.

Figura 19 – Segmentação da trajetória dos participantes.



Durante a revisão sistemática apresentada no Capítulo 3, foram levantadas as métricas de desempenho presentes na literatura que foram empregadas na avaliação de aprendizes em simuladores de RV ou RA. As características calculadas foram definidas com base nessas métricas, além de outras métricas que foram consideradas relevantes ou específicas para o procedimento de anestesia para bloqueio do nervo alveolar inferior e identificadas por uma instrutora.

Para calcular as características, foi primeiramente calculada a velocidade para cada instante da trajetória. Para isso, foi utilizada a Equação 5.1, onde $Vi(n)$ é a velocidade no instante de índice n, $D(n+k, n-k)$ é a distância entre os pontos nos instantes de índices $n+k$ e $n-k$ e $T(n+k)$ é o tempo decorrido nos instantes de índice $n+k$, desde o início da trajetória. Em um instante n da trajetória, esta equação calcula a velocidade média do intervalo $n+k$ a $n-k$. A mesma abordagem foi utilizada para calcular a aceleração, o *jerk* (derivada da aceleração que identifica movimentos bruscos) e a velocidade de rotação para cada instante da trajetória.

$$V(n) = \frac{D(n+k, n-k)}{T(n+k) - T(n-k)} \tag{5.1}$$

No total, foram calculadas 98 características para o conjunto de dados relacionadas ao Tempo, Movimento e Precisão do participante durante o procedimento. As características são descritas a seguir.

Características por segmento: foram calculadas 19 características para cada um dos cinco segmentos de trajetórias definidos. Foram elas:

1. **Distância percorrida:** o simulador coleta 60 pontos por segundo, correspondentes à posição da seringa virtual no espaço tridimensional, enquanto o participante realiza o procedimento. As distâncias euclidianas entre cada par de pontos consecutivos foram calculadas e somadas para encontrar a distância total (D_t) percorrida durante cada segmento, conforme a Equação 5.2, em que P_k^t representa a coordenada k de um instante t da trajetória e n_{max} o número total de pontos do segmento.

$$D_t = \sum_{n=1}^{n_{max}} (\sqrt{(P_x^n - P_x^{(n-1)})^2 + (P_y^n - P_y^{(n-1)})^2 + (P_z^n - P_z^{(n-1)})^2}) \tag{5.2}$$

2. **Ângulo acumulado:** O ângulo acumulado AN_t decorrente da soma dos ângulos AN_n formados pela rotação da seringa entre cada par de instantes consecutivos da trajetória do segmento, descrito na Equação 5.3.

$$AN_t = \sum_{n=1}^{n_{max}} AN_n \tag{5.3}$$

3. **Tempo decorrido:** o tempo decorrido ΔT, do momento em que a seringa virtual inicia a trajetória do respectivo segmento (T_o), até a sua finalização (T_f), como apresentado na Equação 5.4.

$$\Delta T = T_f - T_o \tag{5.4}$$

4. **Velocidade média:** a velocidade média V_m da seringa é calculada somando-se todas as velocidades V_n a cada instante do segmento, e dividindo-as pelo número de pontos n, conforme a Equação 5.5.

$$V_m = \frac{\sum_{n=1}^{n_{max}} V_n}{n} \tag{5.5}$$

5. **Desvio padrão da velocidade:** o desvio padrão $\sigma(V)$ calculado a partir das velocidades V_n a cada instante do segmento, como mostrado na Equação 5.6.

$$\sigma(V) = \sqrt{\frac{\sum_{n=1}^{n_{max}} (V_n - V_m)^2}{n_{max} - 1}} \tag{5.6}$$

6. **Velocidade máxima:** a velocidade máxima V_{max} encontrada entre todas as velocidades $V(n)$ a cada instante do segmento, descrita na Equação 5.7.

$$V_{max} = \max\{V(1), \ldots, V(n_{max})\} \tag{5.7}$$

7. **Aceleração média:** a aceleração média A_m da seringa é calculada somando-se todas as acelerações A_n a cada instante do segmento, e dividindo-as pelo número de pontos n, conforme a Equação 5.8.

$$A_m = \frac{\sum_{n=1}^{n_{max}} A_n}{n} \tag{5.8}$$

8. **Desvio padrão da aceleração:** o desvio padrão $\sigma(A)$ calculado a partir das acelerações A_n a cada instante do segmento, conforme a Equação 5.9.

$$\sigma(A) = \sqrt{\frac{\sum_{n=1}^{n_{max}} (A_n - A_m)^2}{n_{max} - 1}} \tag{5.9}$$

9. **Aceleração máxima:** a aceleração máxima A_{max} encontrada entre todas as acelerações $A(n)$ a cada instante do segmento, como descrita na Equação 5.10.

$$A_{max} = \max\{A(1), \ldots, A(n_{max})\} \tag{5.10}$$

10. ***Jerk* médio:** o *jerk* médio J_m da seringa é calculado somando-se todos os *jerks* Ji_n a cada instante do segmento, e dividindo-os pelo número de pontos n, conforme a Equação 5.11.

$$J_m = \frac{\sum_{n=1}^{n_{max}} J_n}{n} \tag{5.11}$$

11. **Desvio padrão do *jerk*:** o desvio padrão $\sigma(J)$ calculado a partir dos *jerks* J_n a cada instante do segmento, descrito na Equação 5.12.

$$\sigma(J) = \sqrt{\frac{\sum_{n=1}^{n_{max}} (J_n - J_m)^2}{n_{max} - 1}} \tag{5.12}$$

12. ***Jerk* máximo:** o *jerk* máximo J_{max} encontrado entre todos os *jerks* $J(n)$ a cada instante do segmento, como mostrado na Equação 5.13.

$$J_{max} = \max\{J(1), \ldots, J(n_{max})\} \tag{5.13}$$

13. **Velocidade de rotação média:** a velocidade de rotação média VR_m da seringa é calculada somando-se todas as velocidades de rotação VR_n a cada instante do segmento, e dividindo-as pelo número de pontos n, conforme a Equação 5.14.

$$VR_m = \frac{\sum_{n=1}^{n_{max}} VR_n}{n} \tag{5.14}$$

14. **Desvio padrão da velocidade de rotação:** o desvio padrão $\sigma(VR)$ calculado a partir das velocidades de rotação VR_n a cada instante do segmento, descrito na Equação 5.15.

$$\sigma(VR) = \sqrt{\frac{\sum_{n=1}^{n_{max}} (VR_n - VR_m)^2}{n_{max} - 1}} \tag{5.15}$$

15. **Velocidade de rotação máxima:** a velocidade de rotação máxima VR_{max} encontrada entre todas as velocidades de rotação $VR(n)$ a cada instante do segmento, como mostrada na Equação 5.16.

$$VR_{max} = \max\{VR(1), \ldots, VR(n_{max})\} \tag{5.16}$$

16. **Firmeza:** representa a oscilação da seringa durante a execução do procedimento. A firmeza F é calculada pelo desvio padrão S do ângulo formado, a cada instante do segmento, entre a seringa e uma orientação pré definida, e dividindo este valor pela média $\overline{X}$ dos valores deste mesmo ângulo. Esta característica está descrita na Equação 5.17, retirada de (ANJOS, 2014).

$$F = 1 - \frac{S}{\overline{X}} \tag{5.17}$$

17. **Velocidade acima do limiar:** Uma trajetória realizada e cuidadosamente analisada por uma instrutora foi considerada como o “padrão ouro” do procedimento. O limiar superior foi definido como o 75º percentil entre as velocidades a cada instante do segmento da trajetória “padrão ouro”.

Esta característica OL foi calculada como a quantidade de instantes n_o no segmento de trajetória do participante em que a velocidade ficou acima do limiar, dividida pelo número total de observações n naquele segmento, apresentada na Equação 5.18.

$$OL = \frac{n_o}{n} \tag{5.18}$$

18. **Velocidade abaixo do limiar:** de forma similar à "Velocidade acima de um limiar", o limiar inferior foi definido como o 25º percentil entre as velocidades a cada instante do segmento da trajetória "padrão ouro".

    Esta característica UL foi calculada como a quantidade de instantes n_u no segmento de trajetória do participante em que a velocidade ficou abaixo do limiar, dividida pelo número total de observações n naquele segmento, conforme a Equação 5.19.

$$UL = \frac{n_u}{n} \tag{5.19}$$

19. ***Distance Time Warping* (DTW):** o DTW é uma métrica que compara séries temporais de diferentes tamanhos e velocidades. O DTW utiliza a distância euclidiana para parear os pontos de duas trajetórias, no caso, de um participante e o "padrão ouro", com o objetivo de encontrar similaridades na forma das duas séries temporais. A descrição completa da implementação do DTW pode ser encontrada em (GIORGINO, 2009).

Características globais: seis características foram calculadas apenas uma vez para cada trajetória, não sendo recalculadas individualmente para cada segmento. As características globais foram:

1. **Profundidade máxima de penetração da agulha:** a profundidade máxima da agulha é fornecida automaticamente pelo simulador. O valor gerado pertence ao intervalo [0, 1], no qual 1 significa que a agulha inteira foi introduzida na mucosa. Por conta de problemas na coleta de dados, a profundidade máxima da agulha teve valores ausentes em uma pequena quantidade de participantes. Para tais instâncias, foi realizada a imputação pela média dos valores de todas as instâncias do conjunto de dados.

2. **Ângulo de penetração da agulha:** utilizando como base a trajetória considerada "padrão ouro", foi possível identificar o ponto e o ângulo esperados de

entrada da agulha na mucosa. O ângulo entre a seringa do participante e a orientação ideal determinada pela trajetória de referência foi então determinado.

3. **Distância da agulha ao alvo:** utilizando a mesma lógica da característica anterior, calculou-se a distância euclidiana D_a entre o ponto no qual a agulha da seringa do participante penetra na mucosa do paciente e o local ideal para aplicação. Na Equação 5.20, P^p representa o ponto no qual a seringa do participante atinge a região de anestesia, enquanto P^o representa o ponto de anestesia definido pelo "padrão ouro".

$$D_a = \sqrt{(P_x^p - P_x^o)^2 + (P_y^p - P_y^o)^2 + (P_z^p - P_z^o)^2} \tag{5.20}$$

4. **Eficiência da trajetória completa:** A Eficiência é uma métrica que busca associar a precisão do participante no procedimento, com o tempo levado para executar a técnica (EVE et al., 2014). A Eficiência da trajetória completa E foi calculada como a razão entre as características "Distância da agulha ao alvo" e "Tempo decorrido", como descrita na Equação 5.21.

$$E = \frac{D_a}{\Delta T} \tag{5.21}$$

5. **Eficiência no segmento B:** A eficiência no segmento B E_b é similar à característica "Eficiência da trajetória completa", porém o "Tempo decorrido" é apenas considerado para o segmento B, conforme a Equação 5.22.

$$E_b = \frac{D_a}{\Delta T_b} \tag{5.22}$$

6. **Eficiência no segmentos A e B:** A eficiência no segmento B E_{ab} é similar à característica "Eficiência da trajetória completa", porém o "Tempo decorrido" é apenas considerado para os segmentos A e B, como mostrada na Equação 5.23.

$$E_{ab} = \frac{D_a}{\Delta T_a + \Delta T_b} \tag{5.23}$$

Como o quarto segmento do "padrão ouro" era muito curto (Segmento D na figura 19), as características "Velocidade acima/abaixo do limiar" não foram calculadas para este segmento. A métrica DTW entre séries temporais também não foi calculada para este segmento da trajetória do "padrão ouro".

5.3.5 Fusão e seleção de características

A importância da redução da dimensionabilidade é tornar a previsão do algoritmo mais confiável, uma vez que conjuntos de dados com um elevado número de características podem ser muito esparsos (GÉRON, 2019).

A partir dos conjuntos de dados originais, foram criados mais três conjuntos de dados, cada um usando um método de fusão ou seleção de características, a fim de criar conjuntos de dados com dimensão reduzida. A tabela 3 mostra como cada conjunto de dados foi definido.

O conjunto de dados 2 foi criado usando o algoritmo de fusão de características PCA. O terceiro conjunto de dados utilizou o algoritmo de seleção de características *ReliefF* para definir as características. O quarto conjunto de dados foi criado utilizando o GA para a seleção de características. Como a função de aptidão utilizada neste estudo depende de cada classificador, o quarto conjunto de dados apresentou diferente características dependendo do classificador utilizado.

Esta abordagem foi utilizada tanto para o conjunto de dados com rotulação EN quanto para o conjunto de dados com rotulação AE, para um total de 6 novos conjuntos de dados.

5.3.6 Balanceamento dos conjuntos de dados

Como descrito na Seção 5.2, cada uma das instâncias do conjunto de dados foi rotulada com base no nível de experiência do participante, como “Novato” ou “Experiente”, assim também como pela avaliação realizada por uma especialista no procedimento, definindo as instâncias como “Aprovado” ou “Reprovado”.

Os conjuntos de dados com a rotulação EN apresentaram 71% de instâncias rotuladas como Especialistas (classe negativa) e 29% como Novatos (classe negativa). Os conjuntos de dados com as rotulação AE apresentaram 52% de instâncias rotuladas como Aprovadas (classe positiva) e 48% como Reprovadas (classe negativa).

Para os conjuntos de dados com a rótulo EN, foi necessário balancear as instâncias para evitar *overfit* do modelo na classe majoritária (caso em que o modelo se especializa na identificação de instâncias do conjunto de treinamento e se torna incapaz de realizar a generalização para identificar novas instâncias). Para isso, foi utilizada a técnica *SMOTE-Tomek links* antes de treinar os classificadores. Este al-

Tabela 3 – Tabela com a definição de cada conjunto de dados

Nome	Rótulo	Definição	Número de características
Conjunto de dados 1	EN	Características originais	98
Conjunto de dados 2	EN	Fusão com PCA	14
Conjunto de dados 3	EN	Seleção com *ReliefF*	8
Conjunto de dados 4	EN	Seleção com GA	RF: 12 NB: 6 MLP: 12 SVM: 13 XGB: 10
Conjunto de dados 5	AE	Características originais	98
Conjunto de dados 6	AE	Fusão com PCA	14
Conjunto de dados 7	AE	Seleção com *ReliefF*	7
Conjunto de dados 8	AE	Seleção com GA	RF: 9 NB: 13 MLP: 11 SVM: 10 XGB: 10

goritmo combina a técnica de *oversampling SMOTE* (CHAWLA et al., 2002), na qual são criadas instâncias sintéticas da classe minoritária, com a técnica de *undersampling Tomek links*, que reduz o ruído criado pelo *SMOTE*, retirando instâncias que têm como vizinho mais próximo uma instância de uma classe diferente (BATISTA et al., 2003).

Nenhuma técnica para balancear os conjuntos de dados com os rótulos AE foi utilizada, uma vez que os conjuntos de dados já estavam balanceados.

5.3.7 Classificação

Foram implementados algoritmos clássicos de ML para classificação considerados apropriados para o problema, são eles o NB, RF, SVM, MLP e XGB, descritos na Seção 2.3. Todos os algoritmos são conhecidos e utilizados na literatura, porém utilizam abordagens diferentes para realizar a classificação.

Cada algoritmo de classificação foi aplicado usando cada conjunto de dados, tanto para os conjuntos de dados com rotulação EN como para os com rotulação AE. No total, foram obtidas 40 combinações de classificadores e conjuntos de dados.

Os conjuntos de dados foram divididos em grupo de treinamento e teste, em uma proporção de 80% e 20% (124 e 31 instâncias), respectivamente. O grupo de treinamento foi usado para calibrar os hiper parâmetros de cada um dos classificadores utilizando a técnica *Random Search*, por meio de uma validação cruzada estratificada de 10 *folds* e, para a melhor configuração de hiper parâmetros obtida na calibração, o classificador foi treinado novamente usando todo o grupo de treinamento, e testado no grupo de teste.

Para avaliar o desempenho das 40 combinações de classificadores e conjuntos de dados, foi calculada uma série de métricas de erro: acurácia, sensibilidade, especificidade, precisão, acurácia balanceada e *F1-Score*. Para conjuntos de dados balanceados, a acurácia é uma métrica amplamente utilizada, porém pode fornecer informações tendenciosas quando o conjunto de dados é desbalanceado. Dessa forma, a escolha da melhor configuração de hiper parâmetros foi realizada com base nas configurações que obtiveram a melhor acurácia média entre *folds* para dados rotulados como AE, ou melhor *F1-Score* médio, para casos rotulados como EN (conjuntos de dados desbalanceados).

5.3.8 Teste

Foi realizada uma nova etapa de treinamento para cada uma das 40 combinações de classificadores e conjuntos de dados. Nesta etapa, o modelo foi treinado utilizando a melhor configuração de hiper parâmetros encontrada durante a calibração, para cada uma das 40 combinações. O treinamento foi realizado utilizando todo o grupo de treinamento e o modelo treinado foi utilizado para classificar as instâncias do conjunto de teste. Esta etapa é importante para testar o modelo em dados ainda não vistos pelos algoritmos, simulando a situação de novos dados coletados pelo simulador.

5.4 Considerações Finais

Neste capítulo, foram apresentadas as etapas realizadas para criação do modelo de avaliação automática em simuladores médicos em RV. No próximo capítulo serão apresentados os resultados obtidos e a discussão do trabalho.

6 RESULTADOS E DISCUSSÃO

Neste estudo, foram testados diversos algoritmos para definir o modelo para avaliar o desempenho de simuladores médicos que usam dispositivos hápticos em RV. Como estudo de caso, os modelos foram construídos a partir de dados coletados em um simulador odontológico apresentado na Seção 2.5, considerando duas formas distintas de rotulação, conforme apresentado na Seção 5.2.

As Tabelas 4 e 5 apresentam as métricas de acurácia e *F1-Score* obtidas no grupo de teste para as configurações de hiper parâmetros que obtiveram os melhores valores de *F1-Score* médio de calibração, para conjuntos de dados rotulados como EN, ou melhor acurácia média de calibração, para conjuntos de dados rotulados como AE. Para cada Algoritmo, a combinação de Algoritmo/Conjunto de dados com a melhor acurácia é apresentada em negrito. Quando dois conjuntos de dados obtiveram a mesma acurácia para o mesmo algoritmo, o conjunto de dados selecionado como melhor foi o que obteve o maior *F1-Score*.

Tabela 4 – Acurácia/*F1-Score* obtidos para cada combinação de algoritmo/conjunto de dados para os conjuntos de dados rotulados como EN.

	RF	**NB**	**MLP**	**SVM**	**XGB**
Original (1)	0,58/0,32	0,32/0,46	**0,68/0,55**	**0,71/0,61**	0,42/0,18
PCA (2)	0,52/0,40	0,45/0,41	0,52/0,52	0,52/0,52	**0,52/0,40**
***ReliefF* (3)**	**0,58/0,38**	0,61/0,33	0,58/0,32	0,55/0,22	0,48/0,20
GA (4)	0,55/0,22	**0,65/0,27**	0,61/0,50	0,61/0,50	0,48/0,27

6.1 Conjuntos de dados com rotulação EN

Para conjuntos de dados com rotulação EN (Tabela 4), o Conjunto de dados 1 obteve o melhor resultado com o algoritmo SVM (0,71), dentre todas as combinações

Tabela 5 – Acurácia/*F1-Score* obtidos para cada combinação de algoritmo/conjunto de dados para conjunto de dados rotulados como AE.

	RF	**NB**	**MLP**	**SVM**	**XGB**
Original (5)	0,68/0,71	0,52/0,63	0,65/0,72	0,71/0,69	**0,68/0,69**
PCA (6)	0,61/0,65	0,55/0,65	0,65/0,67	0,55/0,61	0,58/0,61
***ReliefF* (7)**	**0,68/0,74**	**0,71/0,78**	**0,71/0,77**	**0,77/0,81**	0,65/0,72
GA (8)	0,68/0,71	0,55/0,63	0,71/0,73	0,68/0,72	0,65/0,67

de conjuntos de dados e classificadores. Já a pior acurácia foi obtida com o Conjunto de dados 1 e o classificador NB (0,32). É possível observar que, para o Conjunto de dados 2, todos os classificadores obtiveram acurácia com valores próximos a 0,5 e a melhor acurácia obtida (0,52) foi a mais baixa entre todas as melhores acurácias para os conjuntos de dados rotulados como EN.

Observando as combinações que obtiveram a melhor acurácia para cada algoritmo (em negrito na Tabela 4), é possível perceber que técnicas de comitê (RF e XGB) obtiveram uma menor acurácia; no entanto, a diferença de desempenho não foi tão diferente em relação aos demais algoritmos. Para o classificador XGB, é possível que o algoritmo tenha sofrido *overfitting*, uma vez que a melhor acurácia alcançada apresentou valor próximo a 0,5.

6.2 Conjuntos de dados com rotulação AE

Pode-se observar que, novamente, a melhor acurácia foi alcançada com o classificador SVM. Assim como notado na Tabela 4, ao observar os valores em negrito na Tabela 5, as técnicas de comitês (RF e XGB) também alcançaram uma menor acurácia, mas o seu desempenho não foi notoriamente inferior aos outros algoritmos, na maioria dos casos. O pior resultado foi obtido, mais uma vez, pelo classificador NB para o Conjunto de dados original (5).

O Conjunto de dados 7 apresentou a maior acurácia para todos os algoritmos, exceto para o XGB, que obteve o melhor resultado com o Conjunto de dados 5. No geral, o Conjunto de dados 6 obteve desempenho pior que todos os outros Conjuntos de dados com rotulação AE.

Para o classificador NB nos conjuntos de dados rotulados como AE, o *ReliefF*

alcançou a melhor acurácia por uma grande margem, o que pode significar que as características selecionadas também apresentam um distribuição mais próxima à normal quando comparado a outros Conjuntos de dados, beneficiando o classificador NB, uma vez que a implementação utilizada deste assume uma distribuição normal para variáveis numéricas.

6.3 Análise geral

Para todos os classificadores, as melhores acurácias foram observadas nos conjuntos de dados com rotulação AE. Isto pode ser resultado do desbalanceamento encontrado nos conjuntos de dados rotulados como EN, pois, para estes casos, as configurações escolhidas foram aquelas que apresentaram o melhor *F1-Score* médio durante a calibração, e não a melhor acurácia. Por este motivo, esta etapa de teste dos conjuntos de dados com rotulação EN foi repetida, porém desta vez a configuração escolhida foi a que apresentou a melhor acurácia de calibração (Tabela 6). Pode-se observar que, embora a acurácia tenha aumentado para o RF e XGB, muitas das combinações de algoritmo/conjunto de dados tiveram valores muito baixos ou zero para o *F1-Score*. Além disso, todos os classificadores ainda obtiveram uma acurácia menor do que os classificadores com os conjuntos de dados rotulados como AE, com exceção do XGB, que obteve um aumento marginal na acurácia, porém com o custo de ter um *F1-Score* de zero. Tal fato é reflexo do *overfitting* que ocorreu na classe majoritária, de forma que estes algoritmos não classificaram nenhuma instância da classe minoritária corretamente, resultando em um *F1-Score* de zero. Em alguns casos, pode-se observar que o classificador considerou todas as instâncias como sendo da classe majoritária, resultando em acurácia de 0,71, mas *F1-Score* de zero, o que é reflexo de um treinamento com dados desbalanceados.

Outra possível explicação para o pior resultado dos conjuntos de dados com rótulo EN é a definição adotada para a rotulação. Os participantes que declararam saber executar o procedimento foram considerados experientes, assim como alunos de pós graduação, professores e dentistas, porém é possível que o nível de habilidade dos participantes para executar a trajetória (da qual são extraídas as características) seja bem discrepante entre eles, não apenas em termos de conhecimentos, mas em frequência de execução do procedimento. Isso pode ter resultado em uma classe de instâncias muito diversas, algumas delas podendo até mesmo sobrepor a classe de inexperientes, o que pode dificultar sobremaneira a classificação de instâncias.

Podemos também observar que, para alguns dos casos dos conjuntos de dados com rotulação EN, a acurácia obtida foi abaixo de 0,5. Para uma classificação binária, isto indica que poderia ser obtida uma acurácia maior caso a classificação fosse invertida, ou seja, as instâncias classificadas como da classe positivas seriam consideradas da classe negativas, e vice-versa, indicando que o modelo classifica as instâncias da forma oposta ao esperado.

Os melhores resultados dos conjuntos de dados rotulados como AE podem também significar que os especialistas são mais capazes de rotular o desempenho de trajetórias de forma que a diferença entre classes seja mais clara (a fronteira de decisão do conjunto de dados é bem definida) em comparação a rotulação por nível de experiência. Esta questão, porém, necessita de mais estudo e pesquisas na área poderiam se beneficiar do uso de um especialista para rotular os dados.

Tabela 6 – Acurácia/*F1-Score* obtidos para cada combinação de algoritmo/conjunto de dados para conjunto de dados rotulados como EN e selecionados com base na maior acurácia de calibração.

	RF	**NB**	**MLP**	**SVM**	**XGB**
Original (1)	0,52/0,29	0,39/0,39	0,29/0,45	0,68/0	0,61/0,33
PCA (2)	**0,65/0,27**	0,61/0,50	0,29/0,45	0,68/0	**0,71/0**
ReliefF (3)	0,61/0,40	0,65/0,15	0,29/0,45	**0,71/0**	**0,71/0**
GA (4)	0,61/0,25	**0,65/0,27**	**0,68/0,55**	0,61/0,50	0,48/0,27

As combinações algoritmo/conjunto de dados com as melhores acurácias nas Tabelas 4 e 5 (valores em negrito) são apresentadas com todas as suas métricas nas Tabelas 7 e 8. As figuras 20 e 21 comparam as métricas calculadas para cada combinação de algoritmo/conjunto de dados.

6.3.1 Seleção e fusão de características

A redução de dimensionalidade do conjunto de dados é muito importante para tornar a previsão do algoritmo mais confiável, já que conjuntos de dados com um alto número de características podem ser muito esparsos (GÉRON, 2019). Enquanto para os conjuntos de dados rotulados como EN parece não haver nenhum conjunto de dados que consistentemente obteve os melhores resultados, o uso do algoritmo *ReliefF* proporcionou o melhor resultado para quatro dos cinco classificadores nos conjuntos de dados rotulados como AE. Isto pode indicar que o *ReliefF* com rotulação AE

Tabela 7 – Métricas para cada combinação de algoritmo e conjunto de dados que obtiveram a melhor acurácia para cada algoritmo, entre todos os conjuntos de dados rotulados como EN. A combinação que obteve o melhor resultado está apresentada em negrito.

Algoritmo indutor	**Algoritmo de seleção/fusão**	**Acurácia**	**Acurácia Balanceada**	**Sensibilidade**	**Especificidade**	**Precisão**	***F1-Score***
RF	*ReliefF*	0,58	0,54	0,44	0,64	0,33	0,38
NB	GA	0,65	0,52	0,22	0,82	0,33	0,27
MLP	Original	0,68	0,67	0,67	0,68	0,46	0,55
SVM	**Original**	**0,71**	**0,73**	**0,78**	**0,68**	**0,50**	**0,61**
XGB	PCA	0,52	0,53	0,56	0,50	0,31	0,40

Tabela 8 – Métricas para cada combinação de algoritmo e conjunto de dados que obtiveram a melhor acurácia para cada algoritmo, entre todos os conjuntos de dados rotulados como AE. A combinação que obteve o melhor resultado está apresentada em negrito.

Algoritmo indutor	**Algoritmo de seleção/fusão**	**Acurácia**	**Acurácia Balanceada**	**Sensibilidade**	**Especificidade**	**Precisão**	**F1-Score**
RF	ReliefF	0,68	0,67	0,88	0,47	0,64	0,74
NB	ReliefF	0,71	0,70	1.00	0,40	0,64	0,78
MLP	ReliefF	0,71	0,70	0,94	0,47	0,65	0,77
SVM	**ReliefF**	**0,77**	**0,77**	**0,94**	**0,60**	**0,71**	**0,81**
XGB	Original	0,68	0,68	0,69	0,67	0,69	0,69

(Conjunto de dados 7) foi capaz de encontrar as características do conjunto de dados original que podem diferenciar melhor entre as classes, pois os resultados foram consistentes entre os classificadores, enquanto o mesmo não foi observado para o modelo com o uso do *ReliefF* em conjuntos de dados com rótulo EN (Conjunto de dados 3). Além disso, enquanto os algoritmos de seleção GA e *ReliefF* proporcionaram resultados similares, é aparente que o modelo com uso do GA obteve um resultado ligeiramente melhor para os conjuntos de dados rotulados como EN, enquanto o modelo com uso do ReliefF obteve resultado levemente superior para os conjuntos de dados rotulados como AE.

A fusão de características com PCA não proporcionou um bom desempenho para o problema em questão, possivelmente porque o PCA cria CPs que maximizam a variação, o que, porém, não resulta necessariamente em uma melhor separação entre as classes, uma vez que o PCA não leva em consideração as classes ao criar os CPs.

6.3.2 Classificadores

Na tabela 7 pode-se notar que,além de ter obtido a maior acurácia, o classificador SVM foi o mais bem sucedido ao identificar instâncias da classe minoritária (classe positiva) quando comparado com os outros classificadores. Isso pode ser observado no *F1-Score* obtido, sendo o maior entre todas as combinações de algoritmos e conjunto de dados para a rotulação EN.

Na tabela 8 pode-se observar que, com exceção do algoritmo XGB, todos os classificadores alcançaram uma sensibilidade muito maior do que a especificidade. Esses resultados foram obtidos usando os Conjuntos de dados com *ReliefF* (Conjunto de dados 7), enquanto o XGB usou o Conjunto de dados Original (Conjunto de dados 5). Isto pode indicar que as características selecionadas com o *ReliefF* para o conjunto de dados com a rotulação de AE favoreceram a identificação de instâncias da classe positiva, mas não da classe negativa.

6.3.3 Balanceamento

A partir das tabelas 7 e 8, pode-se observar que, embora os conjuntos de dados com a rotulação EN utilizassem algoritmo de balanceamento e as melhores configurações fossem selecionados com base no maior *F1-Score* na calibração, os conjuntos de dados que utilizaram a rotulação AE ainda alcançaram um *F1-Score* mais alto. Isto indica que, mesmo após o uso de técnicas que tentam mitigar o desbalanceamento das classes, não foi possível alcançar um resultado semelhante ao obtido com um conjunto de dados naturalmente balanceado. Independentemente deste fato, como mostrado por Batista, Prati e Monard (2004), o algoritmo de balanceamento *Smote-Tomek links* mostra uma melhora nos resultados em geral, quando comparados ao conjunto de dados original desbalanceado, especialmente para conjuntos de dados com um pequeno número de instâncias positivas (classe minoritária). Nesses casos, se não for realizado o *oversampling* da classe positiva, pode ocorrer uma representação limitada da classe minoritária, resultando em uma classe imprópria para o aprendizado. Ainda assim, é razoável assumir que a classe minoritária terá uma maior chance de apresentar dados homogêneos, mesmo que seja realizado o balanceamento, uma vez que os novos dados sintéticos são criados a partir do reduzido número de instâncias já existentes. Isso pode resultar em um pior desempenho do modelo para identificar novas instâncias da classe minoritária. Apesar de ambas as classes terem o mesmo número

de instâncias de treinamento após o balanceamento, as instâncias da classe minoritária seriam menos variadas e representativas da classe, diminuindo a capacidade de generalização do classificador.

Outros algoritmos também poderiam ser aplicados para o mesmo propósito, por exemplo, Batista, Prati e Monard (2004) comparam algoritmos de balanceamento, incluindo os *Smote-Tomek links*, e indicam que outros algoritmos, tal como o *SMOTE-Edited Nearest Neighbor* também mostram uma área sob a curva similar ao realizar a curva ROC (do inglês, *Receiver Operating Characteristic*).

6.3.4 Análise das métricas

Para os conjuntos de dados com a rotulação EN, a maior acurácia balanceada foi alcançada pelo algoritmo SVM com Conjunto de dados 1 (0,73), a maior sensibilidade foi com o NB com o Conjunto de dados 1 (1), a maior especificidade foi alcançada pelo algoritmo NB com Conjunto de dados 4 (0,82), a maior precisão pelo algoritmo SVM com Conjunto de dados 1 (0,50) e o maior *F1-Score* pelo algoritmo SVM com Conjunto de dados 1 (0,61) (Figura 20).

Para os conjuntos de dados com o rótulo AE, a maior acurácia balanceada foi alcançada pelo algoritmo SVM com Conjunto de dados 7 (0,77), a maior sensibilidade foi no NB com Conjunto de dados 7 (1), a maior especificidade foi alcançada pelo algoritmo SVM com Conjunto de dados 5 (0,80), a maior precisão pelo algoritmo SVM com Conjunto de dados 5 (0,77) e o maior *F1-Score* pelo algoritmo SVM com Conjunto de dados 7 (0,81) (Figura 21). Também pode-se ver que o Conjunto de dados com o *ReliefF* (Conjunto de dados 7) proporcionou a maior sensibilidade para todos os algoritmos, além de obter a menor especificidade para todos os algoritmos, exceto para o NB. Isto reforça a noção de que o Conjunto de dados com o *ReliefF* para AE favoreceu a identificação da classe positiva em detrimento da classe negativa.

Para ambos os casos, seria também possível escolher os modelos de forma a reduzir o número de instâncias classificadas erroneamente como especialistas/aprovados, uma vez que estes alunos seriam considerados aptos a realizar o procedimento equivocadamente pelo simulador, porém a sua inaptidão no procedimento poderia ser de grande risco ao paciente real. Um aluno, porém, que está apto para realizar o procedimento, mas foi avaliado como novato/reprovado, não trará riscos a um paciente, além de que este poderia pedir uma revisão de sua avaliação pelo professor. Este modelo reduziria a chance de ocorrer uma classificação errada que poderia tra-

Figura 20 – Gráfico de barras de cada métrica para cada combinação de algoritmo e conjunto de dados rotulados como EN.

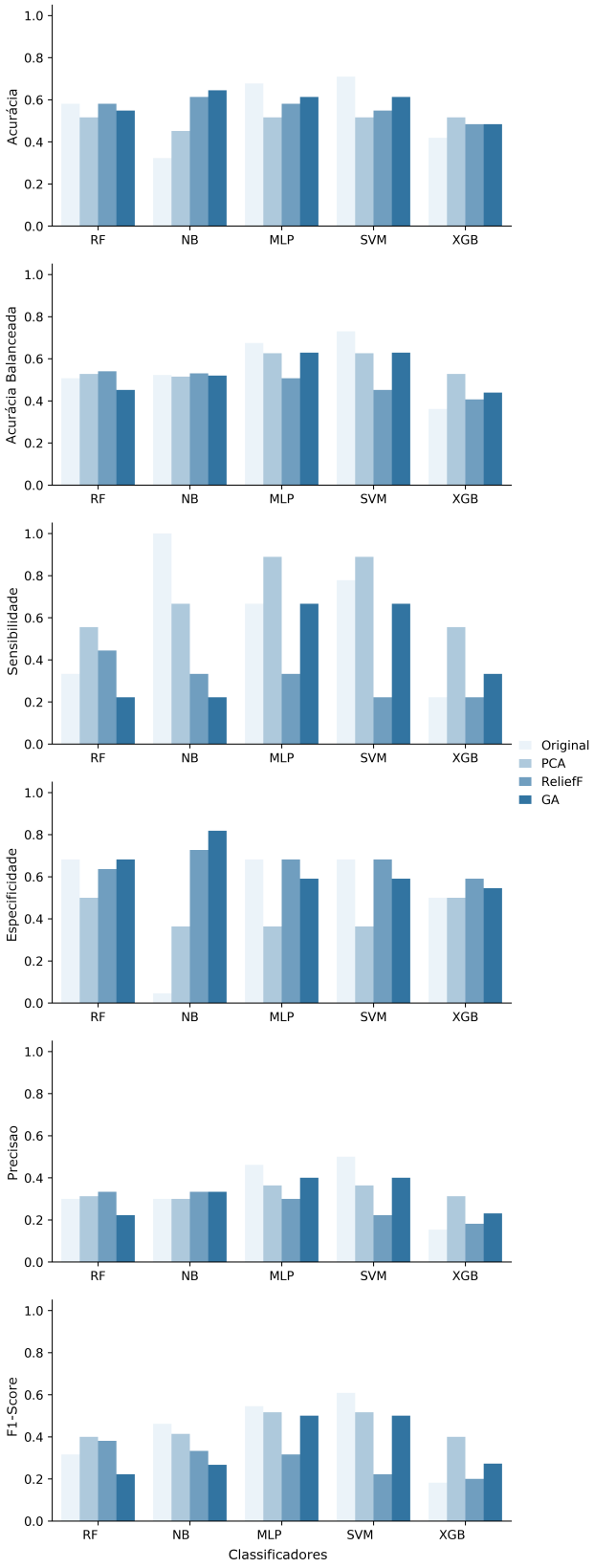
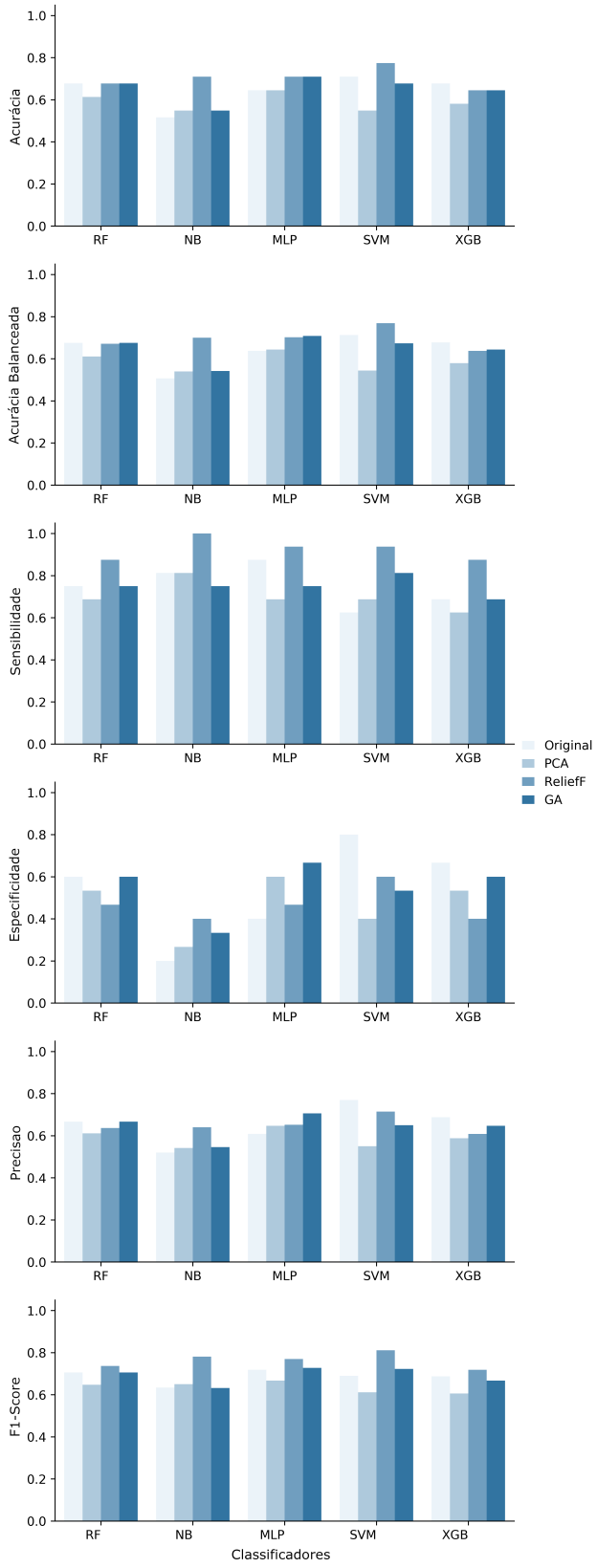


Figura 21 – Gráfico de barras de cada métrica para cada combinação de algoritmo e conjunto de dados rotulados como AE.



zer risco ao paciente, porém com o custo de mais alunos reprovados indevidamente pelo sistema. Dessa forma, para a rotulação EN, tal modelo deve maximizar a especificidade (quantidade menor de falso negativos) e, para a rotulação AE, o modelo deve maximar a sensibilidade (quantidade menor de falso positivos). Assim, uma possível escolha de modelo seria, para rotulação EN, o classificador NB com seleção de características pelo GA (Conjunto de dados 4, acurácia 0,65 e especificidade 0,82), e para a rotulação AE, o classificador NB com o seletor de características *ReliefF* (Conjunto de dados 7, acurácia 0,71 e sensibilidade 1).

Considerando o modelo que maximiza a acurácia, para a rotulação EN, o melhor modelo foi encontrado utilizando o conjunto de dados original e o classificador SVM, obtendo-se uma acurácia de 0,71. Já para a rotulação AE, o melhor modelo utilizou o algoritmo de seleção de características *ReliefF* e o classificador SVM, obtendo acurácia de 0,77. Apesar de o modelo AE obter uma acurácia melhor, a escolha do modelo final é dependente do objetivo que se pretende alcançar com a avaliação automática. Caso o objetivo seja comparar os estudantes com praticantes de diferentes níveis de experiencia e identificar em qual nível ele se encontra, o modelo com rotulação EN é mais adequado. Caso o objetivo seja reproduzir a forma de avaliação realizada por um professor, o modelo AE é mais adequado. Neste caso, a fim de aproximar o modelo da rotina clínica, seria ideal que mais de um professor realizasse a avaliação das instâncias de treinamento e a rotulação final fosse realizada por uma composição de avaliações, de forma que a classificação final não reflita apenas na avaliação de um único instrutor.

O algoritmo de seleção de características *ReliefF* para conjunto de dados rotulados como AE, quando utilizado com o classificador SVM, resultou na melhor acurácia encontrada. Este seletor selecionou as características: Distância da agulha ao alvo, Eficiência nos segmentos A e B, Eficiência da trajetória completa, DTW no segmento C, DTW no segmento D, Eficiência no segmento B e DTW na trajetória completa, nesta ordem. Essas características parecem concordar com dois aspectos considerados de grande importância para a especialista que realizou a avaliação das trajetórias: a distância da agulha ao alvo (presente não apenas na característica de mesmo nome, mas também como parte do cálculo da Eficiência) e a similaridade com a trajetória ideal (presente no DTW). Tal fato, juntamente com os bons resultados em combinação com os classificadores, reforça a capacidade do algoritmo *ReliefF* de identificar as características mais relevantes para a classificação segundo a avaliação de uma especialista. Também com base nas características selecionadas neste caso, não é

possível afirmar que algum dos segmentos da trajetória foi mais importante para a classificação, uma vez que todos os segmentos estão representados entre as características selecionadas.

Com bases nesses resultados, pode-se notar que o classificador SVM apresentou a melhor acurácia para ambas as rotulações e foi o classificador utilizado no modelo final escolhido para ambos os casos. Também é possível observar que, apesar do classificador NB não ter obtido os melhores resultados quando comparado aos outros classificadores testados, fica claro que a sua utilização pode ser de grande relevância caso o contexto de aplicação não seja estritamente a maximização da acurácia.

6.3.5 Comparação com a literatura

A Tabela 9 detalha os trabalhos similares encontrados na literatura. Rhienmora et al. (2011) utilizaram a técnica HMM em um simulador odontológico em RV para o preparo da coroa dentária e descobriram que o modelo era capaz de classificar corretamente novatos e especialistas com 100% de acurácia. Enquanto este estudo mostra que o HMM pode alcançar bons resultados para o problema em questão, os autores também observam que o número de participantes foi relativamente pequeno, com N = 10 (RHIENMORA et al., 2011). Também pode ser observado que o conjunto de dados é naturalmente balanceado, o que pode ter favorecido o bom desempenho do modelo.

Um estudo no campo da ortopedia envolvendo perfuração cirúrgica foi realizado por Zahedi et al. (2020), no qual um modelo usando o algoritmo *Hidden-state Conditional Random Fields* (HCRF) foi treinado com 12 participantes e se mostrou capaz de diferenciar Especialistas e Residentes com uma acurácia de 100% (os movimentos dos especialistas foram corretamente reconhecidas em média em 89,1% do tempo, e 88,4% do tempo para os residentes). Embora o modelo HCRF tenha um desempenho muito bom para o problema, o número de participantes no estudo também foi relativamente pequeno, o que torna o conjunto de dados moderadamente desbalanceado (42% de amostras da classe minoritária), apesar da classe majoritária apresentar apenas dois participantes a mais que a classe minoritária.

Vaughan e Gabrys (2020) treinaram um modelo de vizinhos mais próximos (kNN) usando a métrica DTW (DTW-1-NN), outro modelo usando o algoritmo de centroides mais próximos com a métrica *SoftDTW* e diferentes ANNs para aprendizado profundo, utilizando os dados coletados em um simulador de RV para procedimento de anestesia

epidural. Os autores descobriram que, enquanto o DTW-1-NN atingiu apenas 60% de acurácia, o modelo de centróides mais próximos atingiu 77,5%. Das ANNs treinadas, a Rede Neural Residual, em inglês *Residual Neural Network* (ResNet), obteve o melhor resultado, com 85% de acurácia. Embora o estudo tenha comparado diferentes modelos e encontrado bons resultados, o número de participantes também foi muito limitado. As 271 amostras foram coletadas a partir de 7 participantes que realizaram a simulação múltiplas vezes, o que pode ter resultado em dados mais homogêneos. Além disso, não é clara a distribuição das amostras entre as classes, ou se os dados eram balanceados, porém nenhuma menção à técnicas de balanceamento foi feita.

Bissonnette et al. (2019) compararam diferentes algoritmos de classificação com dados coletados no simulador em RV NeuroVR, durante um procedimento de hemilaminectomia. No total 22 participantes seniores e 19 participantes juniores fizeram parte do experimento e diversas métricas de desempenho foram extraídas e selecionadas. Os autores reportaram que a SVM atingiu a maior acurácia, com 97,6%, seguida pelos kNN, com 92,7%, análise discriminante linear, em inglês *Linear Discriminant Analysis* (LDA), com 87,8%, árvore de decisão com 70,7%, e NB com 65,9%. Este resultado está de acordo com o obtido neste trabalho, no qual a SVM alcançou a melhor acurácia. Os dados utilizados também são balanceados, o que pode ter influenciado no resultado final.

O mesmo grupo de pesquisadores também publicou mais um trabalho usando o simulador em RV Sim-Ortho para um procedimento de discectomia cervical anterior (ALKADRI et al., 2021), com dados de 9 participantes que finalizaram a residência, 5 seniores e 7 juniores. Várias métricas foram novamente extraídas e selecionadas e um modelo utilizando MLP foi treinado com os dados, obtendo uma acurácia de 83,3% ao diferenciar os juniores dos seniores e participantes com residência finalizada. Os autores também utilizaram o *Connection Weights Algorithm* para encontrar a importância de cada característica na classificação e descobriram que o número de contatos com a dura-máter da coluna vertebral e a quantidade máxima de força aplicada no ligamento longitudinal posterior esquerdo foram as características mais importantes para diferenciar os juniores dos participantes com residência finalizada e seniores. Embora os autores reconheçam que o tamanho da amostra foi relativamente pequena, o modelo alcançou bons resultados e o uso do *Connection Weights Algorithm* para reconhecer as métricas que eram mais relevantes foi um passo importante para oferecer um melhor *feedback* ao usuário. O conjunto de dados também é moderadamente desbalanceado entre as classes (43% com residência finalizada, 24% seniores e 33%

juniores), porém nenhuma técnica de balanceamento foi mencionada.

No geral, pode-se notar que, embora estudos semelhantes tenham alcançado uma acurácia maior, o resultado é altamente dependente do conjunto de dados utilizado e os resultados deste estudo podem ser comparáveis à literatura quando considera-se quantidade de amostras do conjunto de dados, com 77% de acurácia. Este estudo também utilizou um maior número de participantes, e os resultados também concordam com os de Bissonnette et al. (2019), no qual a SVM mostrou fornecer bons resultados para o tipo de problema. A maior quantidade de amostras de diferentes participantes do presente trabalho é um diferencial em relação à literatura. No entanto, deve ser ponderado que, se por um lado este fato pode ter contribuído para um treinamento mais adequado dos modelos, por outro o desbalanceamento e a diversidade de voluntários podem ter sido fatores preponderantes para que a acurácia observada apresentasse valores bem discrepantes, de acordo com a combinação de técnicas utilizadas. Apesar disso, levando-se em conta a área de aplicação deste sistema, uma acurácia de 77% pode ser considerada relativamente baixa, uma vez que poderia afetar a nota final dos alunos caso o professor utilizasse a avaliação automática para tal, porém a classificação errada não seria crítica, uma vez que o aluno poderia pedir a revisão do resultado obtido. Outro possível problema seria a classificação incorreta de alunos que ainda não estão aptos para realizar o procedimento. É importante ressaltar que a utilização da avaliação automática deve ser de caráter de auxílio ao professor, de forma que cabe ao especialista tomar a decisão final para aprovar ou não o aluno. Além disso, apesar do número de participantes do estudo aqui apresentado ser maior do que nos estudos relacionados, um número maior de instâncias seria recomendado para aplicação de determinadas técnicas de ML, a fim de aumentar as chances de generalização dos resultados.

Para os estudos citados, nos casos em que foi obtida acurácia acima de 80%, pode-se observar, pela Tabela 9, que foram utilizadas métricas relacionadas à força ou algoritmos capazes de utilizar a trajetória completa, como a ResNet, ou segmentá-la, como o HMM e o HCRF. Tais fatores podem ter influenciado no resultado final obtido, uma vez que, no presente trabalho, esses tipos de características não foram utilizadas. Também é relevante notar que, entre os procedimentos simulados, a anestesia epidural estudada por Vaughan e Gabrys (2020) aparenta ter maior similaridade com o procedimento de BNAI, uma vez que ambos se referem à aplicação anestésica com agulha, e os resultados obtidos foram similares aos apresentados neste estudo, com exceção da ResNet, que realiza aprendizado profundo utilizando a série temporal da

trajetória como entrada. Dessa forma, as técnicas utilizadas em trabalhos similares, tal como de aprendizado profundo, poderiam também ser estudadas com o objetivo de melhorar a acurácia do sistema.

Como ressaltado anteriormente, os trabalhos relacionados utilizaram poucos dados na composição do modelo, com o número de participantes variando entre sete e 41. Um número reduzido de participantes, ainda que estes realizem o procedimento simulado diversas vezes para aumentar o número de instâncias do conjunto de dados, pode limitar a diversidade dos dados, reduzindo a capacidade de generalização do modelo de avaliação.

6.3.6 Limitações e aplicações práticas

Este trabalho apresenta algumas limitações relacionadas principalmente ao escopo considerado. Primeiro, uma quantidade limitada de algoritmos de classificação foi testada, sendo que outros possíveis algoritmos também poderiam ser testados nos conjuntos de dados. O objetivo deste estudo, entretanto, não era testar extensivamente os algoritmos de classificação, mas sim definir um modelo que pudesse ser usado para avaliar o desempenho em simuladores médicos, com uma acurácia que viabilizaria seu uso nas atividades rotineiras de aquisição de habilidades por meio de simulares de Realidade Virtual, dentro do escopo considerado. Segundo, como o conjunto de dados original era pequeno, o conjunto de testes continha apenas 31 instâncias. Isto resultou em uma variação maior entre os resultados dos modelos, o que pode ter influenciado os resultados finais. Um novo experimento foi originalmente planejado para coletar mais dados, porém esses planos foram interrompidos devido à pandemia da COVID-19. Finalmente, o modelo proposto foi baseado em uma classificação binária. Para os conjuntos de dados rotulados como EN, uma classificação multi-rótulo poderia ser aplicada. Para os conjuntos de dados rotulados como AE, modelos de regressão também poderiam ser utilizados. Esses casos estavam fora do escopo deste estudo, mas devem ser considerados em trabalhos futuros.

A arquitetura descrita neste estudo teve a colaboração de especialistas no intuito de reproduzir a avaliação que é feita em procedimentos reais. Dessa forma, o modelo poderia ser adaptado e/ou expandido para ser utilizado em simuladores médicos em RV que utilizam dispositivos hápticos para auxiliar o instrutor, fornecendo um *feedback* em tempo real sobre o desempenho do aluno. Para tal, seria necessário coletar os dados de múltiplos especialistas e novatos para que o modelo pudesse estabelecer uma

distinção entre eles e classificar os novos estudantes de acordo. Alternativamente, o instrutor poderia atribuir notas às trajetórias coletadas e o modelo estabeleceria uma distinção entre aprovados e reprovados. Neste caso, seria possível que os estudantes requisitassem a revisão da classificação atribuída pelo modelo e, caso o instrutor não considerasse que o resultado do modelo foi correto, ele poderia ser reclassificado e seus dados utilizados para treinar o algoritmo novamente, garantindo uma atualização constante do modelo. Esta abordagem, entretanto, exigiria que o instrutor revisasse o desempenho de alguns alunos frequentemente, de forma aleatória, uma vez que a revisão apenas dos casos solicitados por estudantes poderia introduzir viés ao modelo, pois não é provável que um estudante peça uma revisão quando sua classificação foi positiva (aprovada). Tal fato resultaria em uma maior probabilidade dos falsos negativos serem relatados e corrigidos, quando comparado com os falsos positivos.

O sistema de avaliação automática proposto neste trabalho poderia ser utilizado para permitir que o aprendiz possa treinar sem a necessidade de um instrutor presente, proporcionando maior liberdade para que este treine por quanto tempo julgar necessário ou até quando atingir um desempenho adequado. Para tal, o estudante seria avaliado a cada execução do procedimento simulado, portanto seria interessante que o *feedback* proporcionado fosse mais detalhado, fornecendo informações que ajude na identificação dos aspectos da execução do procedimento que devem ser melhorados.

Um sistema de aprendizado adaptativo também poderia ser implementado, no qual o aluno poderia receber um maior número de instruções e ajudas caso seu desempenho não esteja adequado. Quando o aprendiz mostrasse domínio da técnica, obtendo boa classificação pelo modelo, as instruções seriam gradativamente removidas ou ele poderia até mesmo ser conduzido para casos mais complexos do procedimento e situações de emergência. O ensino seria modelado de acordo com a velocidade de aprendizado de cada indivíduo, proporcionando um treinamento personalizado.

A partir do modelo de avaliação automática, um módulo de gamificação poderia também ser projetado. A avaliação automática seria uma das métricas utilizadas para calcular uma pontuação que possibilitaria ao aprendiz avançar para versões mais difíceis do procedimento, como situações de risco, variações anatômicas do paciente ou até condições mais raras, propiciando o treinamento de determinada situação múltiplas vezes, mas ainda criando mecanismos que auxiliam na motivação e possibilitam que o aluno participe no processo do aprendizado.

Tabela 9 – Trabalhos relacionados encontrados na literatura. Não foi feita menção à técnicas de balanceamento nos estudos citados.

Referencia	**Procedimento**	**Dados**	**Tipos de características**	**Resultados**
(RHIENMORA et al., 2011)	Preparo do dente para coroa	10 participantes Classes: - 5 Novatos - 5 Experientes	a) Força no instrumento através do tempo b) Posição/orientação do instrumento através do tempo c) Métricas específicas do instrumento d) Técnica específica corretamente/incorretamente aplicada	Acurácia: HMM: 100%
(ZAHEDI et al., 2020)	Perfuração do fêmur	12 participantes: Classes: - 5 Experientes - 7 Residencia	Posição/orientação do instrumento através do tempo	Acurácia: HCFR: 100%
(VAUGHAN; GABRYS, 2020)	Anestesia epidural	7 participantes: Classes: - Novatos - Intermediários - Experientes 271 coletas total: Não é clara a divisão dos dados dentro das três classes	1) DTW-1- NN: Similaridade entre trajetórias 2) Centróides mais próximos: Similaridade entre trajetórias 3) Aprendizado profundo: séries temporais	Melhores acurácias: 1) DTW-1- NN: 60% 2) Centróides mais próximos: SoftDTW: 77,5%. 3) Aprendizado profundo: ResNet: 85%,
(BISSONNETTE et al., 2019)	Hemilaminectomia	41 participantes: Classes: - 22 seniores - 19 juniores	a) Força máxima aplicada b) Tempo transcorrido em uma etapa chave do procedimento c) Toques errôneos no tecido d) Consistência de movimentos e) Velocidade média f) Aceleração média g) Velocidade máxima h) Variância no ângulo do instrumento	Acurácias: 1) SVM: 97,6% 2) kNN: 92,7% 3) LDA: 87,8% 4) Árvore de decisão: 70,7% 5) NB: 65,9%
(ALKADRI et al., 2021)	Discectomia cervical anterior	21 participantes Classes: - 9 com residência - 5 seniores - 7 juniores	a) Velocidade máxima b) Velocidade média c) Força máxima aplicada d) Quantidade de tecido saudável removido e) Tempo transcorrido em uma etapa chave do procedimento f) Comprimento do contato com a vértebra	Acurácia: MLP: 83,3%
Presente estudo	**BNAI**	**155 participantes** **1) Classes:** **- 110 Experientes** **- 45 Inexperientes** **2) Classes:** **- 80 Aprovados** **- 75 Reprovados:**	**a) Comprimento do trajeto do instrumento** **b) Acúmulo do ângulo do instrumento** **c) Tempo transcorrido durante o experimento por completo** **d) Média, máxima e desvio padrão da velocidade, aceleração, *jerk* e velocidade de rotação** **e) Firmeza** **f) Velocidade acima/abaixo de um limiar** **g) Semelhança entre trajetórias** **h) Profundidade de inserção do instrumento** **i) Ângulo entre o instrumento e o alvo** **j) Distância do instrumento ao alvo** **k) Eficiência**	**Acurácias:** **1) RF: 58%** **NB: 65%** **MLP: 68%** **SVM: 71%** **XGB: 52%** **2) RF: 68%** **NB: 71%** **MLP: 71%** **SVM: 77%** **XGB: 68%**

7 CONSIDERAÇÕES FINAIS

Neste trabalho foi definido um modelo de ML para avaliação automática de desempenho de aprendizes em simuladores médicos em RV com dispositivo háptico. Um estudo de caso foi conduzido com o simulador odontológico VIDA Odonto para validar a viabilidade do modelo.

Foi observado que, embora parte do conjunto de dados não fosse balanceado e uma técnica de balanceamento foi aplicada para tentar mitigar este problema, os resultados obtidos com os conjuntos de dados desbalanceados, rotulados como EN, foram piores quando comparados com os resultados dos conjuntos de dados que eram naturalmente balanceados (conjuntos de dados rotulados como AE). É possível que o desbalanceamento tenha sido um fator relevante durante a classificação e futuras pesquisas na área devem planejar experimentos a fim de gerar conjuntos de dados com classes balanceadas, embora isso seja uma tarefa não trivial.

Em geral, foi observado que o algoritmo SVM obteve os melhores resultados tanto para os dados rotulados como EN quanto para os rotulados como AE. Verificar se métodos que são beneficiados por grandes quantidades de dados, como o MLP, superariam os resultados da SVM caso o número de participantes fosse muito maior pode ser um estudo relevante.

Considerando as formas de rotulação definidas, o melhor modelo para a rotulação EN utilizou o conjunto de dados original e o classificador SVM, obtendo uma acurácia de 0,71 e 0,61 de *F1-Score*. Para a rotulação AE, o melhor modelo utilizou o conjunto de dados com seleção de características por *ReliefF* e o classificador SVM, obtendo 0,77 de acurácia e 0,81 de *F1-Score*. Esses modelos poderiam ser incluídos em simuladores e utilizados na prática do treinamento virtual rotineiro de aprendizes.

7.0.1 Trabalhos futuros

Durante este estudo, foram identificadas oportunidades e desafios para o avanço deste trabalho, e da área de pesquisa no geral. São eles:

- para a rotulação AE, modelos de regressão podem ser utilizados para fornecer uma pontuação numérica para o estudante, ao invés de apenas classificá-lo como aprovado ou reprovado;
- novamente para a rotulação AE, a avaliação de cada trajetória poderia ser realizada por mais de um especialista, e a rotulação final feita por maioria ou média, com objetivo de obter uma rotulação final menos subjetiva. Pode ser também utilizada lógica *fuzzy* para introduzir peso em cada avaliação realizada, de forma, por exemplo, em que avaliadores com maior experiência tenham maior peso na nota final;
- para a rotulação EN, modelos multi-rótulo podem ser implementados, de forma que os aprendizes não apenas sejam classificados entre experientes e inexperientes, mas também em níveis intermediários de experiência;
- além dos algoritmos de classificação apresentados, outros algoritmos de ML podem ser testados, a fim de aperfeiçoar o modelo de avaliação;
- apesar deste trabalho utilizar características calculadas a partir de uma trajetória coletada em um simulador, o estudo de algoritmos capazes de utilizar a trajetória completa como entrada pode fornecer resultados relevantes;
- como algoritmos de ML muitas vezes se beneficiam de grandes quantidades de dados de treinamento, seria importante a realização de novos experimentos no simulador VIDA Odonto, a fim de aumentar a confiabilidade dos modelos e estudar se mais dados implicariam em resultados melhores e quais algoritmos mais se beneficiariam;
- métodos de gamificação poderiam ser introduzidos juntamente com o modelo de avaliação automática desenvolvido, como um meio de aumentar o engajamento dos estudantes e motivá-los a aperfeiçoar sua técnica, como apresentado na Seção 6.3.6;
- o estudo de caso deste trabalho foi desenvolvido com base nos dados coletados no simulador VIDA Odonto, portanto um trabalho futuro de grande relevância é

a implementação do modelo final no simulador, para que estudantes possam ser avaliados automaticamente após o treinamento. Uma vez que o modelo foi implementado, seria possível que o aluno realizasse o treinamento independentemente de um instrutor, ou até mesmo com um sistema aprendizado adaptativo, como descrito na Seção 6.3.6;

- para que o aprendiz possa realizar o treinamento autonomamente, seria relevante tornar o modelo mais interpretável, com o objetivo de entender quais características são mais importantes para a classificação, o que também melhoraria o *feedback* para o usuário ao indicar que pontos do precedimento precisam ser melhorados;

- o modelo de avaliação automática de desempenho poderia levar em consideração os vasos e nervos do paciente virtual na região de aplicação da anestesia, de forma a possibilitar um refinamento maior do local correto de aplicação da anestesia e evitando maiores danos ao paciente;

- um sistema de avaliação automática capaz de identificar erros do aluno poderia ser utilizado para implementar emoções no paciente virtual, podendo mostrar nervosismo, dor, ou outros tipos de reações dependendo da atuação do aluno na simulação;

- a força que o aprendiz realiza no simulador durante o procedimento pode ser uma característica de grande relevância em simuladores com dispositivo háptico, porém não foi possível coletá-la neste estudo. Esta característica é de difícil coleta, tanto em simuladores sem dispositivo háptico, assim como na execução do procedimento real, porém a sua análise pode ser relevante para o modelo de avaliação, dependendo do procedimento médico estudado;

- o modelo foi desenvolvido com dados de um simulador para anestesia odontológica, porém acredita-se que ele poderia ser adaptado para uso em simuladores de outros procedimentos médicos com características similares ao simulador utilizado como estudo de caso neste trabalho;

7.0.2 Publicações

Os artigos elaborados durante o desenvolvimento deste trabalho são apresentados a seguir:

- dois *short papers* com resultados parciais foram publicados na trilha *Workshop of Thesis and Dissertations* (WTD) no congresso *SVR* em 2020 (SALLABERRY; TORI; NUNES, 2020) e 2021 (SALLABERRY; TORI; NUNES, 2021b);
- artigo com resultados parciais obtidos foi publicado no congresso *SVR* (SALLABERRY; TORI; NUNES, 2021a);
- artigo detalhado referente à revisão sistemática da literatura apresentada no Capítulo 3 foi publicado no periódio *ACM Computing Surveys* (SALLABERRY; TORI; NUNES, 2022);
- artigo com os resultados obtidos no estudo está em fase de elaboração e será submetido em breve;

REFERÊNCIAS

ACM. **ACM Digital Library**. 2021. <https://dl.acm.org/>. (accessed 2022-12-09).

AKHTAR, K. et al. Training safer orthopedic surgeons: Construct validation of a virtual-reality simulator for hip fracture surgery. **Acta Orthopaedica**, v. 86, n. 5, p. 616–621, set. 2015. ISSN 1745-3674, 1745-3682.

ALKADRI, S. et al. Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure. **Computers in Biology and Medicine**, v. 136, p. 104770, set. 2021. ISSN 00104825.

ANJOS, A. M. dos. **Um método para avaliar a aquisição de habilidades sensório-motoras em ambientes virtuais interativos tridimensionais para treinamento médico.** Tese (Doutorado em Engenharia de Computação) — Universidade de São Paulo, São Paulo, set. 2014.

Autodesk. **Maya**. 2022. <https://www.autodesk.com.br/products/maya/overview>. (accessed 2022-12-09).

AZARNOUSH, H. et al. The force pyramid: A spatial analysis of force application during virtual reality brain tumor resection. **Journal of Neurosurgery**, v. 127, n. 1, p. 171–181, jul. 2017. ISSN 0022-3085, 1933-0693.

BATISTA, G. E. et al. Balancing training data for automated annotation of keywords: A case study. In: **WOB**. [S.l.: s.n.], 2003. p. 10–18.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20–29, jun. 2004. ISSN 1931-0145, 1931-0153.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. (Information Science and Statistics). ISBN 978-0-387-31073-2.

BISSONNETTE, V. et al. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. **Journal of Bone and Joint Surgery**, v. 101, n. 23, p. e127, dez. 2019. ISSN 0021-9355, 1535-1386.

BITTNER, J. G. et al. Face and construct validity of a computer-based virtual reality simulator for ERCP. **Gastrointestinal Endoscopy**, v. 71, n. 2, p. 357–364, fev. 2010. ISSN 00165107.

BOUAICHA, S. et al. Three days of training with a low-fidelity arthroscopy triangulation simulator box improves task performance in a virtual reality high-fidelity virtual knee arthroscopy simulator. **Knee Surgery, Sports Traumatology, Arthroscopy**, v. 28, n. 3, p. 862–868, mar. 2020. ISSN 0942-2056, 1433-7347.

BREUER, J.; BENTE, G. Why so serious? On the relation of serious games and learning. 2010.

BURDEA, G.; COIFFET, P. **Virtual Reality Technology**. 2nd ed. ed. Hoboken, N.J: J. Wiley-Interscience, 2003. ISBN 978-0-471-36089-6.

CHAN, S. et al. Virtual Reality Simulation in Neurosurgery: Technologies and Evolution. **Neurosurgery**, v. 72, p. A154–A164, jan. 2013. ISSN 0148-396X.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, jun. 2002. ISSN 1076-9757.

CHELLALI, A. et al. Preliminary evaluation of the pattern cutting and the ligating loop virtual laparoscopic trainers. **Surgical Endoscopy**, v. 29, n. 4, p. 815–821, abr. 2015. ISSN 0930-2794, 1432-2218.

CHEN, H.-E. et al. Can Haptic Simulators Distinguish Expert Performance? A Case Study in Central Venous Catheterization in Surgical Education:. **Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare**, v. 14, n. 1, p. 35–42, fev. 2019. ISSN 1559-2332.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. San Francisco California USA: ACM, 2016. p. 785–794. ISBN 978-1-4503-4232-2.

CHIKWE, J.; de Souza, A. C.; PEPPER, J. R. No time to train the surgeons. **BMJ**, v. 328, n. 7437, p. 418–419, fev. 2004. ISSN 0959-8138, 1468-5833.

CHOLLET, F. et al. **Keras**. 2015.

COLLAÇO, E. et al. Immersion and haptic feedback impacts on dental anesthesia technical skills virtual reality training. **Journal of Dental Education**, p. jdd.12503, nov. 2020. ISSN 0022-0337, 1930-7837.

CORRÊA, C. G. et al. Virtual Reality simulator for dental anesthesia training in the inferior alveolar nerve block. **Journal of Applied Oral Science**, v. 25, n. 4, p. 357–366, ago. 2017. ISSN 1678-7757.

CORREA, C. G.; TORI, R.; NUNES, F. L. S. Haptic Simulation for Virtual Training in Application of Dental Anesthesia. In: **2013 XV Symposium on Virtual and Augmented Reality**. Cuiabá - Mato Grosso, Brazil: IEEE, 2013. p. 63–72. ISBN 978-0-7695-5001-5.

DE, S. et al. A Physics-Driven Neural Networks-Based Simulation System (PhyNNeSS) for Multimodal Interactive Virtual Environments Involving Nonlinear Deformable Objects. **Presence: Teleoperators and Virtual Environments**, v. 20, n. 4, p. 289–308, ago. 2011. ISSN 1054-7460.

DEO, D.; DE, S. A higher order polynomial reproducing radial basis function neural network (HOPR-RBFN) for real-time interactive simulations of nonlinear deformable bodies with haptic feedback. In: **2010 IEEE Haptics Symposium**. Waltham, MA, USA: IEEE, 2010. p. 527–530. ISBN 978-1-4244-6821-8.

DIAS, R. D.; GUPTA, A.; YULE, S. J. Using Machine Learning to Assess Physician Competence: A Systematic Review. **Academic Medicine**, v. 94, n. 3, p. 427–439, mar. 2019. ISSN 1040-2446.

DMLC. **XGBoost**. 2022. <https://xgboost.readthedocs.io/en/stable/python/index.html>. (accessed 2022-12-09).

ELSEVIER. **Scopus**. 2021. <https://www.elsevier.com/solutions/scopus>. (accessed 2020-12-30).

ENGELHARDT, A. et al. Comparing classification methods for diffuse reflectance spectra to improve tissue specific laser surgery. **BMC Medical Research Methodology**, v. 14, n. 1, p. 91, dez. 2014. ISSN 1471-2288.

ESEN, H. et al. A Multi-User Virtual Training System Concept and Objective Assessment of Trainings. In: **RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication**. Jeju, South Korea: IEEE, 2007. p. 1084–1089. ISBN 978-1-4244-1634-9.

EVE, E. J. et al. Performance of Dental Students Versus Prosthodontics Residents on a 3D Immersive Haptic Simulator. **Journal of Dental Education**, v. 78, n. 4, p. 630–637, abr. 2014. ISSN 00220337.

FIARD, G. et al. Initial Validation of a Virtual-Reality Learning Environment for Prostate Biopsies: Realism Matters! **Journal of Endourology**, v. 28, n. 4, p. 453–458, abr. 2014. ISSN 0892-7790, 1557-900X.

FUCENTESE, S. F. et al. Evaluation of a virtual-reality-based simulator using passive haptic feedback for knee arthroscopy. **Knee Surgery, Sports Traumatology, Arthroscopy**, v. 23, n. 4, p. 1077–1085, abr. 2015. ISSN 0942-2056, 1433-7347.

Gélinas-Phaneuf, N. et al. Assessing performance in brain tumor resection using a novel virtual reality simulator. **International Journal of Computer Assisted Radiology and Surgery**, v. 9, n. 1, p. 1–9, jan. 2014. ISSN 1861-6410, 1861-6429.

GÉRON, A. **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. Sebastopol, CA: O'Reilly Media, Inc., 2019. ISBN 978-1-4920-3264-9.

GIORGINO, T. Computing and Visualizing Dynamic Time Warping Alignments in R : The **dtw** Package. **Journal of Statistical Software**, v. 31, n. 7, 2009. ISSN 1548-7660.

GUO, Y.-R. et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. **Military Medical Research**, v. 7, n. 1, p. 11, dez. 2020. ISSN 2054-9369.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd ed. ed. New York, NY: Springer, 2009. (Springer Series in Statistics). ISBN 978-0-387-84857-0 978-0-387-84858-7.

HUANG, C. et al. Automated Metrics in a Virtual-Reality Myringotomy Simulator: Development and Construct Validity. **Otology & Neurotology**, v. 39, n. 7, p. e601–e608, ago. 2018. ISSN 1531-7129.

IEEE. **IEEE Xplore**. 2021. <https://ieeexplore.ieee.org/Xplore/home.jsp>. (accessed 2021-01-12).

JERALD, J. **The VR Book: Human-Centered Design for Virtual Reality**. First edition. New York: acm, Association for Computing Machinery, 2016. (ACM Books, 8). ISBN 978-1-970001-12-9 978-1-970001-15-0 978-1-970001-13-6.

KHALIL, H. A basic review on the inferior alveolar nerve block techniques. **Anesthesia: Essays and Researches**, v. 8, n. 1, p. 3, 2014. ISSN 0259-1162.

KONONENKO, I.; ŠIMEC, E.; Robnik-Šikonja, M. Overcoming the myopia of inductive learning algorithms with RELIEFF. **Applied Intelligence**, Springer, v. 7, n. 1, p. 39–55, 1997.

LEBLANC, J. et al. A Comparison of Orthopaedic Resident Performance on Surgical Fixation of an Ulnar Fracture Using Virtual Reality and Synthetic Models. **The Journal of Bone & Joint Surgery**, v. 95, n. 9, p. e60, maio 2013. ISSN 0021-9355.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research**, v. 18, n. 17, p. 1–5, 2017.

MA, X. et al. A flexible digit with tactile feedback for invasive clinical applications. **Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine**, v. 218, n. 3, p. 151–158, mar. 2004. ISSN 0954-4119, 2041-3033.

MAGILL, R. A.; ANDERSON, D. **Motor Learning and Control: Concepts and Applications**. Eleventh edition. New York, NY: McGraw-Hill Education, 2017. ISBN 978-1-259-82399-2.

Manuel Calzolari. **Sklearn-Genetic**. 2022. <https://sklearn-genetic.readthedocs.io/en/latest/index.html>. (accessed 2022-12-09).

MCDOUGALL, E. M. Validation of Surgical Simulators. **Journal of Endourology**, v. 21, n. 3, p. 244–247, mar. 2007. ISSN 0892-7790, 1557-900X.

MILGRAM, P. et al. Augmented reality: A class of displays on the reality-virtuality continuum. In: DAS, H. (Ed.). **Photonics for Industrial Applications**. Boston, MA: [s.n.], 1995. p. 282–292.

MIRGHANI, I. et al. Capturing differences in dental training using a virtual reality simulator. **European Journal of Dental Education**, v. 22, n. 1, p. 67–71, fev. 2018. ISSN 13965883.

MITCHELL, M. **An Introduction to Genetic Algorithms**. Cambridge, Mass.: MIT Press, 1996. ISBN 978-0-262-28001-3.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997. (McGraw-Hill Series in Computer Science). ISBN 978-0-07-042807-2.

MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In: **2020 11th International Conference on Information and Communication Systems (ICICS)**. Irbid, Jordan: IEEE, 2020. p. 243–248. ISBN 978-1-72816-227-0.

MORRIS, D. et al. Visuohaptic simulation of bone surgery for training and evaluation. **IEEE Computer Graphics and Applications**, v. 26, n. 6, p. 48–57, nov. 2006. ISSN 0272-1716.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. Cambridge, MA: MIT Press, 2012. (Adaptive Computation and Machine Learning Series). ISBN 978-0-262-01802-9.

NCBI. **PubMed**. 2021. <https://pubmed.ncbi.nlm.nih.gov/>. (accessed 2021-01-12).

O'TOOLE, R. V. et al. Measuring and developing suturing technique with a virtual reality surgical simulator. **Journal of the American College of Surgeons**, v. 189, n. 1, p. 114–127, jul. 1999. ISSN 10727515.

P, A.; MENON, B. M.; RAO, B. R. Performance Categorization for Personalized Learning in Vocational Training Simulators. In: **2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)**. Mumbai: IEEE, 2018. p. 66–68. ISBN 978-1-5386-6049-2.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PRASAD, R. et al. Face and Construct Validity of a Novel Virtual Reality–Based Bimanual Laparoscopic Force-Skills Trainer With Haptics Feedback. **Surgical Innovation**, v. 25, n. 5, p. 499–514, out. 2018. ISSN 1553-3506, 1553-3514.

Rafii-Tari, H. et al. Objective Assessment of Endovascular Navigation Skills with Force Sensing. **Annals of Biomedical Engineering**, v. 45, n. 5, p. 1315–1327, maio 2017. ISSN 0090-6964, 1573-9686.

REILINK, R. et al. Evaluation of flexible endoscope steering using haptic guidance. **The International Journal of Medical Robotics and Computer Assisted Surgery**, v. 7, n. 2, p. 178–186, jun. 2011. ISSN 14785951.

RHIENMORA, P. et al. A Virtual Reality Simulator for Teaching and Evaluating Dental Procedures. **Methods of Information in Medicine**, v. 49, n. 04, p. 396–405, 2010. ISSN 0026-1270, 2511-705X.

RHIENMORA, P. et al. Intelligent dental training simulator with objective skill assessment and feedback. **Artificial Intelligence in Medicine**, v. 52, n. 2, p. 115–121, jun. 2011. ISSN 09333657.

RIA, S. et al. A Scoring System for Assessing Learning Progression of Dental Students' Clinical Skills Using Haptic Virtual Workstations. **Journal of Dental Education**, v. 82, n. 3, p. 277–285, mar. 2018. ISSN 00220337.

RIBEIRO, M. D. L.; NUNES, F. L. S.; ELIAS, S. Towards Determining Force Feedback Parameters for Realistic Representation of Nodules in a Breast Palpation Simulator. In: **2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)**. Dublin: IEEE, 2016. p. 279–284. ISBN 978-1-4673-9036-1.

ROITBERG, B. Z. et al. Evaluation of Sensory and Motor Skills in Neurosurgery Applicants Using a Virtual Reality Neurosurgical Simulator: The Sensory-Motor Quotient. **Journal of Surgical Education**, v. 72, n. 6, p. 1165–1171, nov. 2015. ISSN 19317204.

ROSEN, J. et al. Generalized Approach for Modeling Minimally Invasive Surgery as a Stochastic Process Using a Discrete Markov Model. **IEEE Transactions on Biomedical Engineering**, v. 53, n. 3, p. 399–413, mar. 2006. ISSN 0018-9294.

ROSEN, J. et al. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. **IEEE Transactions on Biomedical Engineering**, v. 48, n. 5, p. 579–591, maio 2001. ISSN 00189294.

ROSEN, J. et al. Hidden Markov models of minimally invasive surgery. **Studies in Health Technology and Informatics-Medicine Meets Virtual Reality**, Citeseer, v. 70, p. 279–285, 2000.

SALLABERRY, L.; TORI, R.; NUNES, F. Avaliação automática de habilidades sensório-motoras em simulador odontológico. In: **Anais Estendidos Do Simpósio de Realidade Virtual e Aumentada (SVR Estendido 2020)**. Brasil: Sociedade Brasileira de Computação, 2020. p. 5–6.

SALLABERRY, L. H.; TORI, R.; NUNES, F. L. S. Comparison of machine learning algorithms for automatic assessment of performance in a virtual reality dental simulator. In: **Symposium on Virtual and Augmented Reality**. Virtual Event Brazil: ACM, 2021. p. 14–23. ISBN 978-1-4503-9552-6.

SALLABERRY, L. H.; TORI, R.; NUNES, F. L. S. Development of a machine learning model for automatic assessment of performance in virtual reality medical simulators. In: **Anais Estendidos Do XXIII Simpósio de Realidade Virtual e Aumentada (SVR Estendido 2021)**. Brasil: Sociedade Brasileira de Computação, 2021. p. 21–22.

SALLABERRY, L. H.; TORI, R.; NUNES, F. L. S. (Aceito para publicação) Automatic Performance Assessment in Three-dimensional Interactive Haptic Medical Simulators: A Systematic Review. **ACM Computing Surveys**, p. 3539222, jun. 2022. ISSN 0360-0300, 1557-7341.

SCHMIDT, R. A.; LEE, T. D. **Motor Learning and Performance: From Principles to Application**. Fifth edition. Champaign, IL: Human Kinetics, 2014. ISBN 978-1-4504-4361-6.

School of Simulation & Visualization - Glasgow School of Art. **Unity 5 Haptic Plugin for Geomagic OpenHaptics 3.3 (HLAPI/HDAPI)**. 2016. <https://assetstore.unity.com/packages/essentials/tutorial-projects/unity-5-haptic-plugin-for-geomagic-openhaptics-3-3-hlapi-hdapi-34393>. (accessed 2021-01-22).

Scikit Learn. **Grid Search**. 2022. <https://scikit-learn.org/stable/modules/grid_search.html#exhaustive-grid-search>. (accessed 2022-12-09).

Scikit Learn. **Random Search**. 2022. <https://scikit-learn.org/stable/modules/grid_search.html#randomized-parameter-optimization>. (accessed 2022-12-09).

SICKLE, K. R. V. et al. Construct validation of the ProMIS simulator using a novel laparoscopic suturing task. **Surgical Endoscopy And Other Interventional Techniques**, v. 19, n. 9, p. 1227–1231, set. 2005. ISSN 0930-2794, 1432-2218.

SIKDER, S. et al. The use of a virtual reality surgical simulator for cataract surgical skill assessment with 6 months of intervening operating room experience. **Clinical Ophthalmology**, p. 141, jan. 2015. ISSN 1177-5483.

SUEBNUKARN, S. et al. Augmented Kinematic Feedback from Haptic Virtual Reality for Dental Skill Acquisition. **Journal of Dental Education**, v. 74, n. 12, p. 1357–1366, dez. 2010. ISSN 00220337.

Surgicalscience. **LAPSIM® – YOUR FIRST CHOICE FOR LAPAROSCOPY TRAINING**. 2020. <https://surgicalscience.com/systems/lapsim/>. (accessed 2021-01-22).

TAI, Y. et al. A novel framework for visuo-haptic percutaneous therapy simulation based on patient-specific clinical trials. In: **2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**. Banff, AB: IEEE, 2017. p. 3362–3366. ISBN 978-1-5386-1645-1.

TAI, Y. et al. Development of Haptic-Enabled Virtual Reality Simulator for Video-Assisted Thoracoscopic Right Upper Lobectomy. In: **2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**. Miyazaki, Japan: IEEE, 2018. p. 3010–3015. ISBN 978-1-5386-6650-0.

TORI, R. et al. VIDA ODONTO: Ambiente de Realidade Virtual para Treinamento Odontológico. **Revista Brasileira de Informática na Educação**, v. 26, n. 02, p. 80–101, maio 2018. ISSN 2317-6121, 1414-5685.

Unity Technologies. **Unity**. 2020. <https://unity.com/products/unity-platform>. (accessed 2020-09-23).

URBANOWICZ, R. J. et al. Benchmarking relief-based feature selection methods for bioinformatics data mining. **Journal of Biomedical Informatics**, v. 85, p. 168–188, set. 2018. ISSN 15320464.

VÅPENSTAD, C. et al. Limitations of haptic feedback devices on construct validity of the LapSim® virtual reality simulator. **Surgical Endoscopy**, v. 27, n. 4, p. 1386–1396, abr. 2013. ISSN 0930-2794, 1432-2218.

VAUGHAN, N.; GABRYS, B. Scoring and assessment in medical VR training simulators with dynamic time series classification. **Engineering Applications of Artificial Intelligence**, v. 94, p. 103760, set. 2020. ISSN 09521976.

WANG, D. et al. Preliminary evaluation of a virtual reality dental simulation system on drilling operation. **Bio-Medical Materials and Engineering**, v. 26, n. s1, p. S747–S756, ago. 2015. ISSN 18783619, 09592989.

Wing-Yin Chan et al. A Serious Game for Learning Ultrasound-Guided Needle Placement Skills. **IEEE Transactions on Information Technology in Biomedicine**, v. 16, n. 6, p. 1032–1042, nov. 2012. ISSN 1089-7771, 1558-0032.

Winkler-Schwartz, A. et al. Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. **Journal of Surgical Education**, v. 76, n. 6, p. 1681–1690, nov. 2019. ISSN 19317204.

WOLPERT, D. M.; DIEDRICHSEN, J.; FLANAGAN, J. R. Principles of sensorimotor learning. **Nature Reviews Neuroscience**, v. 12, n. 12, p. 739–751, dez. 2011. ISSN 1471-003X, 1471-0048.

ZAHEDI, E. et al. Towards Skill Transfer via Learning-Based Guidance in Human-Robot Interaction: An Application to Orthopaedic Surgical Drilling Skill. **Journal of Intelligent & Robotic Systems**, v. 98, n. 3-4, p. 667–678, jun. 2020. ISSN 0921-0296, 1573-0409.