Universidade Federal da Bahia
Escola Politécnica
Programa de Pós-Graduação em Engenharia Elétrica

SISTEMA PARA RECONHECIMENTO DE EMOÇÃO MULTIMODAL E MULTICLASSE PARA INTERAÇÃO HUMANO-ROBÔ

Lara Toledo Cordeiro Ottoni

Orientador: Prof. Dr. Jés de Jesus Fiais Cerqueira

Salvador (BA), 2024.

Universidade Federal da Bahia
Escola Politécnica
Programa de Pós-Graduação em Engenharia Elétrica

SISTEMA PARA RECONHECIMENTO DE EMOÇÃO MULTIMODAL E MULTICLASSE PARA INTERAÇÃO HUMANO-ROBÔ

Lara Toledo Cordeiro Ottoni

Tese apresentada à banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal da Bahia, como parte dos requisitos necessários à obtenção do grau de Doutor em Engenharia Elétrica.

Salvador (BA), 2024.

Ficha catalográfica elaborada pela Biblioteca Bernadete Sinay Neves, Escola Politécnica - UFBA.

O91 Ottoni, Lara Toledo Cordeiro.
Sistema para reconhecimento de emoção multimodal e multiclasse para a interação humano-robô/Lara Toledo Cordeiro Ottoni. – Salvador, 2024.
139 f.: il. color.

Orientador: Prof. Dr. Jés de Jesus Fiais Cerqueira.

Tese (doutorado) – Programa de Pós-Graduação em Engenharia Elétrica, Escola Politécnica, Universidade Federal da Bahia, 2024.

1. Reconhecimento de Emoção. 2. Sistema Multimodal. 3. Interação Humano-Robô. 4. Aprendizado de Máquina. 5. Sistema Fuzzy. I. Cerqueira, Jés de Jesus Fiais. II. Universidade Federal da Bahia. III. Título.

CDD: 623.76

Universidade Federal da Bahia

Programa de Pós-graduação em Engenharia Elétrica

Curso de Doutorado em Engenharia Elétrica

A Banca Examinadora, constituída pelos professores abaixo listados, leram e recomendam a aprovação da Tese de Doutorado, intitulada "Sistema para reconhecimento de emoção multimodal e multiclasse para interação humano-robô", apresentada no dia 04 de Outubro de 2024, como requisito para obtenção do título de Doutora em Engenharia Elétrica.

gov.br
Documento assinado digitalmente
JES DE JESUS FIAIS CERQUEIRA
Data: 21/10/2024 17:59:26-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Jés de Jesus Fiais Cerqueira
UFBA

gov.br
Documento assinado digitalmente
ADRIAO DUARTE DORIA NETO
Data: 24/10/2024 14:34:36-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Adrião Duarte Doria Neto
UFRN

Mariana S Netto
24/10/2024

Profª. Drª. Mariana Schiavo Netto
Gustave Eiffel University

gov.br
Documento assinado digitalmente
MARCOS YUZURU DE OLIVEIRA CAMADA
Data: 24/10/2024 08:39:25-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcos Yuzuru de Oliveira Camada
IF Baiano

gov.br
Documento assinado digitalmente
ANTONIO CARLOS LOPES FERNANDES JUNIOR
Data: 23/10/2024 22:27:15-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Antônio Carlos Lopes Fernandes Júnior
UFBA

Dedico este trabalho aos meus pais Rodrigo e Janaína,

e ao meu marido André.

Agradecimentos

Ao concluir este trabalho, gostaria de expressar minha profunda gratidão a todos que contribuíram para o sucesso desta jornada.

Em primeiro lugar, agradeço a Deus, cuja orientação e força me sustentaram ao longo de todo o processo. Sua sabedoria e graça foram fundamentais para superar os desafios e alcançar este marco significativo em minha vida.

À minha família, meu mais sincero agradecimento. Aos meus pais, Rodrigo e Janaína, sou imensamente grata pelo apoio constante e pelo incentivo inabalável. Agradeço também aos meus irmãos, Júlia e João Paulo, e a todos da família Toledo Cordeiro. À família Ottoni, minha sincera gratidão pelo suporte e apoio nessa jornada.

Ao meu marido André, minha profunda gratidão. Sua paciência, compreensão e apoio foram a luz que iluminou meu caminho durante toda esta jornada. Sua presença ao meu lado me fortaleceu nos momentos de incerteza e celebrou comigo cada conquista.

Ao meu orientador Jés, expresso meu sincero agradecimento. Suas orientações, insights e apoio foram inestimáveis, enriquecendo imensamente minha trajetória acadêmica. Agradeço também à UFBA e pelo apoio financeiro obtido através da FAPESB.

De todo o coração, agradeço à equipe da Esquecer para Descobrir, especialmente à mentora Érica Andrade, por compartilhar sua sabedoria e ensinamentos. Sou profundamente grata por me ajudar a reconhecer meu potencial, capacidade e por me mostrar que é possível seguir os direcionamentos do coração em uma jornada de amor.

Obrigada a todos.

Resumo

O desafio da Interação Humano-Robô (IHR) é construir sistemas inteligentes que possam se adaptar às mudanças dos usuários e do ambiente, a fim de melhorar a interação em tempo real. Desta forma, uma abordagem crescente é o uso de emoções na IHR. Neste sentido, existem os sistemas de reconhecimento de emoção multimodal, nos quais, realizam a classificação das emoções em várias modalidades (expressão facial, gestos, fala, e outros). No entanto, embora existam estudos que tratam do reconhecimento multimodal de emoções, eles ainda apresentam limitações na metodologia da classificação das emoções, além de considerar as emoções como binárias e ignorando as várias emoções que podem estar presentes no usuário. Assim, o objetivo deste trabalho foi propor um sistema de reconhecimento de emoções multimodal e multiclasse para a interação humano-robô. É proposto o uso das modalidades de expressão facial e fala, assim como a fusão das emoções. O Módulo de Reconhecimento de Emoção da Fala (MREF) é responsável por inferir a emoção na fala do usuário, no qual é utilizado um modelo de aprendizado profundo para classificar a emoção. Também é proposto o Módulo de Reconhecimento de Emoção da Expressão Facial (MREEF), que classifica a emoção pela face do usuário utilizando rede neural convolucional (CNN). Por fim, propõe-se a fusão das emoções reconhecidas utilizando sistema nebuloso. O sistema proposto utiliza da base de dados MELD, obtendo um resultado de 73% de acurácia usando apenas o MREF, 78,06% utilizando apenas o MREEF, e 78,94% de acurácia usando a fusão dos módulos.

Palavras-chave: Reconhecimento de Emoção. Sistema Multimodal. Interação Humano-Robô. Aprendizado de Máquina. Sistema Fuzzy.

Abstract

The challenge of Human-Robot Interaction (HRI) is to build intelligent systems that can adapt to user and environmental changes in order to enhance real-time interaction. In this regard, an emerging approach is the use of emotions in HRI. There are multimodal emotion recognition systems that classify emotions across various modalities (facial expression, gestures, speech, among others). However, despite studies on multimodal emotion recognition, they still have limitations in emotion classification methodology, often considering emotions as binary and overlooking the various emotions that may be present in the user. Therefore, the aim of this work is to propose a multimodal and multiclass emotion recognition system for human-robot interaction. The use of facial expression and speech modalities, as well as emotion fusion, is proposed. The Speech Emotion Recognition Module (MREF) is responsible for inferring the user's emotion from speech, utilizing a deep learning model for emotion classification. Additionally, the Facial Expression Emotion Recognition Module (MREEF) is proposed, which classifies the user's emotion from facial expressions using convolutional neural networks (CNNs). Finally, emotion fusion is proposed using fuzzy systems. When the proposed system was tested using the MELD database, the MREF achieved an accuracy of 73%, the MREEF 78.06%, and the fusion of modules achieved an accuracy of 78.94%. Thus, it can be observed that a multimodal system is more effective.

Keywords: Emotion Recognition. Multimodal System. Human-Robot Interaction. Machine Learning. Fuzzy System.

Lista de Figuras

Lista de Tabelas

Sumário

CAPÍTULO 1

Introdução

A detecção de emoções é um processo natural nas interações humanas, influenciando diretamente a tomada de decisões e as ações durante a comunicação. Quando aplicada à interação humano-robô (IHR), essa capacidade pode ser incorporada ao robô, permitindo que ele interaja com as pessoas de maneira mais natural e harmoniosa (Lan et al., 2020). Para isso, os robôs podem detectar a emoção dos seres humanos através da expressão facial (Kim e Lee, 2023a; Lu e Wan, 2023), fala (Gupta e Chandra, 2021; Hazra et al., 2022), gestos (Camada et al., 2021; Rad et al., 2018) e outros métodos (Filippini et al., 2020; Xu et al., 2018; Andreasson et al., 2018).

Os robôs sociais utilizam seu sistema de percepção para captar as informações necessárias para detectar emoções humanas. As principais fontes de percepção em um robô incluem as capacidades sensoriais visuais, auditivas e fisiológicas, entre outras (Heredia et al., 2022). Por meio da visão, os robôs podem capturar imagens e vídeos; com sua capacidade auditiva, podem perceber a fala; e, utilizando sensores fisiológicos, podem obter informações como temperatura corporal e frequência cardíaca. Com esses dados, é possível analisar emoções a partir de expressões faciais, fala, gestos, textos, temperatura corporal e outras formas (Mittal et al., 2020).

Como as emoções são fenômenos psicofisiológicos complexos associados a muitas pistas não verbais, é difícil construir modelos robustos de reconhecimento de emoções usando uma única modalidade (Lan et al., 2020). Desta forma, ao combinar os diversos modais é possível obter um sistema com maior assertividade na emoção do usuário, levando a uma abordagem multimodal (Heredia et al., 2022).

Embora existam estudos que tratam do reconhecimento de emoções multimodais para robôs

sociais (Tzirakis et al., 2021; Mittal et al., 2020), eles ainda apresentam limitações significativas na classificação das emoções. Em geral, a maioria dos trabalhos na literatura aborda as emoções de forma binária, identificando apenas a emoção predominante. No entanto, a percepção humana é subjetiva por natureza e não exata, o que torna a abordagem multiclasse mais adequada para a análise das emoções, semelhante a uma distribuição de probabilidade. Além disso, as abordagens de *deep learning* enfrentam desafios na otimização e ajuste dos modelos, uma vez que muitos estudos não exploram de forma abrangente as combinações entre arquitetura, hiperparâmetros, métodos de extração de características e outros fatores. Essa falta de investigação pode comprometer a precisão e a robustez do sistema (Ottoni et al., 2023b).

Nesse contexto, este trabalho propõe um sistema de reconhecimento de emoções multimodal e multiclasse para ser utilizado na interação humano-robô. O sistema integra as modalidades de fala e expressão facial, além de realizar a fusão das emoções detectadas. O Módulo de Reconhecimento de Emoções da Fala (MREF) emprega uma arquitetura híbrida para classificar as emoções expressas na fala. O Módulo de Reconhecimento de Emoções da Expressão Facial (MREEF) é responsável por inferir as emoções a partir das expressões faciais do usuário, utilizando uma CNN para a classificação. Por fim, a fusão das emoções será realizada por um sistema *fuzzy* do tipo Mamdani (Mamdani e Assilian, 1975), que combinará os dois módulos e apresentará as emoções multiclasse do usuário.

1.1 Contexto da Tese

A proposta desta tese faz parte do projeto desenvolvido pelo Laboratório de Robótica do Departamento de Engenharia Elétrica e de Computação da Universidade Federal da Bahia. O Hibot pretende ser uma plataforma para experimentos de interação humano-robô (IHR). O projeto, apresentado na Figura 1.1, demostra como se espera que o robô trabalhe futuramente.

A ideia é que o robô possua quatro módulos de reconhecimento de emoções: (1) Módulo de Reconhecimento da Emoção da Expressão Facial (MREEF), (2) Módulo de Reconhecimento de Emoção da Expressão Corporal (MREEC), (3) Módulo de Reconhecimento de Emoção da Fala (MREF) e (4) Módulo de Reconhecimento de Emoções por Eletroencefalograma (MREEG).

As informações dos sensores afetivos serão processadas por um cérebro artificial que possuirá emoção e personalidade. O cérebro artificial será formado a partir de uma arquitetura

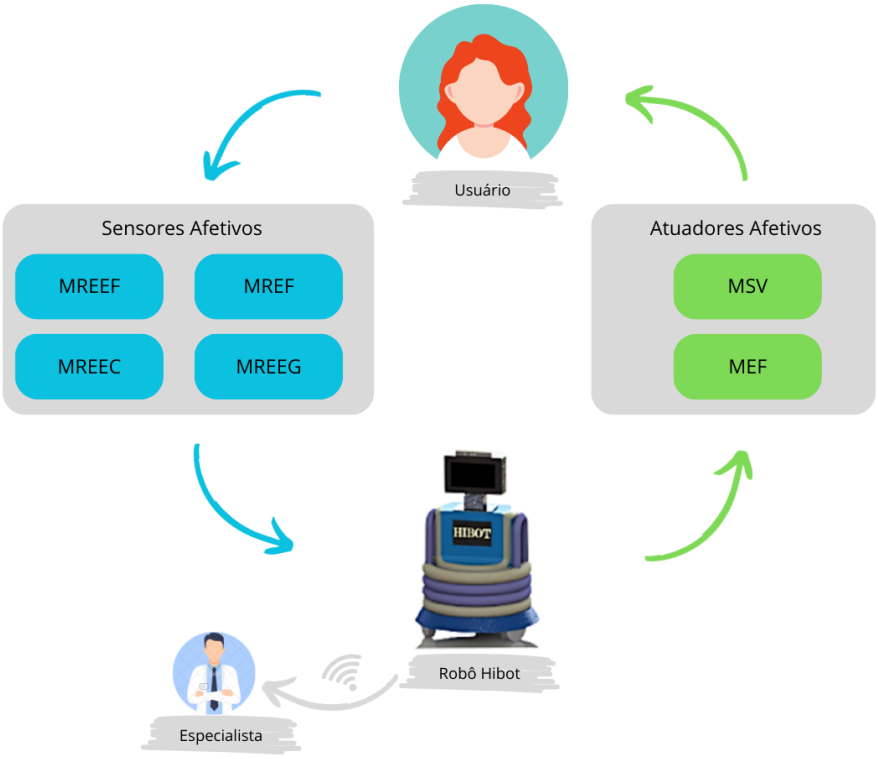


Figura 1.1: Esquema do funcionamento do projeto HiBot.

cognitiva que tomará as decisões com base na emoção e personalidade do usuário e também na personalidade do próprio robô.

Além disso, o robô possuirá comunicação com profissionais (terapeutas ou responsáveis), conforme mostrado pela Figura 1.1. Por fim, o robô Hibot poderá demostrar suas emoções através de dois atuadores afetivos: (1) Módulo de Síntese de Voz (MSV) e (2) Módulo de Expressão Facial (MEF). Além do mais, o rosto do robô é composto por uma tela, na qual será possível realizar jogos, passar vídeos e uma série de brincadeiras com o usuário.

1.2 Objetivos

O objetivo deste trabalho foi desenvolver um sistema de reconhecimento de emoções multimodal, que combina expressão facial e fala, utilizando *deep learning* para classificar as emoções do usuário durante a interação humano-robô. Além disso, propõe-se a utilização de um sistema *fuzzy* do tipo Mamdani para a fusão das emoções detectadas, integrando as informações multimodais e multiclasse.

1.2.1 Objetivos Específicos

Os objetivos específicos desse trabalho são:

1. Propor uma abordagem de *deep learning* utilizando CNN e LSTM, para o módulo de reconhecimento de emoção da fala, otimizando a combinação dos diferentes parâmetros da classificação das emoções na fala do usuário;

2. Propor uma abordagem de *deep learning* utilizando CNN, para o módulo de reconhecimento de emoção da expressão facial utilizando redes neurais convolucionais para classificação da emoção multiclasse da face do usuário;

3. Propor a fusão das emoções de fala e expressão facial utilizando o sistema *fuzzy* para inferência das emoções do usuário.

1.3 Contribuições

Com isso, está tese tem como contribuições principais: (i) Propor o Módulo de Reconhecimento de Emoção da Fala; (ii) Propor o Módulo de Reconhecimento de Emoções da Expressão Facial; (iii) Propor a realização da fusão dos módulos de reconhecimento MREF e MREEF utilizando sistema *fuzzy*; e o (iv) Tratamento das emoções de forma multiclasse, ou seja, trataremos de porcentagens de emoções (não-binário).

Até o presente momento, o estudo desenvolvido por esse doutorado contribuiu com a literatura da área com as seguintes publicações:

- Ottoni L. T. C., Ottoni A. L. C., Cerqueira J. J. F. A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta-Learning. Electronics - 2023.

- Ottoni, L. T. C. e Cerqueira, J. J. F. Human-Robot Interaction of Emotion: A Systematic Review and Future Directions. International Journal of Social Robotics - 2024.

- Ottoni, L. T. C., Oliveira, M., Cerqueira, J. J. F. e Simas Filho, E. Perceptive Artificial Hearing Module for User Interface in Socially Interactive Robots: Speaker Identification, Emotion Classification, and Spatial Localization. Autonomous Robots - 2024 (submetido).

- Ottoni, L. T. C. e Cerqueira, J. J. F. A Review of Emotions in Human-Robot Interaction. XVIII Latin American Robotics Symposium (LARS) - 2021.

- Ottoni, L. T. C. e Cerqueira, J. J. F. Optimizing Speech Emotion Recognition: Evaluating Combinations of Databases, Data Augmentation, and Feature Extraction Methods. XVI Brazilian Congress on Computational Intelligence (CBIC) - 2023.

1.4 Organização do Texto

Os assuntos discutidos neste trabalho estão organizados conforme mostrado a seguir.

Capítulo 2 - Robótica Social: Neste capítulo são apresentados as informações sobre robótica social. É discutido sobre a ética na robótica e em seguida abordado conceitos sobre interação humano-robô com uso de emoções e personalidade.

Capítulo 3 - Fundamentação Teórica: Neste capítulo é apresentado a teoria de redes neurais convolucionais e sistema *fuzzy*.

Capítulo 4 - Módulo de Reconhecimento de Emoção da Fala: Neste capítulo é apresentado com detalhes o sistema de reconhecimento de emoção da fala, é apresentado uma revisão de literatura sobre o tema, a metodologia proposta e os resultados obtidos.

Capítulo 5 - Módulo de Reconhecimento de Emoção da Expressão Facial: Neste capítulo é apresentado com detalhes o sistema de reconhecimento de emoção da expressão facial, é apresentado uma revisão de literatura sobre o tema, a metodologia proposta e os resultados obtidos.

Capítulo 6 - Fusão das Emoções: É apresentado a metodologia da escolha das funções de pertinência do sistema *fuzzy* e os resultados obtidas da fusão.

Capítulo 7 - Considerações Finais: Por fim, são apresentados as considerações finais do trabalho e propostas de continuidade.

A tese foi estruturada de forma a proporcionar uma compreensão clara e detalhada de cada módulo de reconhecimento investigado, optando-se por separar esses módulos em capítulos distintos. Dessa forma, o Capítulo 4 e o Capítulo 5 são dedicados, cada um, a um módulo

específico de reconhecimento e o Capítulo 6 dedicado a fusão das emoções. Em vez de consolidar toda a metodologia em um único capítulo e os resultados em outro, preferiu-se apresentar a metodologia e os resultados dentro de cada capítulo, individualmente. Essa abordagem permite uma análise mais profunda e focada, na qual é possível compreender diretamente como a metodologia aplicada levou aos resultados obtidos para cada módulo de reconhecimento.

Além disso, optou-se por utilizar algumas palavras em inglês, como *data augmentation*, *meta-learning*, *convolutional neural network*-CNN e *long short-term memory*-LSTM, *fuzzy* e outras palavras. A escolha se deu devido à sua ampla aceitação e uso frequente entre os leitores brasileiros da área.

CAPÍTULO 2

Robótica Social

2.1 Introdução

A Robótica Social (RS) é o ramo da robótica em que o robô passa a interagir de forma mais próxima dos seres humanos (Belo et al., 2017). Isto é, envolve os robôs que realizam Interação Humano-Robô (IHR) por meio de fala, gestos, atividades ou outras mídias (Scassellati et al., 2012).

Na robótica social é possível classificar os robôs de diversas formas conforme: o objetivo, aparência, capacidades do robô, autonomia, locomoção, etc (Romero et al., 2014). Por exemplo, os robôs sociais podem ser: autônomos ou tele-operados; fixos, aquáticos, terrestres ou aéreos; aparência humanoide ou não humanoide. Além disso, o robô social pode ter o objetivo de realizar busca e resgate ou ter o objetivo de divertir os usuários. Enfim, são muitas as aplicações da robótica social (Boada, 2021).

Para um melhor entendimento desta área de pesquisa nas seções a seguir serão tratados os conceitos importantes de ética na robótica. Posteriormente, serão apresentado os conceitos de robótica socialmente assistiva e o conceito de interação humano-robô com o uso de emoções e em seguida de personalidade.

2.2 Ética na Robótica

A robótica do século passado utilizava os robôs, na grande maioria das vezes, no âmbito industrial. Os grandes manipuladores robóticos se moviam nas linhas de montagens das empresas multinacionais dentro de grades ou "jaulas", para garantir a segurança dos trabalhadores.

Com o desenvolvimento da tecnologia e da inteligência artificial, os robôs foram migrando das indústrias e aproximados cada vez mais do cotidiano das pessoas (Boada et al., 2021).

Hoje em dia, no século XXI, existem alguns exemplos de robôs que estão presentes no dia-a-dia da população. A inteligência artificial chamada Alexa (Amazon) é um dos exemplos. Podendo ser utilizada como uma assistente virtual, a Alexa é conectada na internet e é capaz de responder diversas perguntas, tocar músicas e realizar jogos para interagir com os usuários. Outro exemplo, é o robô aspirador de pó que realiza o mapeamento inteligente do ambiente para limpa-lo e é capaz de se recarregar de forma autônoma. E mais recentemente, podemos citar o ChatGTP [1] que é capaz de responder de forma natural e rápida a diversos questionamentos.

Neste sentido, os pesquisadores da robótica social buscam desenvolver ferramentas centradas no ser humano, que visam priorizar o indivíduo para melhorar sua qualidade de vida. No entanto, algumas questões relevantes de ética em inteligência artificial e robótica vem sendo levantadas nos últimos anos. Algumas delas são: quais as limitações da interação humano-robô? As pessoas podem insultar uma inteligência artificial e quais são as consequências dessa atitude para o ser humano? Ele se tornará mais agressivo nas interações humano-humano? Quais as consequências da IHR nos seres humanos? E se um ser humano tiver laços afetivos por um robô? Em fim, são muitas as questões que envolvem a ética na robótica social.

As primeiras diretrizes de ética para a robótica foi proposta por Isaac Asimov em sua ficção *"Runaround"*, e ficaram internacionalmente conhecidas a partir da literatura e filmes. As três leis da robótica de Asimov são (dos Reis Alves, 2016):

1. Um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano sofra algum mal;

2. Um robô deve obedecer as ordens que lhe sejam dadas por seres humanos, exceto nos casos em que tais ordens entrem em conflito com a Primeira Lei;

3. Um robô deve proteger sua própria existência desde que tal proteção não entre em conflito com a Primeira e Segunda Lei.

No entanto, essas leis contribuem pouco para a elaboração de uma regulação que limite as atitudes do robô enquanto agente e também as pessoas e empresas envolvidas no desenvolvimento desses robôs. Neste sentido, há um esforço crescente de diversas instituições de elaborar

[1] https://chat.openai.com/

uma legislação que proteja a integridade e privacidade das pessoas. Alguns exemplos de instituições dedicadas as questões éticas são as: *European Robotics Research Network* (EURON), *Open Roboethics* e o comitê técnico para ética robótica da IEEE (Boada et al., 2021).

2.3 Robótica Socialmente Assistiva

A Robótica Socialmente Assistiva (RSA) é uma área nova, no entanto tem se desenvolvido rápido. As características sociais dos robôs são particularmente importantes porque, ao contrário das aplicações típicas da robótica social, os robôs socialmente assistivos tem como objetivo ajudar o usuário. Portanto, devem ser capazes de motivar e influenciar a mudança de comportamento (Cano et al., 2021).

Um exemplo de robô socialmente assistivo é o robô Matilda (Khosla et al., 2013), apresentado na Figura 2.1 [2]. Matilda é capaz de desempenhar uma série de tarefas para melhorar a saúde emocional de pessoas em clínicas dedicadas a cuidado de idosos. Para isso, o robô contém um sistema de detecção de emoção e cuida de forma personalizada de cada idoso. Assim, ao detectar que o paciente está triste, sugere jogos e até mesmo serviços de telefonia para combater o isolamento social. Outro exemplo é o robô Gerda (Landowska, 2013) voltado para o ambiente pedagógico. O robô Gerda contém um sistema de percepção de emoções de estudantes e a partir das informações extraídas dos alunos oferece uma intervenção para otimizar o processo de aprendizagem.

Além disso, o uso da robótica social assistiva para terapia de criança com TEA (transtorno do espectro autista) tem ganhado cada vez mais espaço e alcançado resultados positivos (Coeckelbergh et al., 2016; Dautenhahn e Billard, 2002; Duquette et al., 2008; Goulart et al., 2014; Iacono et al., 2011; Kim et al., 2012b, 2013; Michaud e Clavet, 2001; Robins et al., 2005, 2010). Acredita-se que a natureza artificial e previsível dos robôs os tornam facilmente compreensíveis para as crianças com TEA, ajudando a estimular as habilidades sociais (Duquette et al., 2008), uma vez que eles podem utilizar da fala, sons, elementos visuais e gestos para estabelecer uma comunicação significativa (Michaud e Clavet, 2001).

Esse êxito das aplicações de robótica social assistiva com crianças TEA tem incentivado diversas pesquisas. Dentre diversos trabalhos desenvolvidos, destaca-se o robô ROBUS (Michaud e Clavet, 2001), Keepon (Kozima et al., 2009a), PLEO (Kim et al., 2013), KASPAR (Wainer

[2] Disponível em: https://www.smh.com.au/

Figura 2.1: Robô Socialmente Assistivo Matilda, utilizado para melhorar a saúde emocional de idosos (Khosla et al., 2013).

et al., 2014), PROBO (Simut et al., 2016), ROBOTA (Robins et al., 2004), MARIA (Valadão et al., 2016) e Pomodoro (dos Reis Alves e Ferasoli Filho, 2016).

O ROBUS (*Robot of University of Sherbrooke*), usado para pesquisas de crianças com TEA no Canadá é um exemplo desses robôs. Seu objetivo é ajudar crianças autistas a se abrirem para o ambiente, melhorar a imaginação e experimentar padrões de comportamento menos repetitivos. O ROBUS possui diversas fantasias, o que lhe dá a possibilidade de interagir de forma lúdica com as crianças (Michaud e Clavet, 2001).

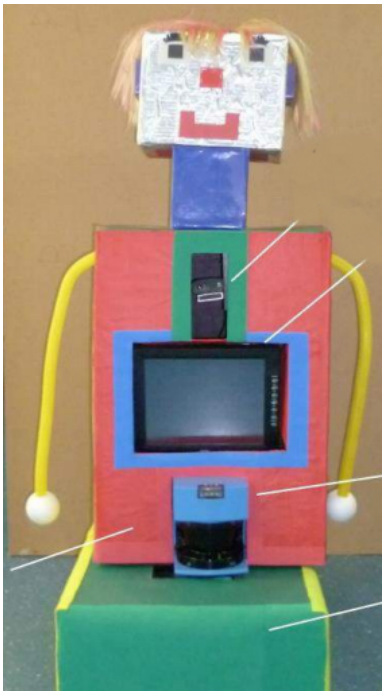
O robô PLEO, construído em formato de dinossauro, conforme visto na Figura 2.2(a), foi desenvolvido para realizar a comunicação entre crianças diagnosticadas com TEA com adultos, como por exemplo os pais, terapeutas, professores, etc. O estudo realizado por Kim et al. (2013) demonstrou que crianças autistas se comunicam com maior facilidade com o robô PLEO do que diretamente com um adulto.

Em Wainer et al. (2014), o autor utiliza o robô KASPAR (*Kinesics and Synchronization in Personal Assistant Robotics*), visto na Figura 2.2(b). O robô opera de forma totalmente autônoma e usa informações sobre o estado do jogo e o comportamento das crianças para envolver, motivar, incentivar e aconselhar pares de crianças jogando um jogo de imitação.

O robô MARIA (*Mobile Autonomous Robot for Interaction with Autistics*) (Valadão et al., 2016; Valadão, 2016), Figura 2.2(c), tem como objetivo melhorar a interação com crianças autistas para que desta forma possam propor novos tratamentos para o transtorno de espectro autista. O robô MARIA consiste de um robô móvel que contém um tela para realizar a interação com a criança por meio de vídeos.

(a) Robô Pleo (Kim et al., 2013).

(b) Robô Kaspar (Wainer et al., 2014).



(c) Robô Maria (Valadão, 2016).

(d) Robô Keepon (Kozima et al., 2009b).

Figura 2.2: Exemplos de robôs socialmente assistivos utilizados para interagir com crianças que possuem o transtorno do especto autista.

Outro exemplo é o robô Keepon (Kozima et al., 2009a) (Figura 2.2(d)). Para chamar a atenção de crianças com TEA esse robô conta com um sistema que reage aos estímulos do ambiente. Ao ser tocado reage com sons, luzes e até movimentos divertidos. Nesse aspecto, esse sistema se mostrou eficaz para provocar uma motivação para compartilhar estados mentais das crianças.

Já em Simut et al. (2016) é usado o robô PROBO para imitar interações sociais que os humanos teriam com as crianças. Aspectos como o olhar e a atenção, dentre outras variáveis como atenção conjunta, fala, afetividade positiva eram analisadas, porém apenas o olhar apresentou maior diferença nas crianças com TEA, enquanto as outras variáveis se mantiveram sem grande diferença.

Em Robins et al. (2004) o robô ROBOTA é utilizado para promover a comunicação entre crianças com TEA e terapeutas. Com o tempo, crianças se sentiram mais confortáveis com o robô, que tem formato de boneca, e procuraram interagir mais e compartilhar experiências com os adultos.

E por fim, o robô Pomodoro (dos Reis Alves, 2016) é composto por um robô de baixo custo controlado com um smartfone incorporado. O objetivo do robô é interagir com crianças autistas por meio de controle da fala, reconhecimento de imagem e toque, além de um subsistema complementar de rastreamento de movimento manual para tele-operar o sistema usando o sistema real e virtual.

2.4 Interação Humano-Robô com Emoções

A interação humano-robô (IHR) é um campo de estudo recente, que vem ganhando cada vez mais espaço na literatura. Além disso, a IHR busca compreender, desenvolver e aperfeiçoar a forma com o qual humanos e robô interagem (Picard, 2003). O campo de estudo IHR surgiu em 1990 e é um área de estudo multidisciplinar que envolve pesquisadores da robótica, ciências cognitivas, linguagem natural, psicologia, etologia entre outros (Goodrich e Schultz, 2008; Berry, 2015).

Os estímulos para crescimento da área de IHR são muitos. Seus principais desafios podem ser agrupados em cinco atributos (Goodrich e Schultz, 2008):

Natureza da tarefa A natureza da tarefa diz respeito aos desafios relacionados ao alcance dos

objetivos da interação humano-robô (Goodrich e Schultz, 2008; dos Reis Alves e Fera-soli Filho, 2016). Neste segmento existem diversos estudos já desenvolvidos, como por exemplo, na área da saúde (Johnson et al., 2014; Remazeilles et al., 2008; Valadão et al., 2016), entretenimento (Koch et al., 2017; Arkin et al., 2003), tarefas domesticas (Nieuwe-nhuisen e Behnke, 2013; Koay et al., 2009), educação (Kozima et al., 2009b; Berry, 2015).

Nível de autonomia O nível de autonomia do robô se refere a sua dependência com o humano, quanto mais independente é o robô maior seu nível de autonomia. Desafios ligados a esse atributo envolvem a elaboração de robôs com as habilidades cognitivas apropriadas para interagir de forma natural com o ser humano (Goodrich e Schultz, 2008).

Estrutura da equipe A estrutura da equipe está voltada para interações que não se restringem a apenas um humano e um robô. Em geral, apresenta-se em quatro categorias (Burke et al., 2004): uma pessoa e um robô, uma pessoa e muitos robôs, muitas pessoas e um robô e muitas pessoas e muitos robôs (dos Reis Alves, 2016). Desafios nesta área envolvem: relações de liderança, i.e., quais as tarefas que cada robô e humano irá desempenhar e o grau de dependência entre eles (Howard e Cruz, 2006).

Comunicação A comunicação (verbal e não-verbal) é a forma que se permite a troca de informações dentro da IHR (Kozima et al., 2009b). Um dos desafios mais recentes desta área envolve realizar uma comunicação natural e amigável, fazendo com que o robô expresse suas emoções de diversas maneiras. Algumas dessas pesquisas buscam demostrar emoções na fala (James et al., 2020; Crumpton e Bethel, 2016; Read e Belpaeme, 2016), nos gestos (Law et al., 2020; Rosenthal-von der Pütten et al., 2018; McColl e Nejat, 2014), na forma em que o robô anda (Jerčić et al., 2018a; Li e Chignell, 2011), nas expressões faciais (Liu et al., 2013; De Beir et al., 2016; Reyes et al., 2019).

Adaptação e aprendizagem A adaptação e aprendizagem se preocupa com a capacidade do robô de se adaptar automaticamente a diversas situações durante a interação (Goodrich e Schultz, 2008). Alguns exemplos nesta área são: os problemas de navegação autônoma (Bajracharya et al., 2008), detecção de novidade (Bove et al., 2020), o robô seguir a pessoa com o olhar (Hoffman et al., 2006), etc.

Desta forma, uma abordagem crescente é o uso de emoções na interação humano-robô. As emoções aumentam a adaptação e cria uma relação amigável, do ponto de vista humano,

entre pessoas e robôs (Rairán e Nino, 2017). Na IHR é essencial que além do robô reconhecer emoções ele seja capaz de transmitir emoções tornando a interação o mais natural possível (Kozima et al., 2009b).

O conceito de emoção é motivo de muitas divergências entre os pesquisadores devido ao seu caráter subjetivo. As emoções representam uma resposta afetiva de curto prazo e tendem a decair rapidamente quando sua causa é removida. Sua intensidade depende da emoção específica provocada, humor atual e personalidade (Ortega et al., 2020). Uma vez que o humor é uma resposta afetiva de médio prazo e a personalidade é de longo prazo.

A noção de emoção básica foi proposta pelos psicólogos norte-americanos Paul Ekmam e Wallace Friesen e é frequentemente associada a simulações de expressões faciais (Ekman e Friesen, 1986). Essas emoções são inatas e compartilhadas por todas as culturas. São eles: tristeza, raiva, felicidade, medo, desgosto e surpresa. Na literatura, é o modelo mais comum utilizado na pelos trabalhos da área de interação humano-robô.

Nesse sentido, foi realizada uma revisão de literatura sistemática na base de dados da SCOPUS, utilizando-se as palavras chaves: “human-robot interaction” e “emotion”. A seguir, serão apresentadas algumas formas encontradas na literatura para o reconhecimento e expressão de emoções na IHR.

2.4.1 Reconhecimento de emoções humanas

O reconhecimento de emoções humanas é uma etapa importante no processo da interação humano-robô, pois permite que o robô adapte suas ações de acordo com a emoção do usuário. A seguir serão apresentadas algumas formas de detecção de emoções encontradas na literatura, como: o reconhecimento de emoções nas expressões faciais (Bagheri et al., 2020; Menne e Schwab, 2018; Ke et al., 2020), na fala (Johnson et al., 2014; Sajjad et al., 2020; Devillers et al., 2015), no toque (Yohanan e MacLean, 2012; Andreasson et al., 2018; Silvera-Tawil et al., 2014), expressões corporais (Goodwin et al., 2011; Rad et al., 2018; Camada et al., 2021) e em sinais fisiológicos (Filippini et al., 2020; Xu et al., 2018; Yang e Dorneich, 2017).

Reconhecimento de expressões faciais

O ser humano possui a habilidade de expressar suas emoções através das expressões faciais. Conforme mostrado na Figura 2.3, as expressões mais comuns de serem demonstradas pelo

Figura 2.3: Exemplo das emoções demostradas pela expressão facial. Fonte: banco de dados WSEFEP (Olszanowski et al., 2015).

ser humano no cotidiano são a felicidade, tristeza, surpresa, medo, desgosto e raiva. Devido a importância do impacto da expressão facial na interação social existem diversos trabalhos na literatura que desenvolvem ferramentas e métodos para realizar o reconhecimento de emoções pela expressão facial humana. De forma geral, eles realizam estudos utilizando câmeras e algoritmos de aprendizado de máquina para que o robô seja capaz de detectar as emoções humanas de forma assertiva e eficiente.

Neste sentido, o robô humanoide Pepper (Bagheri et al., 2020; Val-Calvo et al., 2020) é um exemplo de utilização do reconhecimento de emoções humanas através da face. Em Bagheri et al. (2020) os autores utilizaram o aprendizado por reforço em um experimento para que o robô identifique emoções de 28 humanos. Ao reconhecer a emoção, o robô Pepper é capaz de proporcionar conforto e ajudar a pessoa a se sentir melhor através de uma tela interativa.

Já em Menne e Schwab (2018) e Rosenthal-von der Pütten et al. (2013) o robô dinossauro Pleo foi utilizado para estudar as expressões faciais dos humanos. O experimento consistia em 62 participantes assistindo um vídeo do robô Pleo sendo abraçado. Em um segundo momento, os participantes assistiam vídeos do robô sendo agredido. Com isso, foram captadas as faces dos participantes e realizado o reconhecimento das emoções predominante, que foram de felicidade e raiva.

O trabalho de Ke et al. (2020) é um outro exemplo em que o robô humanoide SHFR-II apresenta um modelo de reconhecimento emocional baseado na expressão facial e na fala. São

utilizados os dois módulos para lidar com situações em que um único modal falharia. Ao mesmo tempo, o algoritmo *fuzzy* é usado para simular a tomada de decisão emocional. Outros exemplos podem ser encontrados em Xu et al. (2018); Jerčić et al. (2018b); Basori (2013).

Reconhecimento de emoção da fala

A fala é a forma mais natural de comunicação entre humanos-humanos e também entre humanos-robôs. Para uma comunicação eficaz é preciso que os robôs entendam as verdadeiras intenções dos falantes. A detecção de emoção pela fala é um campo crescente na IHR, em geral, são utilizados fontes sensoriais de áudio, como microfones, para captar fala e sons não linguísticos. Em seguida são utilizados algoritmos de aprendizado de máquina para extrair as emoções de alegria, tristeza, medo, raiva, desgosto e surpresa. Neste sentido, alguns trabalhos da literatura vem explorando a detecção de emoções humanas na fala, como é o caso dos trabalhos de Devillers et al. (2015); Peng et al. (2020); Sajjad et al. (2020); Prado et al. (2012); Gupta e Chandra (2021); Asiya e Kiran (2021); Hazra et al. (2022).

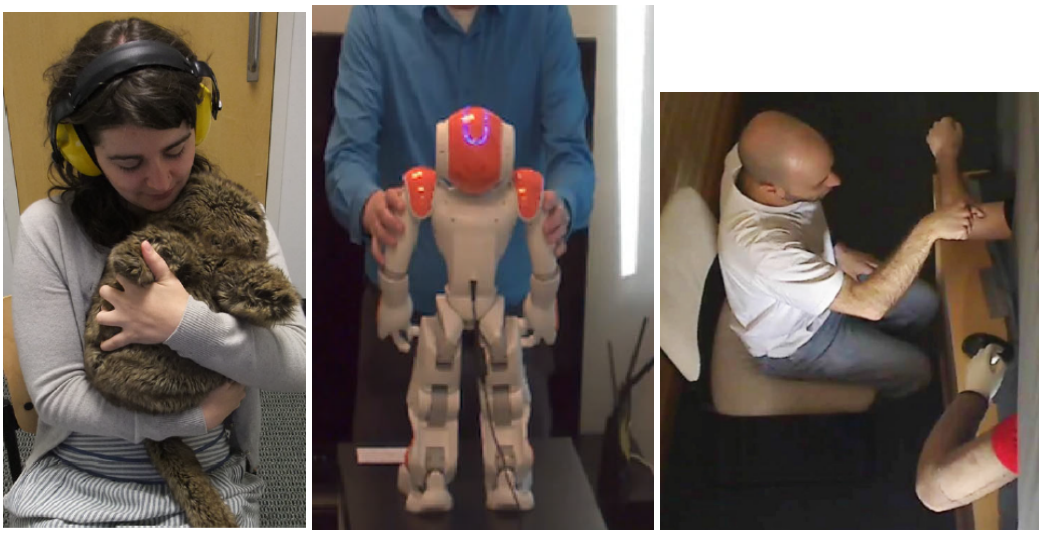
Reconhecimento de emoção no toque

O toque afetivo é um elemento crucial do desenvolvimento humano nos laços sociais e do suporte emocional (Eid e Al Osman, 2015). É considerada uma técnica difícil de estudar, e por isso, tem recebido pouca atenção da pesquisa (Eid e Al Osman, 2015; Andreasson et al., 2018).

Os robôs da Figura 2.4 são exemplos do reconhecimento do toque afetivo dos humanos. Em (a) o robô Haptic Creature (Yohanan e MacLean, 2012), (b) robô NAO (Andreasson et al., 2018) e (c) braço de manequim de tamanho real (Silvera-Tawil et al., 2014). Essas pesquisas buscam compreender o toque humano e futuramente inserir essa capacidade no robô. Em geral, são utilizados dispositivos sensíveis ao toque humano para captar as emoções expressadas.

Reconhecimento por expressões corporais

Outra forma importante de reconhecer as emoções em humanos é através da expressão corporal e gestos (Camada et al., 2021; Goodwin et al., 2011). A partir dos gestos é possível identificar emoções como: medo, alegria, tristeza e outras. De forma geral, são utilizados sensores de movimento e algoritmos de aprendizado de máquina. Em Vu et al. (2011) o robô



(a) Robô Haptic Creature. (b) Robô NAO. (c) Robô Manequim.

Figura 2.4: Exemplos robôs que reconhecem emoções humanas através do toque.

Mascote identifica as emoções a partir de 8 tipos de gestos que são registrados por participantes da pesquisa.

Além disso, alguns gestos estereotipados podem ajudar a entender as emoções do ser humano. Um exemplo são as pessoas que têm transtorno do espectro do autismo (TEA) (Rad et al., 2018). Em Camada et al. (2021) o robô HiBot identifica a emoção das crianças com TEA por meio de gestos. Outros exemplos são Goodwin et al. (2011); Rad et al. (2018); Sadouk et al. (2018).

Reconhecimento por sinais fisiológicos

A pressão arterial, frequência respiratória, temperatura corporal e pulsação são alguns exemplos de sinais fisiológicos que podem indicar as emoções humanas. Um tipo de robô que utiliza os sinais fisiológicos para reconhecer a emoção humana é o robô educacional Mio Amico (Filippini et al., 2020). Ele utiliza de um sensor infravermelho térmico capaz de realizar a leitura das atividades neuro-vegetativa periférica de crianças que estão interagindo. A partir da leitura do sensor é realizado uma decodificação do estado emocional.

Outro exemplo é o trabalho de Xu et al. (2018) que, por sua vez, propõe um sistema de reabilitação de pacientes que sofreram AVC e são assistidos por robótica. É utilizado um sistema de reconhecimento do nível de ansiedade do paciente por meio da leitura dos sinais fisiológicos. Os participantes foram equipados com biossensores para monitorar as atividades do sistema

nervoso relacionadas à ansiedade, incluindo eletromiograma facial, eletrocardiograma, condutância da pele e respiração.

Já em Yang e Dorneich (2017) foram realizados experimentos para estudar a influência do *delay* nas emoções humanas. Os participantes do experimento foram convidados a navegar em um veículo robótico teleguiado, com um atraso na resposta, por um labirinto. Foi utilizado um sensor EDA para mostrar a excitação emocional, uma vez que pode medir a ativação do sistema nervoso simpático. Os resultados apontaram que a raiva e frustração foram as emoções predominantes.

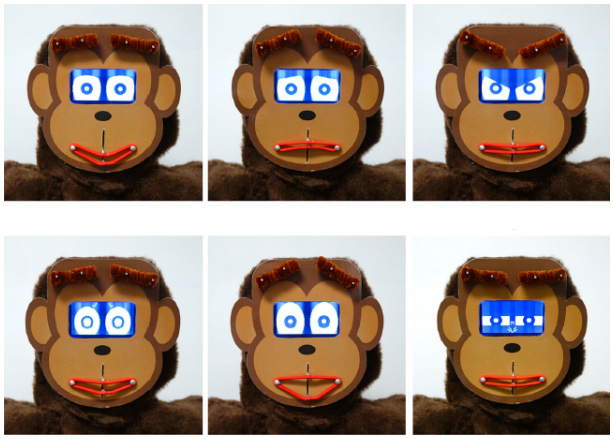
O reconhecimento de emoções através de sinais fisiológicos tem sido utilizado também no mundo virtual. Os avatares de jogos como *Serious Game* (Jerčić et al., 2018b) se comportam de maneiras diferentes conforme a emoção do jogador. Em Basori (2013) é utilizado um método para verificar a tensão muscular facial, sinais cerebrais do jogador. Desta forma a expressão facial e o andar do avatar mudam conforme a emoção do jogador.

2.4.2 Emoções dos Robôs

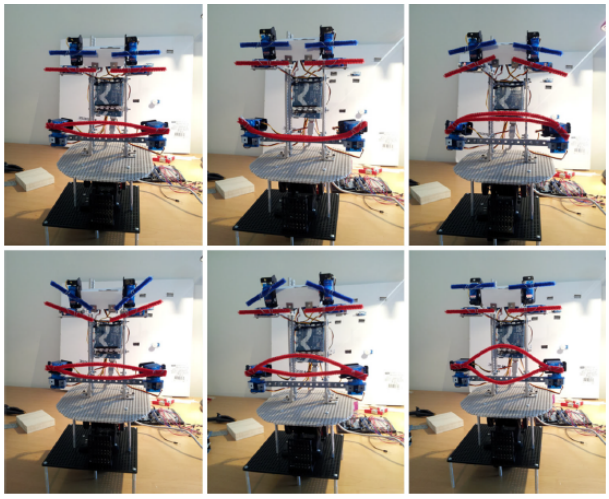
Assim como na iteração humano-humano há uma troca de expressões que demostram a emoção ao longo da conversa, é importante que os robôs sejam equipados de emoções na IHR. Desta forma, a interação se torna mais natural, aumentando a aceitabilidade dos robôs sociais e tornando o processo amigável. Em seguida são apresentados trabalhos que trazem emoções nas expressões faciais do robô (Saldien et al., 2010; Johnson et al., 2014), na linguagem corporal (Banik et al., 2013; Jerčić et al., 2018a), na fala (Tsiourti et al., 2019; Crumpton e Bethel, 2016) e na aparência (Reich-Stiebert et al., 2019; Wang e Huang, 2012; Lee et al., 2012).

Expressões Faciais dos Robôs

O robô macaco SAM (Koch et al., 2017), Figura 2.5(a), é capaz de expressar seis emoções faciais, são elas: felicidade, tristeza, raiva, medo, surpresa e desgosto. Essas expressões trazem maior naturalidade e sutileza na interação com crianças de desenvolvimento típico e crianças portadoras do transtorno do especto autista (TEA). Para isso SAM é equipado com um *display* LCD para exibir os olhos e está equipado com quatro pequenos motores que controlam as sobrancelhas e a boca. Suas expressões faciais são controladas por um microcontrolador Arduino Mega 2560.



(a) Robô SAM(Koch et al., 2017).



(b) Robô Mirae (Bennett e Šabanović, 2014).

(c) Robô Probo (Saldien et al., 2010).

Figura 2.5: Exemplos emoções transmitidas pelos robôs através de expressões faciais.

Outro exemplo de robô social para crianças é o robô Probo, Figura 2.5(c). Em Saldien et al. (2010) é realizado um estudo para verificar se as emoções de Probo são reconhecidas pelas crianças, pois o reconhecimento das expressões faciais são muito importantes para estabelecer uma boa comunicação não verbal entre um humano e um robô.

O robô NAO também é bastante utilizado na interação humano-robô. No entanto, sua característica física não permite a expressão de emoções faciais como por exemplo, sorrir ou mover a sobrancelha. Devido a essa dificuldade, nos trabalhos Johnson et al. (2013, 2014) foi analisado a capacidade de transmitir emoções apenas alterando a cor dos olhos do robô NAO, composto por um LED RGB. Já no trabalho de De Beir et al. (2016), os autores acrescentaram ao robô um dispositivo de sobrancelha articulada, fazendo com que o robô expresse raiva ou tristeza.

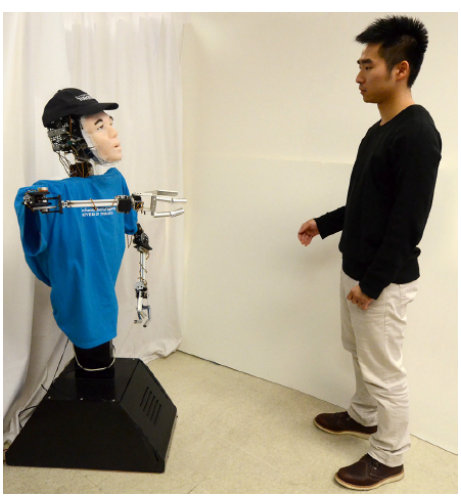
Diversos trabalhos abordaram um estudo de emoções de cabeças interativas robóticas (Bennett e Šabanović, 2014; Kühnlenz et al., 2013; Jiang et al., 2013; Liu et al., 2013; Pais et al., 2013; Reyes et al., 2019; Nieuwenhuisen e Behnke, 2013) em que foram submetidos a uma série de experimentos com seres humanos para verificar o reconhecimento das emoções do robô, como mostrado na Figura 2.5(b) pelo robô Mirae (Bennett e Šabanović, 2014). A face robótica era capaz de expressar: felicidade, tristeza, raiva, medo, surpresa e desgosto na face e no movimento com o pescoço.

E por fim, no artigo de Trovato et al. (2013), os autores se preocupam com o fato das expressões faciais serem manifestadas de formas diferentes em determinadas culturas. Com isso, o robô humanoide KOBIAN-R pode expressar duas versões da mesma emoção reconhecível por japoneses e ocidentais.

Linguagem corporal dos Robôs

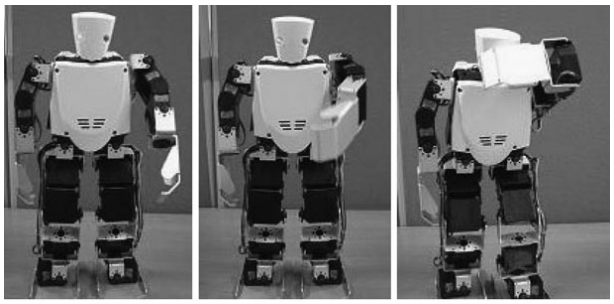
A linguagem corporal pode ser uma ferramenta importante para que os robôs realizem as interações com humanos (Banik et al., 2013; Martín et al., 2017). Há evidências de que as pessoas podem interpretar o comportamento não verbal em entidades artificiais (Jerčić et al., 2018a). Uma vez que os humanos são orientados empática e emocionalmente, eles são propensos a reagir também emocionalmente ao comportamento emocional não verbal artificial (Rosenthal-von der Pütten et al., 2018).

Já existem vários estudos na literatura que buscam a melhor forma de transmitir emoções através da linguagem corporal, como mostrado na Figura 2.6. No trabalho de McColl e Nejat



(a) Robô Brian 2.0 (McColl e Nejat, 2014).

(b) Robô Puppeteers (Li e Chignell, 2011).



(c) Robô RobovieX (Nomura e Nakao, 2010).

Figura 2.6: Exemplos de robôs que expressam as emoções através da expressão corporal.

(2014), o robô Brian 2.0 explora sua capacidade corporal para demostrar emoções através de gestos. Em Law et al. (2020), o mesmo acontece com o robô Cozmo. Em Li e Chignell (2011) é estudada a capacidade de como um robô com movimentos restritos (apenas cabeça e braços) pode transmitir emoções pela linguagem corporal. E em Richardson et al. (2012) é pesquisado qual o melhor número de graus de liberdade para um robô atingir uma linguagem corporal mais realista. Em Johnson et al. (2016), por sua vez, o robô NAO expressa suas emoções utilizando gestos ao brincar com humanos.

Outra vertente das pesquisas é a busca em associar a linguagem corporal com outras formas de expressar emoções. Em Wang e Huang (2012), linguagem corporal e linguagem verbal são controladas utilizando o algoritmo genético, desta forma é possível combinar discursos com gestos de forma automática. Outro trabalho que trata do mesmo tema é apresentado em Claret et al. (2017), em que o robô Pepper é capaz de conciliar gestos com outras atividades (ex.: carregar caixas e objetos pesados).

Comportamento verbal dos Robôs

Existe uma preocupação crescente com relação a voz do robô social (Crumpton e Bethel, 2016). Isso porque, é importante que ela expresse afeto e confiança ao usuário. Com essa motivação os trabalhos desenvolvidos por James et al. (2020); Tsiourti et al. (2019); Crumpton e Bethel (2016) buscam melhorar a voz do robô para que ela seja mais simpática. Em James et al. (2020) o robô Healthbot interage com pacientes de um hospital e são investigadas as emoções necessárias para uma voz empática. Em Tsiourti et al. (2019) são incorporadas no estudo informações da psicologia, neurociência e interação humano-computador para examinar como as pessoas respondem a emoções pela voz do robô. E em Crumpton e Bethel (2016) traz pesquisas informando que as mudanças na prosódia vocal é uma forma de tornar a voz sintética mais natural e amigável.

Em outra vertente, Read e Belpaeme (2016) investigam se seres humanos adultos conseguem identificar emoções nas elocuções não linguísticas (NLUs) dos robôs. As NLUs podem ser definidas como sons do tipo bipes e zumbidos. Seus resultados indicaram que é possível transmitir afeto através de elocuções não linguísticas.

Aparência dos Robôs

Alguns estudos foram realizados demonstrando que a forma física do robô influencia diretamente na interação humano-robô (Broadbent et al., 2011; Nomura e Nakao, 2010; Lee et al., 2017; Xia e LeTendre, 2020; Reich-Stiebert et al., 2019).

Um exemplo é o estudo realizado por Broadbent et al. (2011). Foi verificado que pessoas com mais de 40 anos preferem interagir com robôs que não são semelhantes ao ser humano (robôs humanoides).

Já Reich-Stiebert et al. (2019) demostram que estudantes universitários preferem interagir com robôs com características humanas. Segundo o trabalho, o robô deve interagir principalmente por meio da fala e ser capaz de exibir emoções básicas, especialmente as positivas. Além disso, do ponto de vista dos estudantes universitários, o robô deve exibir um comportamento marcado por gentileza.

2.5 Personalidade na Interação Humano-Robô

A personalidade pode ser definida como sendo um padrão de comportamento característico de um indivíduo. Esse comportamento se estende em amplos sentidos, como por exemplo: ações, pensamentos, sentimentos e motivação (Robert, 2018). Além disso, um indivíduo pode ter a personalidade estável por um longo período de tempo, mudando poucas vezes ao longo da vida. Já o humor, por um tempo intermediário e emoção variando por um curto período de tempo (Tkalčič et al., 2016).

Nas últimas décadas, pesquisadores buscaram entender a personalidade humana e as diferenças básicas entre os indivíduos. Uma das teorias que ganhou grande aceitação na comunidade científica foi a de abordagens de traços de personalidade (Neto e da Silva, 2012; Esterwood e Robert, 2021). A análise de personalidade fundamentada em traços baseia-se na medição de padrões gerais de comportamentos, pensamentos e emoções (Giritlioğlu et al., 2021). Esses padrões (traços) são relativamente estáveis ao longo do tempo e em diferentes contextos.

A partir da abordagem da personalidade utilizando o conceito de traços, surgiram algumas teorias como: o Modelo dos 16 Fatores de Cattell (16F) (Cattell, 1946), o Modelo dos três Super Fatores de Eyssenck (Eysenck, 1991) e o Modelo dos 5 Grandes Fatores (Digman, 1990).

Dessas teorias, uma em questão, tem recebido a atenção dos pesquisadores na área de robótica social que é o Modelo dos 5 Grandes Fatores (Esterwood e Robert, 2021). Esse Modelo é descrito com maiores detalhes na sessão a seguir (Sessão 2.5.1).

2.5.1 Modelo dos Cinco Grandes Fatores

O Modelo dos Cinco Grandes Fatores começou a ser elaborado no ano de 1930 pelo psicólogo britânico Willian McDougall que sugeriu analisar a personalidade com relação a cinco fatores (Neto e da Silva, 2012; Digman, 1990). No entanto, esse modelo só ganhou visibilidade a partir dos anos de 1980, quando os pesquisadores e psicólogos começaram a comprovar a existência dos cinco traços de personalidade (Neto e da Silva, 2012).

Esta teoria indica que a personalidade pode ser descrita utilizando cinco traços: abertura a experiência, responsabilidade, extroversão, cordialidade e neuroticismo (Digman, 1990). Podendo ser definidos da seguinte forma:

Abertura a experiência Indivíduos que são abertos a experiências são caracterizados por se-

rem criativos, inteligentes, originais, perspicazes e curiosos. O oposto é descrito como pessoas que possuem pouco interesse, que são simples, superficiais e que não gostam de ideias que provocam mudanças.

Responsabilidade É caracterizado por pessoas responsáveis, organizadas e persistentes. O oposto é definido por pessoas que não possuem objetivos claros, que são irresponsáveis e preguiçosos.

Extroversão Um indivíduo extrovertido é assertivo, falante e sociável. O oposto de extroversão é a introversão, caracterizada por uma pessoa que é quieta, reservada ou tímida.

Cordialidade É caracterizada por pessoas amáveis, cordiais, prestativas e altruístas. O oposto são pessoas frias, egocêntricas, rude e irritáveis.

Neuroticismo Indivíduos com traços altos de neuroticismo são caracterizados por serem nervosos, tensos, sem paciência. O oposto são pessoas equilibradas e calmas.

A personalidade é definida por um vetor de cinco posições (veja Equação (2.1)): A (abertura a experiência), R (responsabilidade), E (extroversão), C (cordialidade) e N (neuroticismo). Cada uma das posições do vetor variam de $-1{,}0$ até $+1{,}0$.

$$P = (A{,}R{,}E{,}C{,}N) \tag{2.1}$$

Valores próximos à $-1{,}0$ indicam baixo nível do traço em questão. E valores próximos à $1{,}0$ indicam um alto nível do traço. Por exemplo, um indivíduo que possui como traço de personalidade $P = (1{,}1{,}1{,}1{,}-1)$, significa que essa pessoa possui um alto nível de abertura a experiência, responsabilidade, extroversão e cordialidade; e um baixo nível de neuroticismo.

Esses traços estão associados principalmente à cognição, afeto e comportamento não verbal, como olhar, movimento da cabeça, pose do corpo e aparência facial (Giritlioğlu et al., 2021). Segundo Zillig et al. (2002), a Responsabilidade é dominada mais pelo comportamento, o Neuroticismo pelo afeto negativo. Já a Extroversão é dominada pelo lado afetivo e o comportamental e, finalmente, a Abertura e a Cordialidade pelas cognições (Zillig et al., 2002; Giritlioğlu et al., 2021). Ainda no trabalho de Zillig et al. (2002), sugere que certos traços de personalidade são mais visíveis aos olhos do que outros. Nesse sentido, traços como Extrover-

são, Responsabilidade ou Neuroticismo seriam mais aparentes ao observarmos um indivíduo à primeira vista.

Psiquiatras e psicólogos clínicos estudam extensivamente os traços de personalidade e seu papel, tanto no diagnóstico quanto no prognóstico de transtornos psiquiátricos (Giritlioğlu et al., 2021). Alguns estudos na literatura indicam forte relação à depressão, transtorno afetivo bipolar, esquizofrenia e ansiedade aos 5 traços de personalidade.

Segundo os artigos de Mulder (2002) e Kim et al. (2012a), os traços de personalidade podem afetar os resultados do tratamento da depressão e do transtorno afetivo bipolar. Além disso, podem fornecer orientação sobre os resultados do tratamento, bem como o prognóstico da doença, exibindo até mesmo uma associação com mudanças nos estados do transtorno (Kim et al., 2011; Giritlioğlu et al., 2021).

Traços de personalidade também têm uma estreita relação com transtornos psicóticos e de ansiedade. Segundo Ohi et al. (2016), pacientes com esquizofrenia têm um nível mais alto de Neuroticismo e níveis mais baixos de Extroversão, Abertura, Cordialidade e Responsabilidade. Estudos anteriores mostram que os Cinco Grandes Traços estão associados ao funcionamento social (Lysaker e Davis, 2004), satisfação com a vida (Boyette et al., 2014) e não adesão e/ou atraso no tratamento em pacientes com esquizofrenia (Compton et al., 2015; Lecomte et al., 2008).

Além disso, os traços de personalidade parecem ter um papel no desenvolvimento de transtornos de ansiedade. No trabalho de Wauthia et al. (2019), descobriram que um alto nível de Neuroticismo prediz fobia social, transtorno do pânico e transtorno de ansiedade generalizada. Níveis mais altos de Neuroticismo e níveis mais baixos de Responsabilidade estão relacionados a um aumento da ansiedade e uma menor Extroversão está associada à fobia social (Kotov et al., 2010).

Por tanto, pode-se perceber que os traços de personalidade têm um papel fundamental tanto para o diagnóstico quanto para o prognóstico de grandes transtornos mentais. Assim, existe a necessidade de mais trabalhos na literatura que se dediquem aos estudos da importância da detecção dos traços de personalidade do indivíduo em tratamentos clínicos. Principalmente se tratando de outros transtornos, como é o exemplo do autismo.

O Modelo dos 5 Grandes Fatores tem sido aplicado para analisar de que forma a personalidade influencia a interação entre humanos e robôs em dois aspectos distintos. O primeiro se

relaciona à personalidade do indivíduo humano, enquanto o segundo aborda a personalidade do robô. A seguir serão apresentados mais detalhes dos dois aspectos citados.

2.5.2 Personalidade Humana

A primeira área de pesquisa é sobre a personalidade humana na interação humano-robô. Alguns trabalhos desta área são Gockley e MatariĆ (2006); Syrdal et al. (2006, 2007); Walters et al. (2008); Salem et al. (2015). Segundo a revisão realizada por Robert (2018), os pesquisadores da área tem investigado como os traços da personalidade humana influenciam na aceitação e confiança no robô.

O traço do Modelo dos Cinco Grandes Fatores mais estudado pela literatura é a Extroversão/Introversão. Segundo Gockley e MatariĆ (2006); Syrdal et al. (2007), pessoas extrovertidas se sentem mais confortáveis com a presença dos robôs sociais. Além disso, extrovertidos tendem a humanizar os robôs, isto é, representando os robôs como humanos e/ou atribuindo qualidades humanas aos robôs.

Extrovertidos se sentem mais confortáveis com robôs autônomos, enquanto os introvertidos preferem estar no controle do robô, segundo o trabalho de Syrdal et al. (2006). Os introvertidos também preferem robôs que parecem mais mecânicos do que aqueles que parecem humanos (humanoides) (Walters et al., 2008).

Além da Extroversão/Introversão, pesquisadores em vários estudos examinaram os outros Cinco Grandes Traços de personalidade. O baixo Neuroticismo, ou seja, alta estabilidade emocional foi positivamente correlacionado com robôs humanoides e com sentimentos de simpatia em relação aos robôs (Salem et al., 2015). Indivíduos neuróticos (ou seja, baixa estabilidade emocional) tendem a preferir robôs de aparência mecânica em comparação com robôs de aparência humanoide (Walters et al., 2008).

2.5.3 Personalidade do Robô

A segunda área de pesquisa diz respeito a personalidade do robô na Interação Humano-Robô, pois a exibição de personalidades de um robô aumenta a facilidade de uso e diminui o humor negativo das pessoas, segundo Moshkina e Arkin (2005). Esta é uma área de estudo recente e possui poucos trabalhos na literatura, em sua maioria, dizem a respeito ao traço de personalidade da Extroversão/Introversão do robô (Robert, 2018).

Os robôs extrovertidos são considerados divertidos e engraçados pelas pessoas, causando uma impressão positiva (Goetz e Kiesler, 2002). Em interações diádicas, ou seja, apenas o robô e uma pessoa, os robôs extrovertidos foram considerados mais inteligentes (Leuwerink, 2012).

Já os robôs introvertidos foram considerados mais sérios, menos brincalhões e, por sua vez, menos divertidos (Goetz e Kiesler, 2002). Em interações de grupo, o robô extrovertido foi percebido como o mais inteligente, e na interação diádica o robô extrovertido foi considerado o mais inteligente (Leuwerink, 2012).

No entanto, algumas pesquisas sugerem que a personalidade do robô depende de sua aplicação. Como é o caso do estudo realizado em Tay et al. (2014), em que foi investigado se a aplicação do robô (segurança *versus* cuidador) influenciaram o impacto das personalidades extrovertidas e introvertidas do robô. Os participantes tiveram uma resposta mais positiva ao robô de saúde extrovertido do que ao robô de saúde introvertido. No entanto, os participantes tiveram uma resposta mais positiva ao robô de segurança introvertido do que ao robô de segurança extrovertido. Concluindo que robôs de segurança que são introvertidos passam mais credibilidade, assim como o robô cuidador ser extrovertido.

2.6 Discussão

A partir da extensa revisão de literatura realizada sobre o uso de emoções e personalidade na interação humano-robô é possível extrair algumas conclusões. No que diz respeito ao uso de emoções, é possível notar que existem numerosos estudos na literatura que abordam essa área, tanto no reconhecimento de emoções humanas quanto na expressão de emoções pelos robôs.

O objetivo principal de 53% dos trabalhos analisados foi o desenvolvimento de métodos ou técnicas para detectar emoções em humanos, reconhecendo emoções a partir da expressão facial, fala, toque ou sinais fisiológicos. Quanto aos outros estudos, 47% dos artigos analisaram a melhor forma de representar as emoções em robôs para que a IHR fosse natural e amigável. Portanto, há um equilíbrio nos trabalhos considerados nesta revisão, com o propósito de impulsionar pesquisas em IHR (Interação Humano-Robô) para a transmissão e reconhecimento de emoções.

No que diz respeito aos estudos que desenvolvem técnicas de reconhecimento de emoções humanas, o reconhecimento de expressões faciais, fala e sinais fisiológicos tem recebido uma

grande atenção nos últimos anos. No entanto, é importante destacar que a identificação de emoções por meio do toque e dos gestos tem sido menos explorada, conforme evidenciado pela escassez de trabalhos encontrados na revisão sistemática. O reconhecimento por meio do toque tem se revelado um novo campo de pesquisa, investigando como os seres humanos transmitem emoções por meio do contato físico. Além disso, essa área tem mostrado grande expectativa para modelar e incorporar comportamentos de toque emocional em robôs no futuro.

Dos artigos analisados, 47% abordam técnicas para que os robôs expressem suas emoções. A forma mais comum de expressão emocional por parte dos robôs é através da face e gestos. Além disso, a maioria dos estudos utiliza o modelo de emoções proposto por Paul Ekman, que inclui as emoções de tristeza, raiva, felicidade, medo, desgosto e surpresa.

Os estudos sobre a transmissão de emoções por meio da fala é uma área com poucos trabalhos publicados e tendem a se desenvolver mais rapidamente. Há uma forte tendência nos estudos de qual seria a melhor aparência que o robô deveria ter para ganhar mais empatia dos humanos.

É possível observar algumas ausências que podem se tornar rumos futuros. Observou-se que a grande maioria dos trabalhos trata de protótipos de robôs, ou seja, uma abordagem formativa. Existem poucos artigos que apresentam o robô com mais de uma capacidade de expressar ou reconhecer emoções. Portanto, espera-se que trabalhos futuros apresentem mais de uma forma de reconhecer ou expressar emoções (abordagens multimodais). Além disso, as emoções são trabalhadas de forma binária, ou seja, apenas uma classe de emoção é determinada. Devido a natureza subjetiva da emoção humana, esperava-se que a emoção fosse tratada de forma probabilística, ou seja, determinada de forma multiclasse, na qual existissem mais de uma emoção presente.

Com relação à personalidade na interação humano-robô, é possível observar que é uma área recente e com poucos trabalhos. Ainda há muita pesquisa a ser feita neste campo, principalmente no reconhecimento automático das personalidades. Já existem teorias matemáticas para extrair a personalidade a partir da emoção e humor, como a teoria ALMA apresentada em Gebhard (2005). No entanto, como se sabe a personalidade é uma característica que o ser humano apresenta a longo prazo. Há muito a ser estudado ainda sobre a personalidade na IHR.

CAPÍTULO 3

Fundamentação Teórica

3.1 Introdução

Neste capítulo será realizada uma fundamentação teórica sobre Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN) e Sistemas *Fuzzy* (SF), ambas são técnicas da inteligência computacional e muito utilizadas em trabalhos na área da interação humano-robô. A Rede Neural Convolucional é um tipo de Redes Neurais Artificiais (RNA), na qual é um método de aprendizado profundo (do inglês: *deep learning*). As CNNs tem ganhado grande destaque na literatura pelos resultados alcançados, principalmente na área de visão computacional. Na Seção 3.2 serão apresentados mais informações sobre as CNNs.

Já o Sistema *Fuzzy* ou Nebuloso, é uma técnica clássica da inteligência computacional. Na qual, é utilizada a lógica *fuzzy* que é uma lógica não-binária, além disso, realiza-se a inferência a partir de um banco de regras baseado em conhecimento de especialista. Na Seção 3.3 serão apresentados mais informações sobre os sistemas *fuzzy*.

3.2 Redes Neurais Convolucionais

As Redes Neurais Artificiais são modelos computacionais inspirados na maneira como os sistemas nervosos biológicos (como o cérebro humano) operam. As RNAs são compostas principalmente por um grande número de nós computacionais interconectados (denominados de neurônios), que trabalham interligados para aprender coletivamente com a entrada, a fim de otimizar sua saída final (Silva et al., 2016). As características mais relevantes envolvidas com a aplicação de redes neurais artificiais, segundo Silva et al. (2016) são:

Adaptação por experiência: as adaptações dos parâmetros internos da rede, tipicamente seus pesos sinápticos, são executados a partir da apresentação sucessiva de exemplos (padrões, amostras, medidas) relacionados ao comportamento do processo, possibilitando a obtenção do conhecimento por experimentação;

Capacidade de aprendizado: por meio da aplicação de alguns métodos de treinamento a rede consegue extrair a relação entre as diversas variáveis que compõem a aplicação;

Habilidade de generalização: após o treinamento da rede, ela é capaz de generalizar o conhecimento adquirido, possibilitando estimar soluções que eram até então desconhecidas;

Organização de dados: baseada em características envolvendo determinado conjunto de informações a respeito de um processo, a rede é capaz de realizar sua organização interna visando possibilitar o argumento de padrões que apresentam particularidades em comum;

Tolerância a falhas: devido ao elevado nível de interconexões entre os neurônios artificiais, a rede neural torna-se um sistema tolerante a falhas;

Armazenamento distribuído: o conhecimento a respeito do comportamento de determinado processo dentro de uma arquitetura neural é representado de forma distribuída entre as diversas sinapses de seus neurônios artificiais;

Facilidade de prototipagem: a implementação da maioria das arquiteturas neurais pode ser prototipada em hardware ou em software, pois, após o processo de treinamento, os resultados são normalmente obtidos por algumas operações matemáticas.

A estrutura básica de um neurônio pode ser modelada conforme mostrado na Figura 3.1. Os valores de x_n são as entradas do neurônio, ou seja, são os sinais ou medidas vindas do meio externo. Os valores w_n são os pesos sinápticos, eles servirão para ponderar cada uma das variáveis de entrada da rede, permitindo quantificar as suas relevâncias em relação ao funcionamento do respectivo neurônio. O limiar de ativação, representado por θ, é responsável por especificar qual o estágio apropriado para que o resultado produzido pelo combinador linear possa gerar um valor de disparo de ativação. A função de ativação f, cujo objetivo é limitar a saída do neurônio dentro de um intervalo de valores razoáveis. E por fim, y é o sinal de saída, no qual, consiste no valor final produzido pelo neurônio em relação a um determinado conjunto de sinais de entrada (Silva et al., 2016).

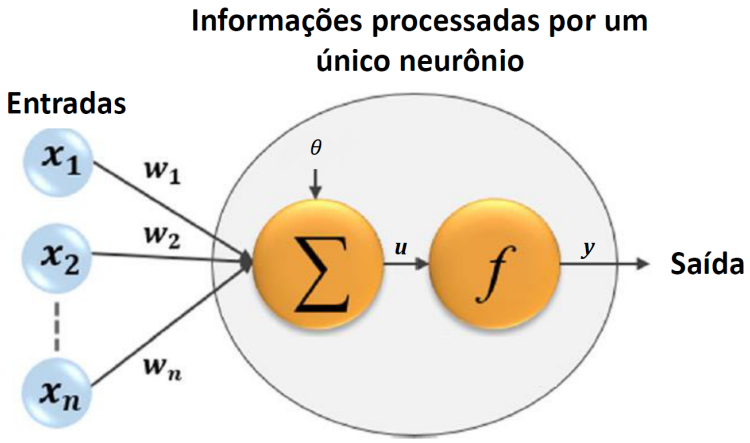


Figura 3.1: Representação de uma Rede Neural Artificial. Fonte: adaptado de Talpur et al. (2022).

Desta forma, podemos sintetizar o resultado produzido pelo neurônio artificial com as Equações (3.1) e (3.2). Em que, u é dado pelo somatório das entradas (x_i) multiplicado pelos seus respectivos pesos (w_i) e subtraído o valor do limiar de ativação (θ). Por fim, o valor de saída (y) de um neurônio é dado pela função de ativação de u, ou seja, $f(u)$.

$$u = \sum_{i=1}^{n} w_i x_i - \theta \tag{3.1}$$

$$y = f(u) \tag{3.2}$$

Por tanto, uma rede neural artificial, pode possuir apenas um neurônio (denominada Rede Neural *Perceptron*), ou diversos neurônios com diversas camadas. Na Figura 3.2 é possível observar uma Rede Neural Profunda, possuindo múltiplas camadas ocultas, ou seja, contem U_n camadas ocultas e mais neurônios que uma Rede *Perceptron* (Silva et al., 2010).

Rede Neural Convolucional - CNN

As Redes Neurais convolucionais são um tipo de rede neural, especializada em processamento de dados. Os exemplos incluem dados de séries temporais, que podem ser considerados como uma grade 1-D coletando amostras em intervalos de tempo regulares, e dados de imagem, que podem ser considerados como uma grade 2-D de pixels (Goodfellow et al., 2016).

As redes convolucionais têm sido bem-sucedidas em aplicações práticas e seu nome indica que a rede emprega uma operação matemática chamada convolução. A convolução é um tipo

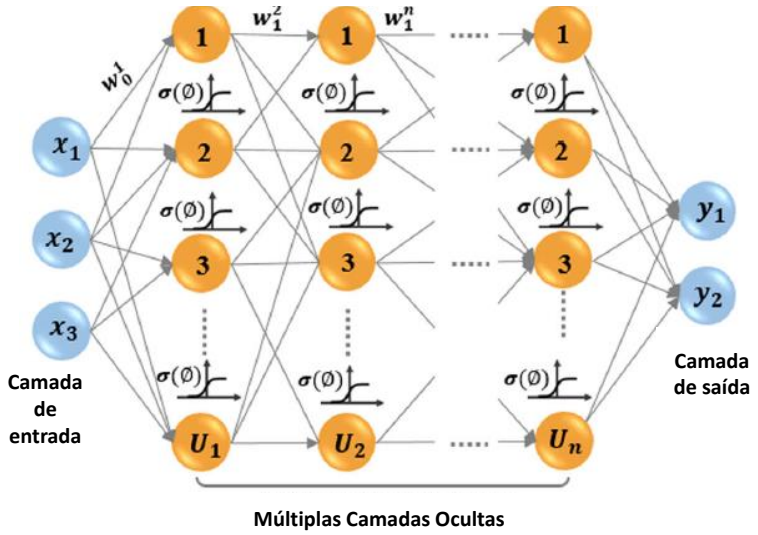


Figura 3.2: Representação da Rede Neural Profunda. Fonte: adaptado de Talpur et al. (2022).

especializado de operação linear que pode ser definida conforme a Equação (3.3). Em que na terminologia de redes convolucionais, o primeiro argumento (neste exemplo, a função x) da convolução é frequentemente referido como a entrada (*input*), e o segundo argumento (neste exemplo, a função w) como o filtro (*kernel*). A saída é, às vezes, chamada de mapa de características (*feature map*) (Goodfellow et al., 2016).

$$s(t) = (x * w)(t) \tag{3.3}$$

As redes convolucionais são simplesmente redes neurais que usam convolução no lugar da multiplicação geral de matrizes em pelo menos uma de suas camadas (Goodfellow et al., 2016). De forma geral, as redes neurais convolucionais seguem o padrão mostrado na Figura 3.3, possuem a entrada, as camadas de convolução, as camadas totalmente conectadas e por fim a saída.

Uma camada convolucional é o bloco central de uma rede neural convolucional, conforme mostrado na Figura 3.3. Na matemática, a convolução é a operação de duas funções para produzir uma terceira função modificada. No contexto das CNNs, por exemplo, a primeira função pode ser uma imagem de entrada e a segunda função é o filtro convolucional (Elgendy, 2020). Ao deslizar o filtro convolucional sobre a imagem de entrada, a rede quebra a imagem em pequenos pedaços e processa esses pedaços individualmente para montar a imagem modificada, um mapa de características (Elgendy, 2020).

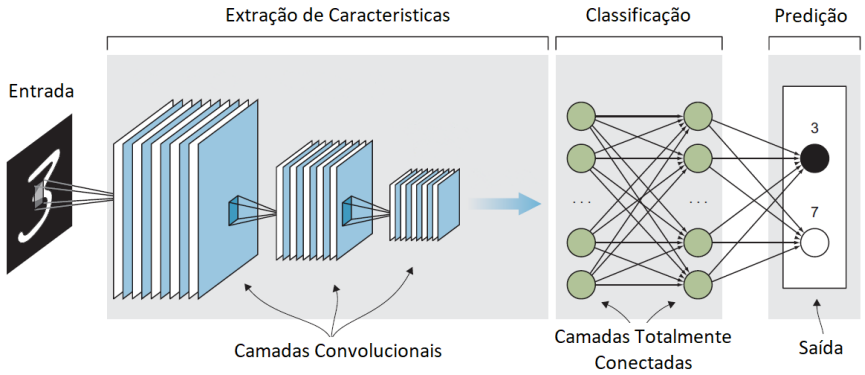


Figura 3.3: Representação da Rede Neural Convolucional. Fonte: adaptado de Elgendy (2020).

A diferença fundamental entre uma camada totalmente conectada e uma camada de convolução é que as camadas densas aprendem padrões globais em seu espaço de características, enquanto as camadas de convolução aprendem padrões locais. A operação de convolução extrai fragmentos do mapa de características de entrada e aplica a mesma transformação a todos esses fragmentos, produzindo um mapa de características de saída (Chollet, 2021).

Após a extração das características, as camadas totalmente conectadas (ou camadas densas) realizam a classificação. Na Figura 3.3, é realizada a classificação da imagem em duas classes: o número 3 ou 7. Após as camadas densas é por fim realizada a predição final, sendo por tanto a saída da rede neural convolucional (Chollet, 2021).

Rede Neural Recorrente - RNN

As redes neurais recorrentes (RNNs) (Rumelhart et al., 1986) formam uma classe de redes neurais projetadas para lidar com dados sequenciais. Enquanto as redes convolucionais são especializadas no processamento de grades de valores, como imagens, as RNNs são projetadas para processar sequências de valores. De maneira semelhante às redes convolucionais, que podem ser escaladas para lidar com imagens de grandes dimensões e, em alguns casos, imagens de tamanhos variáveis, as RNNs conseguem lidar com sequências consideravelmente longas, algo que seria inviável para redes sem essa especialização. Além disso, a maioria das RNNs tem a capacidade de processar sequências de comprimentos variáveis (Goodfellow et al., 2016).

A *Long Short-Term Memory* (LSTM) é um tipo de rede neural recorrente (RNN) projetada para superar o problema do desaparecimento e explosão de gradientes, comuns em RNNs tradicionais ao lidar com dependências de longo prazo em dados sequenciais. Introduzida por

Hochreiter e Schmidhuber em 1997, a LSTM utiliza uma estrutura de portas (*gate mechanisms*) para controlar o fluxo de informações na rede, permitindo que ela armazene, atualize ou descarte informações de maneira eficiente ao longo do tempo (Goodfellow et al., 2016).

A LSTM tem se mostrado extremamente bem-sucedida em muitas aplicações, como reconhecimento de escrita à mão sem restrições, reconhecimento de fala, geração de escrita à mão, tradução automática, legendagem de imagens e análise sintática. O diagrama de blocos da LSTM é ilustrado na Figura 3.4. As equações correspondentes de propagação direta são apresentadas abaixo, para uma rede recorrente rasa (Goodfellow et al., 2016).

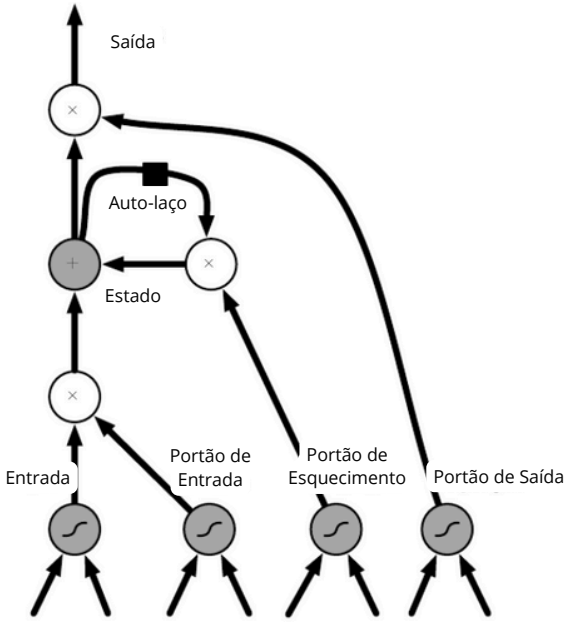


Figura 3.4: Representação da LSTM. Fonte: adaptado de Goodfellow et al. (2016).

O diagrama de blocos da célula da rede recorrente LSTM (Figura 3.4) mostra como as células estão conectadas recorrente e diretamente entre si, substituindo as unidades ocultas habituais das redes recorrentes convencionais. Uma característica de entrada é calculada usando uma unidade de neurônio artificial comum. Seu valor pode ser acumulado no estado se o portão de entrada sigmoidal permitir. A unidade de estado possui um auto-laço linear cujo peso é controlado pelo portão de esquecimento. A saída da célula pode ser interrompida pelo portão de saída. Todas as unidades de portão usam uma não linearidade sigmoidal, enquanto a unidade de entrada pode utilizar qualquer função de ativação que esprema o valor. A unidade de estado também pode ser usada como uma entrada extra para os portões. O quadrado preto no diagrama indica um atraso de um único passo temporal (Goodfellow et al., 2016).

3.2.1 Métricas de Desempenho

As métricas de desempenho permitem avaliar o sistema desenvolvido, isto é, avaliar como ele está funcionando. A maneira mais simples de medir o modelo de classificação é por meio da acurácia. A métrica de acurácia mede quantas vezes o modelo fez a previsão correta. Portanto, ao testar o modelo com 100 amostras de entrada for realizada a previsão correta 90 vezes, isso significa que o modelo é 90% preciso. A Equação (3.4) é utilizada para calcular a acurácia do modelo.

$$Acurácia = \frac{\textit{Predições Corretas}}{\textit{Número total de amostras}} \times 100 \tag{3.4}$$

No entanto, para situações ou aplicações com a base de dados desbalanceada a acurácia se torna problemática. Por exemplo, para uma pesquisa em que se deseja realizar o diagnóstico médico de uma doença rara, na qual para um milhão de pessoas, uma tenha a doença, o sistema alcançará uma acurácia de 99,9%, mas não significa que ele seja bom (Elgendy, 2020).

Para elaborar outras métricas de desempenho utiliza-se o conceito da matriz de confusão. Uma matriz de confusão é uma tabela que descreve o desempenho do modelo de classificação. Por exemplo, suponha um classificador para prever se um paciente está doente ou saudável, as classificações esperadas são positivas (o paciente está doente) ou negativas (o paciente está saudável). Desta forma, cria-se a matriz de confusão demostrada na Figura 3.5 (Elgendy, 2020).

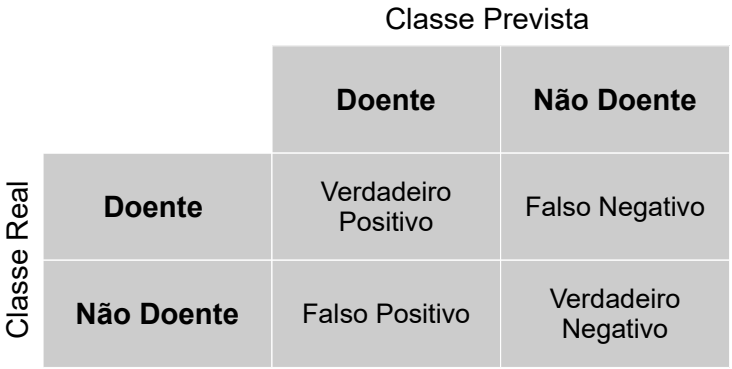


Figura 3.5: Representação da Matriz de Confusão, exemplo para diagnóstico de paciente doente ou não doente. Fonte: adaptado de Elgendy (2020).

Conforme mostrado na Figura 3.5, os verdadeiros positivos acontecem quando o modelo previu corretamente, o paciente tem a doença . Já os verdadeiros negativos, o modelo previu corretamente que o paciente não tem a doença. Os falsos positivos, são as situações em que

o modelo previu erroneamente que o paciente tem a doença, mas o paciente na verdade não tem a doença. E por fim, os falsos negativos correspondem a situação em que o modelo previu erroneamente que não possui a doença, mas o paciente realmente tem a doença.

Com isso, os pacientes que o modelo prevê serem negativos (sem doença) são aqueles que o modelo acredita serem saudáveis. Já os pacientes que o modelo prediz serem positivos (têm doença) são os que seriam enviados para uma investigação mais aprofundada. Qual erro prefere-se cometer? Diagnosticar erroneamente alguém como positivo (tem doença) e encaminhá-lo para mais investigação ou diagnosticar erroneamente alguém como negativo (saudável) e mandá-lo para casa com risco de vida. A escolha da métrica de avaliação mais preocupante é com os números de falsos negativos. Desta forma, busca-se encontrar todas as pessoas doentes, mesmo que o modelo acidentalmente classifique algumas pessoas saudáveis como doentes. Essa métrica é chamada de *recall* ou sensibilidade (Elgendy, 2020).

O *recall* informa quantos pacientes doentes o modelo diagnosticou incorretamente. Em outras palavras, quantas vezes o modelo diagnosticou incorretamente um paciente doente como negativo (falso negativo)? O *recall* é calculado pela Equação (3.5) (Elgendy, 2020).

$$Recall = \frac{\textit{Verdadeiro Positivo}}{\textit{Verdadeiro Positivo + Falso Negativo}} \times 100 \tag{3.5}$$

A precisão (também conhecida como especificidade) é o oposto do *recall*. Ela revela quantos pacientes saudáveis o modelo diagnosticou incorretamente como doentes. Em outras palavras, quantas vezes o modelo diagnosticou incorretamente um paciente saudável como positivo (falso positivo)? A precisão é calculada pela Equação (3.6) (Elgendy, 2020).

$$Precisão = \frac{\textit{Verdadeiro Positivo}}{\textit{Verdadeiro Positivo + Falso Positivo}} \times 100 \tag{3.6}$$

Em muitos casos, deseja-se unir o desempenho de um classificador com uma única métrica que represente tanto o *recall* quanto a precisão. Para fazer isso, pode-se converter precisão e *recall* em uma única métrica *F1-score*, conforme demostrado na Equação (3.7), na qual é realizado a média harmônica de p e r (Elgendy, 2020).

$$\textit{F1-score} = \frac{2 \times precisão \times recall}{precisão + recall} \times 100 \tag{3.7}$$

3.2.2 Hiperparâmetros da Rede

Os algoritmos de aprendizado de máquina, normalmente são definidos por hiperparâmetros que controlam muitos aspectos do comportamento do modelo. Alguns hiperparâmetros afetam o tempo e o custo de memória da execução do algoritmo e outros afetam a capacidade de previsão do modelo (Elgendy, 2020). De modo geral, podemos categorizar os hiperparâmetros das redes neurais convolucionais em hiperparâmetros de arquitetura; hiperparâmetros de aprendizado e otimização; e os hiperparâmetros de regularização.

Hiperparâmetros de Arquitetura

Em se tratando dos hiperparâmetros da arquitetura da rede neural, tem-se: o número de camadas ocultas (representando a profundidade da rede), o número de neurônios em cada camada, também conhecidos como unidades ocultas (representando a largura da rede) e as funções de ativação.

O número de camadas e unidades ocultas descrevem a capacidade de aprendizado da rede. O objetivo é definir um número grande o suficiente para que a rede aprenda as características dos dados. Uma rede menor pode ser insuficiente e uma rede maior pode ser superadaptada. Quanto mais complexo o conjunto de dados, mais capacidade de aprendizado o modelo precisará para aprender. Se o seu modelo estiver superajustado, talvez seja um sinal para diminuir o número de unidades ocultas (Elgendy, 2020; Chollet, 2021).

As funções de ativação introduzem a não linearidade nos neurônios. Sem ativações, os neurônios passariam combinações lineares (somas ponderadas) uns para os outros e não resolveriam nenhum problema não linear. Esta é uma área de pesquisa muito ativa e há muitas funções de ativação disponíveis. Mas no momento, ReLU (*Rectified Linear Unit*) têm o melhor desempenho em camadas ocultas (Elgendy, 2020). E na camada de saída é muito comum usar a função softmax para problemas de classificação, com o número de neurônios igual ao número de classes do seu problema (Elgendy, 2020).

A função de ativação ReLu, ativa um nó somente se a entrada estiver acima de zero. Se a entrada estiver abaixo de zero, a saída será sempre zero. Mas quando a entrada é maior que zero, ela tem uma relação linear com a variável de saída. A função ReLU é representada por (Elgendy, 2020):

$$f(x) = max(0,x). \tag{3.8}$$

A função softmax é uma generalização da função sigmoide. É usada para obter probabilidades de classificação quando se tem mais de duas classes. Ela transforma os valores de entrada em valores de probabilidade entre 0 e 1, sendo que a soma final de todas as probabilidades é 1. Um caso de uso muito comum em problemas de aprendizado profundo é prever uma única classe entre muitas opções (mais de duas) (Elgendy, 2020). A equação softmax é descrita por (3.9), em que $\sigma(x_j)$ é a probabilidade do neurônio de saída, x_j é o vetor do neurônio de saída e i são os índices de todos os neurônios.

$$\sigma(x_j) = \frac{e^{x_j}}{\sum_i e^{x_i}} \tag{3.9}$$

Hiperparâmetros de Aprendizado e Otimização

Os hiperparâmetros de aprendizado e otimização são responsáveis por determinar como a rede aprende e otimiza seus parâmetros para atingir o erro mínimo. Dentre eles estão os algoritmos de otimização, a taxa de aprendizado e os critérios de parada.

O algoritmo de otimização determina como a rede será ajustada com base na função de perda. Já a função de perda, por sua vez, calcula a quantidade que será minimizada durante o treinamento (Chollet, 2021). Existem diversas funções de otimização, como por exemplo o SGD, Adam e Adagrad.

O *Stochastic Gradient Descent* (SGD) é uma técnica de otimização amplamente empregada em aprendizado de máquina, especialmente em aprendizado profundo. Ao contrário do gradiente descendente regular, que calcula a perda e o gradiente em todo o conjunto de dados de treinamento antes de ajustar os parâmetros, o SGD adota uma abordagem mais ágil. Ele seleciona aleatoriamente um ponto de dados do conjunto de treinamento para cada etapa e calcula o gradiente usando apenas aquela instância (Elgendy, 2020).

O otimizador Adam, abreviação de *Adaptive Moment Estimation*, é um algoritmo de otimização amplamente utilizado em aprendizado de máquina e treinamento de redes neurais. Adam combina conceitos do SGD e do método *momentum* calculando médias móveis de gradientes passados e quadrados. Isso permite lidar com problemas de otimização não estacionários de maneira eficaz e ajuda a evitar oscilações indesejadas no caminho de convergência. O Adam

provou ser particularmente eficiente na aceleração do treinamento de redes profundas e é amplamente adotado na comunidade de aprendizado de máquina (Elgendy, 2020).

O Adagrad, uma abreviatura de *Adaptive Gradient Descent*, é uma técnica de otimização baseada em gradiente descendente. É um otimizador que utiliza taxas de aprendizagem adaptadas a parâmetros específicos, adaptando-as de acordo com a frequência com que cada parâmetro é atualizado durante o treinamento. Parâmetros sujeitos a atualizações frequentes apresentam taxas de aprendizagem reduzidas, levando a ajustes de parâmetros progressivamente menores à medida que o treinamento se desenvolve (Duchi et al., 2011).

Um dos parâmetros de entrada do otimizador é a taxa de aprendizado. Teoricamente, uma taxa de aprendizado muito pequena garante o erro mínimo (se treinar por tempo infinito). Uma taxa de aprendizado muito alta acelera o aprendizado, mas não garante encontrar o erro mínimo (Elgendy, 2020).

Um critério comum é o número de iterações ou épocas de treinamento. Uma iteração de treinamento, ou época, é quando o modelo faz um ciclo completo e vê todo o conjunto de dados de treinamento de uma só vez. O número de épocas é definido para ajustar quantas iterações a rede vai continuar treinando. Quanto mais iterações de treinamento, mais o modelo aprende as características dos dados de treinamento (Elgendy, 2020). No entanto, vale ressaltar um dos problemas mais comuns das redes neurais: o *overfitting*, em que a partir de um certo número de iterações de treinamento, a rede começa a aprender padrões específicos dos dados, mas que são enganosos ou irrelevantes quando se trata de novos dados (Chollet, 2021). Para evitar o *overfitting* são aplicados os hiperparâmetros de regularização.

Hiperparâmetros de Regularização

Se a rede neural está superajustando (*overfitting*) a partir dos dados de treinamento, a rede pode estar com alta capacidade e precisa ser simplificada. Uma das técnicas é a regularização. As técnicas mais comuns de regularização são: *dropout* e *data augmentation*.

O *dropout* é uma técnica de regularização muito eficaz para simplificar uma rede neural e evitar o *overfitting*. O algoritmo de *dropout* é bastante simples: a cada iteração de treinamento, cada neurônio tem uma probabilidade de ser temporariamente ignorado durante aquela iteração. Embora seja contra-intuitivo pausar intencionalmente o aprendizado em alguns dos neurônios da rede, é bastante surpreendente o quão bem essa técnica funciona. A probabilidade é um

hiperparâmetro chamado de taxa de *dropout* (Elgendy, 2020).

Além do *dropout*, existem as técnicas de regularização L1 e L2. A regularização L1 (ou *Lasso*) adiciona a soma dos valores absolutos dos coeficientes ao termo de perda, incentivando que alguns dos coeficientes sejam reduzidos exatamente a zero, o que resulta em modelos esparsos (com menos variáveis). Já a regularização L2 (ou *Ridge*) adiciona a soma dos quadrados dos coeficientes, penalizando grandes valores dos pesos, mas sem zerar os coeficientes completamente, mantendo todos os parâmetros, porém com valores menores (Elgendy, 2020).

Uma outra maneira de evitar o *overfitting* é obter mais dados. Como essa nem sempre é uma opção viável, pode-se aumentar os dados de treinamento gerando novas instâncias das mesmas imagens com algumas transformações. O *data augmentation* é uma maneira de fornecer ao algoritmo de aprendizado mais dados de treinamento e, portanto, reduzir o *overfitting*. As muitas técnicas de aumento de imagem incluem inversão, rotação, dimensionamento, zoom, condições de iluminação e muitas outras transformações que podem ser aplicadas ao conjunto de dados para fornecer uma variedade de imagens para treinar (Elgendy, 2020).

3.3 Sistema Fuzzy

O Sistema *Fuzzy* (SF), também conhecido por Sistema Nebuloso, foi proposto por Zadeh, sugerindo uma ideia diferente da lógica binária (0 ou 1). A lógica *fuzzy* leva em consideração elementos qualitativos, ao invés de valores quantitativos. Ou seja, são utilizadas variáveis linguísticas como: “muito", “pouco", “médio", “baixo"e “alto"(Caiado et al., 2021).

Um SF é um sistema que utiliza a lógica *fuzzy* para mapear entradas e saídas. A estrutura do sistema de inferência *fuzzy* (SIF) compreende quatro componentes principais, conforme mostrado na Figura 3.6: (i) Fuzzificação, (ii) Raciocínio baseado em regras, (iii) Mecanismo de inferência e (iv) Defuzificação (Talpur et al., 2023).

Na etapa de Fuzzificação, as variáveis de entrada do sistema são convertidas em conjuntos *fuzzy*, atribuindo-lhes graus de pertinência. A fuzzificação mapeia valores numéricos precisos em conjuntos *fuzzy*, representando a incerteza associada às variáveis em questão. Isso é feito usando funções de pertinência, que descrevem como cada valor numérico pertence a um conjunto *fuzzy* (Talpur et al., 2023).

Na etapa do Raciocínio baseado em regras, são utilizadas regras *fuzzy* que definem o com-

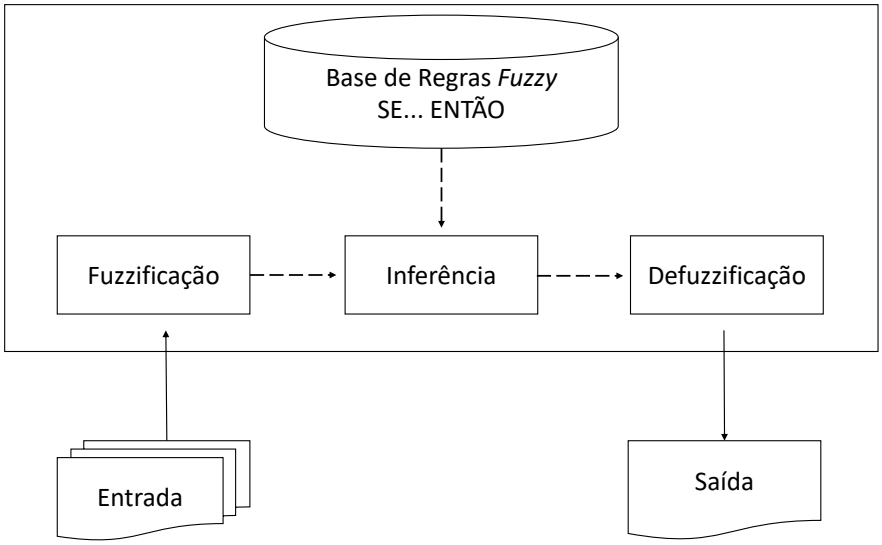


Figura 3.6: Representação do Sistemas de Inferência *Fuzzy*. Adaptado de (Talpur et al., 2022).

portamento do sistema. Essas regras são expressas na forma "Se X é A, então Y é B", onde X e Y são variáveis e A e B são conjuntos *fuzzy*. O raciocínio é baseado na lógica *fuzzy*, que permite a combinação e avaliação das regras *fuzzy* para determinar a saída do sistema (Talpur et al., 2023).

O Mecanismo de inferência é responsável por combinar as regras *fuzzy* e avaliar a contribuição de cada regra para a determinação da saída do sistema. Os dois principais métodos de inferência *fuzzy* são Mamdani e Takagi-Sugeno, amplamente utilizados (Talpur et al., 2023). O método Mamdani, introduzido por Ebrahim Mamdani, é baseado em regras *fuzzy* onde tanto as entradas quanto as saídas são variáveis linguísticas, o que torna o processo mais intuitivo e interpretável (Mamdani e Assilian, 1975). Por outro lado, o método Takagi-Sugeno, desenvolvido por Takagi e Sugeno, utiliza funções matemáticas lineares ou constantes para definir as saídas, tornando-o mais eficiente em termos computacionais, especialmente em sistemas de controle adaptativo (Takagi e Sugeno, 1985). Entretanto, devido à natureza subjetiva e imprecisa da emoção no reconhecimento de emoções, o método Mamdani foi o escolhido, pois permite lidar melhor com incertezas e nuances emocionais ao fornecer saídas mais interpretáveis e alinhadas com a subjetividade humana.

E por fim, na Defuzzificação é necessário converter a saída *fuzzy* em um valor numérico, compreensível para o usuário. Existem várias técnicas de defuzzificação, como o centro de gravidade, que calcula o valor médio ponderado das contribuições *fuzzy* para determinar o valor final de saída (Talpur et al., 2023).

De forma geral, o sistema de inferência *fuzzy* utiliza a fuzzificação para transformar as variáveis de entrada em conjuntos *fuzzy*, o raciocínio baseado em regras para combinar e avaliar as regras, o mecanismo de inferência para determinar a saída e a defuzzificação para converter a saída *fuzzy* em um valor numérico. Essa estrutura permite que o sistema lide com a incerteza e imprecisão, facilitando a tomada de decisões em situações complexas (Talpur et al., 2023).

CAPÍTULO 4

Módulo de Reconhecimento de Emoção da Fala

4.1 Introdução

O reconhecimento das emoções na fala (SER - *Speech Emotion Recognition*) tem se tornado cada vez mais popular no campo em crescimento da Interação Humano-Robô (IHR) (de Lope e Graña, 2023). Isso ocorre devido à importância de analisar as emoções expressas na fala, a fim de elevar o nível de inteligência dos sistemas de conversação robóticos. Ao interagir com as pessoas e compreender as emoções transmitidas na fala, é possível oferecer serviços de melhor qualidade e criar uma experiência de interação entre humanos e computadores mais inteligente, natural e personalizada (Wang e Shen, 2023).

A importância do reconhecimento de emoções na fala se expandiu significativamente em diversos setores, abrangendo aplicações em automação residencial, atendimento ao cliente, aplicações médicas, e até o entretenimento (Baek e Lee, 2023). Em todas essas aplicações, para uma comunicação eficaz entre humanos e computadores/máquinas/robôs é necessário a compreensão das verdadeiras intenções dos humanos identificada pelas emoções da fala (Ottoni e Cerqueira, 2021).

Nos primórdios da pesquisa sobre o reconhecimento de emoções na fala, o foco estava principalmente em modelos probabilísticos, como os Modelos Ocultos de Markov e Modelos de Mistura Gaussiana. No entanto, com o advento da aprendizagem profunda, o cenário do reconhecimento de emoções mudou significativamente em direção a abordagens baseadas em redes neurais (Baek e Lee, 2023). As Redes Neurais Convolucionais (CNN - *Convolutional Neural*

Network) e Redes Neurais Recorrentes (RNN *Recurrent Neural Network*), desempenham um papel predominante no avanço do reconhecimento de emoções na fala atualmente (Kim e Lee, 2023b).

Com o progresso contínuo das técnicas de Aprendizado Profundo, surge algumas dificuldades como necessidade de ter disponíveis dados para apoiar novos avanços e permitir comparações entre modelos de previsão (de Lope e Graña, 2023). No campo da SER existem muitos conjuntos de dados pequenos, surgindo a necessidade de agrupar bancos de dados diferentes e a aplicação de técnicas de aumento de dados, do inglês, *data augmentation* (DA), que desempenham um papel fundamental enriquecendo os conjuntos de dados, permitindo que os modelos aprendam com uma ampla gama de variações na fala emocional (Baek e Lee, 2023).

Outra necessidade é o ajuste das várias configurações de hiperparâmetros, como otimizador e taxa de aprendizado. Em (Ottoni et al., 2023a) fica evidente que o ajuste dos hiperparâmetros influenciam no desempenho do modelo de aprendizado profundo, por tanto é essencial que sejam realizados teste para que os melhores hiperparâmetros sejam encontrados para cada problema. Além disso, a investigação minuciosa das diversas técnicas de extração de características é essencial no campo do reconhecimento de emoções na fala, uma vez que as características extraídas desempenham um papel crítico na qualidade e no desempenho dos modelos (Ahmed et al., 2023).

Por último, mas não menos importante, é importante explorar uma variedade de arquiteturas neurais, como CNNs, RNNs, e arquiteturas híbridas, para determinar quais proporcionam o melhor desempenho e generalização em diferentes conjuntos de dados e cenários de aplicação. Essa abordagem abrangente é essencial para impulsionar o progresso na área do reconhecimento de emoções na fala (Wang e Shen, 2023).

É notório que a otimização de um grande número de combinações requer um alto tempo computacional antes que uma boa solução seja encontrada (Mantovani et al., 2019; Aguiar et al., 2022; Khare et al., 2023). Nesse contexto, uma abordagem recentemente destacada e que tem atraído considerável atenção é o uso de *meta-learning* (MtL). O MtL envolve a construção de estratégias que permitem aos modelos aprender com conjuntos de dados diversos e transferir esse conhecimento para tarefas relacionadas, acelerando o processo de aprendizagem e melhorando a generalização (Brazdil et al., 2022). Um exemplo é a transferência dos hiperparâmetros para bases de dados diferentes de um mesmo problema (Reif et al., 2012).

No entanto, apesar da quantidade considerável de trabalhos em SER, observam-se lacunas nos estudos da literatura existente. A literatura ainda carece de uma abordagem baseada em aprendizagem profunda que explore as combinações entre otimizador, taxas de aprendizagem, técnicas de *data augmentation*, extração de características e arquitetura neural. A maioria dos trabalhos apresenta comparações limitadas, focando em aspectos específicos, como arquiteturas neurais ou técnicas de extração de características isoladamente. Além disso, é notável a ausência de estudos que utilizem MtL e otimização de hiperparâmetros, como otimizadores e taxas de aprendizado. Esses aspectos servem de motivação para a realização deste trabalho.

Desta forma, este Capítulo tem como propósito relatar o desenvolvimento de uma abordagem de aprendizado profundo que investiga as melhores configurações de otimizadores, taxas de aprendizado, técnicas de *data augmentation*, métodos de extração de características e arquiteturas neurais. Para isso, a abordagem é aplicada em quatro bases de dados: RAVDESS, TESS, SAVEE e R+T+S (RAVDESS+TESS+ SAVEE). Uma vez que a melhor combinação seja identificada, é então realizado o *meta-learning*, i.e, a transferência da melhor configuração para duas bases de dados adicionais: CREMA-D e R+T+S+C (RAVDESS+TESS+ SAVEE+CREMA-D).

Em resumo, as principais contribuições deste estudo são:

- Propor uma abordagem de aprendizado profundo que investiga as melhores configurações de otimizadores, taxas de aprendizado, *data augmentation*, extração de características e a arquitetura neural para diferentes conjuntos de dados.

- Aplicação do *meta-learning*, transferindo a melhor configuração encontrada para o otimizador, taxa de aprendizado, *data augmentation*, extração de características e arquitetura neural para dois outros conjuntos de dados SER.

- Testes dos modelos otimizados em uma base de dados áudio-visual.

Este capítulo está estruturado da seguinte forma: A seção 4.2 apresenta uma revisão dos trabalhos na área do reconhecimento de emoções da fala que realizam investigações para encontrar a melhor configuração SER. A seção 4.3 define a abordagem completa para obtenção dos resultados para cada base de dados e a aplicação do *meta-learning* proposto. Os resultados são apresentados e discutidos na seção 4.4. E por fim, na seção 4.5 o módulo de reconhecimento de emoção da fala é testado utilizando a base de dados MELD.

4.2 Trabalhos Relacionados

O campo da pesquisa SER é altamente dinâmico e tem visto muitas inovações e avanços significativos ao longo do tempo. Especificamente, técnicas avançadas de aprendizado profundo trouxeram progressos substanciais nesta área. A implementação desses algoritmos gerou a necessidade de investigar e ajustar configurações para o melhor desempenho do modelo.

Esta seção apresenta uma revisão dos trabalhos relevantes no campo do Reconhecimento de Emoção da Fala (SER). Neste contexto, exploramos uma série de trabalhos recentes que se concentram em investigar a combinação ótima das configurações do SER. Esses artigos, realizam a comparação de pelo menos um parâmetro SER, eles são: otimizador, taxa de aprendizado, extração de características, *data augmentation* (DA), arquitetura neural e *meta-lerning* (MtL). Na Tabela 4.1, pode-se observar os artigos e as configurações que são investigadas em seus trabalhos.

Tabela 4.1: Trabalho relacionado na área SER que investiga as melhores combinações de otimizadores, taxas de aprendizado, extração de características, arquiteturas neurais e meta-aprendizado. Os artigos marcados com um check indicam que realizaram uma comparação com a respectiva configuração.

Artigo	Otimiz.	Taxa de Aprend.	Extração de Caract.	DA	Arquit. Neural	MtL
Pan e Wu (2023)	-	-	-	✓	✓	-
Ahmed et al. (2023)	-	-	-	✓	✓	-
Asiya e Kiran (2021)	-	-	-	✓	-	-
Bautista et al. (2023)	-	-	-	✓	✓	-
Bhangale e Kothandaraman (2023)	✓	-	✓	-	-	-
Chitre et al. (2022)	-	-	-	-	✓	-
Gupta et al. (2022)	-	-	✓	✓	✓	-
Jothimani e Premalatha (2022)	-	-	-	✓	✓	-
Proposta	✓	✓	✓	✓	✓	✓

Em Pan e Wu (2023), os autores utilizam a base de dados RAVDESS, utilizando a técnica MFCC para extrair as características dos áudios. O artigo realiza uma comparação entre o uso de *data augmentation*, acrescentando ruído e mudança de tonalidade, além disso, é realizado uma comparação da arquitetura neural entre uma CNN e CNN+LSTM. Os autores não detalham como foi a escolha do otimizador e taxa de aprendizado e não informam qual desses hiperparâmetros formam utilizados.

Já em Ahmed et al. (2023), os autores também comparam o uso e não uso do *data augmentation* e a arquitetura neural. As bases de dados utilizadas foram a RAVDESS, TESS, SAVEE, CREMA-D e EMO-DB. As técnicas comparadas no *data augmentation* foram o uso de ruído,

mudança de tom e alongamento, a comparação realizada foi entre não utilizar DA e utilizar todas as técnicas juntas. No que diz respeito a comparação realizada com as extrações de características, foram avaliadas de forma separada o método Mel, ZCR, RMS, MFCC e Chromagram. Além disso, os autores também realizam uma investigação quanto a arquitetura neural, avaliando entre CNN, CNN+LSTM e CNN+GRU. Foi utilizado o otimizador Adam com a taxa de aprendizado ajustável, não realizando investigações neste sentido.

Em Asiya e Kiran (2021), é utilizado uma CNN 1D para classificar as emoções da fala. As bases de dados utilizadas são a RAVDESS e TESS. Para extrair as características dos áudios, é utilizado as técnicas MFCC, Mel, Chroma e ZCR em conjunto. As operações de *data augmentation* utilizadas são ruído, mudança de tom e alongamento em conjunto. Os hiperparâmetros utilizados são o otimizador Adam com taxa de aprendizado adaptável.

Já em Bautista et al. (2023), os autores realizam uma comparação entre diferentes tipos de *data augmentation* e diferentes arquiteturas neurais. Os DA utilizados para comparação foram as técnicas de Ruído Gaussiano, *SpecAugment*, *Room Impulse Response* (RIR), e *Tanh Distortion*. As arquiteturas neurais comparadas foram a CNN 2D, CNN+BiLSTM+Attention e CNN+Transformer. As informações de extração de características utilizadas pelos modelos foi o espectrograma mel. A proposta do artigo foi avaliada utilizando a base de dados RAVDESS. Não foram apresentadas informações sobre o otimizador e a taxa de aprendizado.

O trabalho de Bhangale e Kothandaraman (2023) utiliza a base de dados RAVDESS para avaliar seu modelo composto por uma rede neural convolucional. O trabalho realiza comparações entre as técnicas de extração de características e otimizadores. Neste sentido, são comparados as técnicas de extrair características como MFCC, LPCC, WPT, ZCR, RMS e outras. Já os otimizadores utilizados foram SGD, Adam e RMSProp com taxa de aprendizado igual a 0,001.

O artigo de Chitre et al. (2022), compara três arquiteturas de Redes Neurais Convolucionais 2D conhecidas na literatura, a AlexNet, VGG16 e ResNet50. Utiliza das bases de dados RAVDESS e a junção entre RAVDESS+TESS+SAVEE+CREMAD para avaliar o modelo. Além disso, como técnica de extração de características utiliza imagens de espectrograma mel sem ruído. Os autores não informam qual otimizador e taxa de aprendizado foram utilizados. Não foi utilizado técnicas de *data augmentation*.

Em Gupta et al. (2022), utilizou-se a junção entre as bases RAVDESS+TESS+ SAVEE+CREMAD

(R+T+S+C) para avaliar como as diferentes técnicas de extração de características influenciam no desempenho do modelo. Para isso, foi utilizado MFCC, mel, chromagram, ZCR, RMS e *roll off*. Também realizaram o uso do *data augmentation* aplicando ruído, mudança de tom e alongamento. Os algoritmos comparados foram CNN, SVM, MLP, LSTM e CNN+LSTM. O otimizador utilizado foi o Adam com taxa de aprendizado variável.

Por fim, em Jothimani e Premalatha (2022), os autores avaliam duas arquiteturas neurais, uma CNN e uma CNN+LSTM. Também é realizado uma comparação entre o uso do DA com o acréscimo de ruído e mudança de tom. As técnicas de extração de características utilizadas foram as ZCR, RMS e MFCC combinadas. Para avaliar o desempenho da metodologia, utilizaram as bases de dados, RAVDESS, TESS, SAVEE, CREMA-D e a combinação entre elas (R+T+S+C). Foi utilizado o otimizador SGD com a taxa de aprendizado igual a 0,001.

Com o levantamento realizado, descrito na Tabela 4.1, é possível observar algumas tendências e lacunas em alguns trabalhos da literatura do SER. Dos 8 trabalhos levantados, seis realizam uma investigação do uso do *data augmentation*, em sua maioria, os trabalhos comparam o uso e não uso das técnicas de DA, que são geralmente ruído, mudança de tom e alongamento. Apenas os trabalhos de Pan e Wu (2023) e Bautista et al. (2023), realizam uma investigação testando as operações separadamente e em combinações. Outro aspecto que pode ser observado é a tendência em se avaliar e testar diferentes arquiteturas neurais para o problema SER. Dos oito trabalhos mencionados, seis realizam uma investigação da melhor arquitetura. Dentre as mais utilizadas estão o algoritmo CNN, e a combinação da CNN com LSTM ou GRU.

Já com relação as lacunas, o primeiro ponto que pode-se perceber a partir da Tabela 4.1 é que os trabalhos citados da área SER não realizam os ajustes dos hiperparâmetros da rede, como analisar o melhor otimizador e taxa de aprendizado. Apenas Bhangale e Kothandaraman (2023) investigou o melhor otimizador, e nenhum trabalho analisou a taxa de aprendizado. O segundo aspecto que pode-se observar é a falta de trabalhos que avaliam e testam a melhor técnica de extrair características dos áudios para o problema SER. Dentre os oito trabalhos mencionados, apenas dois artigos (Bhangale e Kothandaraman (2023) e Gupta et al. (2022)) investigam a melhor extração de características. Os demais, utilizam uma ou mais de uma técnica combinadas, sem avaliar. O terceiro ponto é com relação ao uso do *meta-learning*. Nenhum trabalho realiza transferência de configurações entre bases de dados para o problema do reconhecimento de emoções na fala. Portanto, este trabalho tem como principais motivações

atender a essas lacunas apontadas pelos trabalhos da Tabela 4.1.

4.3 Metodologia Proposta

Diante das lacunas identificadas na literatura, o propósito deste estudo é apresentar uma abordagem de aprendizado profundo destinada a identificar as combinações mais eficazes para o reconhecimento das emoções da fala. Para garantir a replicabilidade e explicabilidade da abordagem, elaborou-se uma sequência de passos em que deixa claro a tomada de decisão em cada etapa. Como ilustrado na Figura 4.1, a abordagem começa por escolher as bases de dados a serem empregadas na avaliação do desempenho do método. Neste contexto, foram escolhidas as bases RAVESS, TESS, SAVEE e CREMA-D.

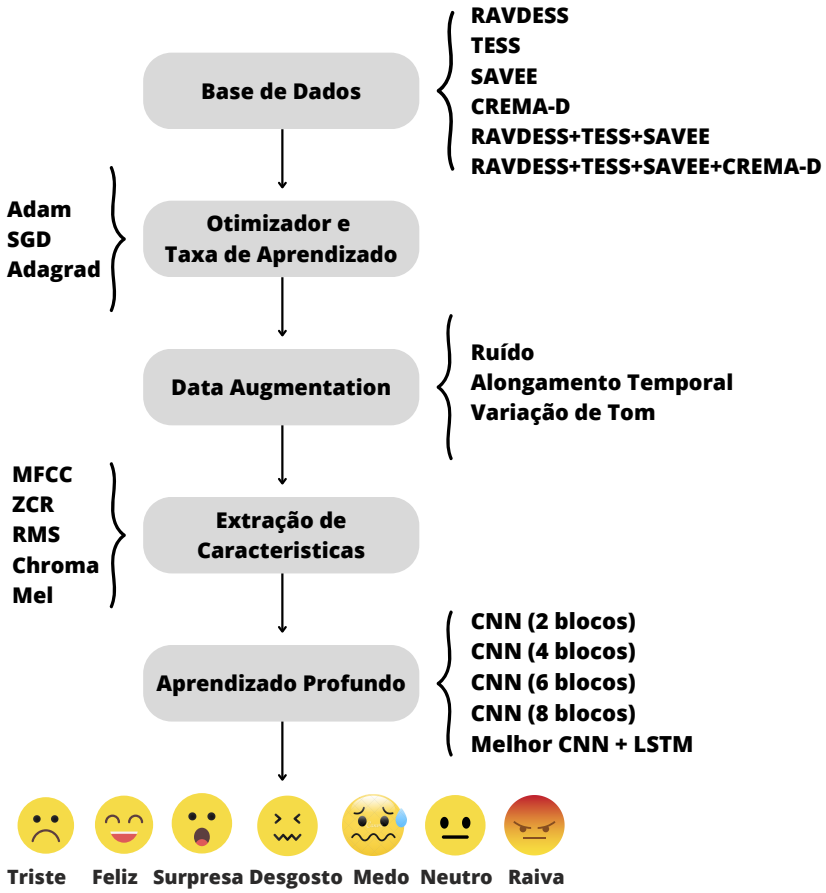


Figura 4.1: Fluxograma da abordagem proposta para buscar a melhor configuração de reconhecimento de emoções de fala.

Para conduzir esta pesquisa, foi estabelecida uma estrutura padrão que envolve a utilização de uma rede neural convolucional (CNN) com 2 blocos (a arquitetura está detalhada na Seção 4.3.5). Para a extração de características, foi empregado o método MFCC, e não foi aplicada

a técnica de *data augmentation*. Com base nessa configuração padrão, a primeira análise a ser realizada diz respeito à influência do otimizador e da taxa de aprendizado. Foram avaliados três otimizadores diferentes: Adam, SGD e Adagrad, juntamente com três taxas de aprendizado distintas: 0,01, 0,001 e uma taxa ajustável.

Após a seleção do otimizador mais adequado e da melhor taxa de aprendizado para cada conjunto de dados, a segunda fase da pesquisa aborda a implementação do *data augmentation*. Essa técnica é empregada com o objetivo de aumentar a diversidade dos áudios, melhorando assim a capacidade de generalização da CNN. Nesse contexto, foram avaliadas diferentes técnicas que incluem a adição de ruído, o mudança de tonalidade, o alongamento temporal bem como a combinação de todas essas técnicas em conjunto.

Em seguida, realizou-se a avaliação para determinar qual seria a melhor técnica de extração de características de áudio para cada conjunto de dados. Neste estudo, foram testados os seguintes métodos de extração de características: *Mel frequency cepstral coefficient* (MFCC), *Zero Crossing Rate* (ZCR), *Root Mean Square* (RMS), *Chromagram*, e espectrograma Mel.

Por fim, a última etapa envolve a avaliação das arquiteturas de aprendizado profundo. Nesse contexto, foram propostas quatro configurações diferentes para a rede neural convolucional (com 2 blocos, 4 blocos, 6 blocos e 8 blocos). A CNN que apresentar o melhor desempenho será integrada ao algoritmo LSTM. Portanto, serão investigadas qual das arquiteturas da CNN é a mais eficaz e qual é o desempenho quando se utiliza uma abordagem híbrida (CNN+LSTM).

Uma vez definidas as melhores combinações envolvendo otimizador, taxa de aprendizado, *data augmentation*, métodos de extração de características e arquitetura de aprendizado profundo, essas configurações serão submetidas a uma avaliação do *meta-learning*. Isto é, a combinação que demonstrar o melhor desempenho geral será transferida para a aplicação nas bases de dados CREMA-D e RAVDESS+TESS+ SAVEE+CREMA-D. Isso permitirá que a abordagem mais eficaz seja aplicada a diferentes conjuntos de dados, e permitindo verificar se o *meta-learning* é uma boa solução para o SER. A seguir serão apresentados com mais detalhes as etapas da metodologia proposta.

4.3.1 Bases de Dados

A escolha de um conjunto de dados desempenha um papel fundamental no reconhecimento das emoções da fala. A seleção do conjunto de dados impacta diretamente as capacidades de

treinamento e generalização do modelo. Idealmente, o conjunto de dados deve abranger um amplo espectro de expressões emocionais, vários fatores demográficos e diversos estilos de fala para garantir a robustez e aplicabilidade do modelo em cenários do mundo real (Ahmed et al., 2023). No entanto, encontrar bases de dados com uma grande diversidade disponíveis online não é fácil.

Nesse sentido, este trabalho utiliza quatro bases de dados amplamente utilizadas na literatura: RAVDESS, TESS, SAVEE e CREMA-D, para avaliar o desempenho da abordagem proposta. Além disso, para aumentar a disponibilidade de dados para algoritmos de aprendizagem profunda, os bancos de dados também foram combinados em dois conjuntos: R+T+S (RAVDESS+TESS+ SAVEE) e R+T+S+C (RAVDESS+TESS+SAVEE+CREMAD). A seguir, são descritos informações mais detalhadas sobre os bancos de dados usados.

RAVDESS: O *Ryerson Audiovisual Database of Emotional Speech and Music* (RAVDESS) (Livingstone e Russo, 2018) é um recurso amplamente utilizado no reconhecimento de emoções de fala. É composto por gravações de 24 atores profissionais, divididos igualmente entre 12 mulheres e 12 homens. Esses atores fazem duas declarações cada, cantando e falando. Os áudios duram 3 segundos e são rotulados com emoções, feliz, triste, raiva, medo, surpresa, neutro, calmo e desgosto, cada um apresentado em dois níveis de intensidade emocional: normal e forte, totalizando 2076 gravações de áudio. Neste trabalho, foi retirada a emoção calma do banco de dados para padronizá-la em sete emoções. A Tabela 4.2 descreve a distribuição do áudio por emoção.

TESS: O *Toronto Emotional Speech Set* (TESS) (Pichora-Fuller e Dupuis, 2020) traz gravações de duas atrizes inglesas, uma de 26 e outra de 64 anos. Os áudios duram dois segundos; as emoções rotuladas são raiva, desgosto, medo, feliz, neutro, surpresa e triste. O conjunto de dados consiste em 2800 arquivos de áudio, com 400 gravações de áudio alocadas para cada categoria de emoção, conforme ilustrado na Tabela 4.2. Vale ressaltar que esta base de dados é balanceada, garantindo igual número de arquivos de áudio para cada categoria de emoção.

SAVEE: O SAVEE (*Surrey Audio–Visual Expressed Emotion dataset*) (Jackson e Haq, 2014) consiste em 480 áudios falados por quatro atores ingleses com idades entre 27 e 31 anos. Os áudios duram em média 3 segundos e são rotulados com sete emoções: raiva, feliz,

neutro, desgosto, triste, medo e surpresa. No entanto, é importante notar que este conjunto de dados apresenta um problema de desequilíbrio de classes. Especificamente, a classe “neutra"contém quase o dobro de amostras que todas as outras classes combinadas, conforme descrito na Tabela 4.2.

CREMA-D: O *Crowdsourced Emotional Multimodal Actors Dataset*(CREMA-D) (Cao et al., 2014) abrange 7442 amostras de áudio, todas gravadas por 91 atores representando diversas origens raciais e étnicas. Dentre esses atores, 48 eram do sexo masculino e 43 do sexo feminino; cada um pronunciou 12 frases. Os áudios duram em média 2 segundos e expressam seis emoções distintas: raiva, feliz, neutro, desgosto, triste e medo. Na Tabela 4.2 é possível observar a quantidade de gravações de áudio para cada emoção dentro do banco de dados.

Tabela 4.2: Número de áudios em cada base de dados.

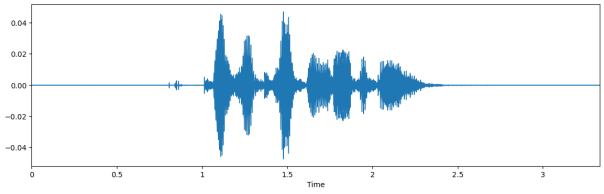
Base de dados	Feliz	Triste	Raiva	Medo	Desgosto	Surpresa	Neutro	Total
RAVDESS	376	376	376	376	192	192	188	2076
TESS	400	400	400	400	400	400	400	2800
SAVEE	60	60	60	60	60	60	120	480
CREMA-D	1271	1271	1271	1271	1271	0	1087	7442
R+T+S	836	836	836	836	652	652	708	5356
R+T+S+C	2107	2107	2107	2107	1923	652	1795	12798

4.3.2 Passo 1: Ajuste dos otimizadores e taxa de aprendizado

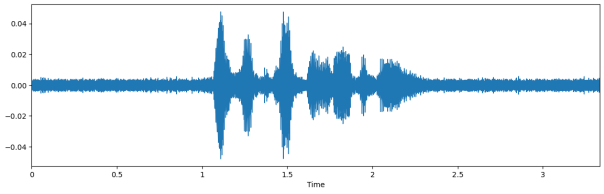
Conforme descrito na Seção 3.2.2, os hiperparâmetros de aprendizado e otimização determinam como a rede aprende e otimiza seus parâmetros para atingir o erro mínimo. Entre eles estão algoritmos de otimização e taxas de aprendizagem (Elgendy, 2020). Devido à importância desses hiperparâmetros para o modelo CNN, a primeira investigação do modelo proposto busca o melhor otimizador e taxa de aprendizado para cada banco de dados utilizado. Para tanto, serão utilizados os três otimizadores mais utilizados na literatura, Adam, SGD e Adagrad, sendo as taxas de aprendizado mais encontradas em artigos da área SER (0,01, 0,001, e taxa de aprendizado adaptativa com fator de 0,4 e um LR mínimo de 0,000001) (Jothimani e Premalatha, 2022; Gupta et al., 2022).

4.3.3 Passo 2: Otimização do *data augmentation*

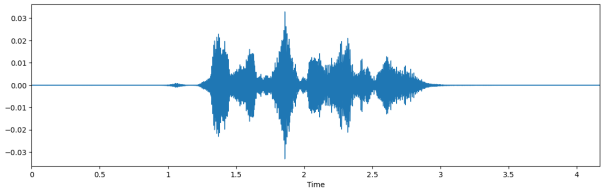
A segunda fase da abordagem proposta envolve a investigação do impacto do *data augmentation* no desempenho do algoritmo de aprendizagem profunda. O DA gera novas amostras de treinamento sintéticas por meio de pequenas perturbações nos exemplos existentes. O objetivo é tornar o modelo invariante a essas perturbações e aumentar sua capacidade de generalização. Várias técnicas são usadas para aumentar os dados de áudio, sendo algumas das mais comuns incluindo ruído, alongamento de tempo e variação de tom (Gupta et al., 2022). A representação visual dos efeitos destas técnicas é ilustrada na Figura 4.2.



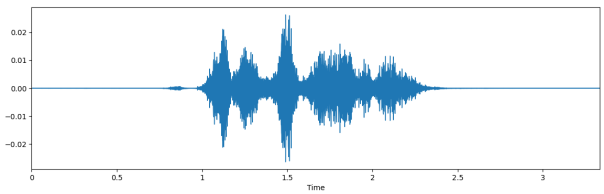
(a) Original audio



(b) Ruído



(c) Alongamento temporal



(d) Variação de tom

Figura 4.2: Exemplo de modificações usadas para *data augmentation*.

Neste estudo, a técnica de injeção de ruído foi empregada para introduzir valores aleatórios nos dados usando os métodos normal e uniforme do NumPy com uma taxa de 0,035 (Dolka et al., 2021). A técnica de alongamento também foi usada para alongar séries temporais a uma

taxa fixa de 0,8, implementada através do método time-stretching da biblioteca Python Librosa (Ahmed et al., 2023). Por fim, foram aplicadas alterações aleatórias de mudança de tom com fator de 0,7, utilizando o método de pitch shifting de Librosa (Ahmed et al., 2023). As técnicas de DA foram testadas individualmente e em combinação.

4.3.4 Passo 3: Seleção da técnica de extração de características

A extração de características de sinais de áudio constitui uma etapa fundamental nas atividades de reconhecimento de emoções de fala (SER) (Ashok et al., 2022). A terceira etapa desta investigação emprega especificamente os cinco atributos espectrais mais usados na pesquisa SER (Ahmed et al., 2023; Asiya e Kiran, 2021; Gupta et al., 2022; Jothimani e Premalatha, 2022). Eles são os valores do *Mel Frequency Cepstral Coefficient* (MFCC), *Zero Crossing Rate* (ZCR), *Root Mean Square* (RMS), *Chromagram* e espectrograma Mel. Mais detalhes de cada técnica de extração de recursos serão apresentados abaixo.

Mel Frequency Cepstral Coefficient (MFCC)

Para extrair as características MFCC, a etapa inicial envolve dividir o sinal de fala em pequenos quadros de 20 a 30 ms cada, avançados a cada 10 ms para capturar características temporais de sinais de fala individuais. A Transformada Discreta de Fourier (DFT) é posteriormente aplicada a cada quadro de janela, convertendo-os em espectros de magnitude. A seguir, 26 filtros são empregados no sinal obtido na etapa anterior para calcular o banco de filtros em escala Mel (MSFB). O MSFB, baseado na percepção de frequência do ouvido humano, produz 26 valores que descrevem a energia de cada quadro. As energias logarítmicas são calculadas para obter as energias do banco de filtros logarítmicos. A Equação (4.1) quantifica a estimativa de Mel a partir de uma frequência física (Ahmed et al., 2023; Gupta et al., 2022; Singh et al., 2023):

$$f_{Mel} = 2590 \log_{10} 1 + \frac{f}{700} \tag{4.1}$$

Aqui, f denota a frequência física (em Hz) e f_{Mel} representa a frequência percebida pelo ouvido humano. Após a obtenção das energias do banco de filtros logarítmicos, a Transformada Discreta de Cosseno (DCT) é aplicada para gerar os MFCCs (Ahmed et al., 2023; de Lope e

Graña, 2023). A extração dos valores de MFCC dos conjuntos de dados foi realizada utilizando a biblioteca Librosa.

Zero Crossing Rate (ZCR)

O ZCR é um recurso comumente usado no SER. Ele quantifica o número de vezes que a amplitude de um sinal de voz cruza o limite do valor zero dentro de um período de tempo especificado. O ZCR provou ser eficaz na distinção entre expressões sonoras e surdas. Matematicamente, o ZCR é definido pela Equação (4.2), onde S representa um sinal de comprimento T, e $1_{\mathbb{R}<0}$ é uma função indicadora. Os valores ZCR dos conjuntos de dados foram extraídos usando a biblioteca Librosa (Ahmed et al., 2023).

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}<0}(S_t S_{t-1}) \tag{4.2}$$

Chromagram

O recurso *Chromagram* (Chroma) caracteriza o conteúdo tonal de um sinal de áudio, intimamente relacionado às 12 classes de altura. Os recursos Chroma são excelentes na captura de características de áudio harmônicas e melódicas. Os recursos do Chromagram são derivados da aplicação da *Short Time Fourier Transform* (STFT) à forma de onda de áudio do conjunto de dados (Nasim et al., 2021). A extração dos valores de Chroma dos conjuntos de dados foi realizada utilizando a biblioteca Librosa.

Mel Espectrograma

Um espectrograma visualiza o espectro de frequência de um sinal ao longo do tempo por meio da análise *Fast Fourier Transform* (FFT). Ele divide o espectro de frequência em frequências da escala Mel, produzindo um espectrograma Mel para cada janela. Os componentes de magnitude correspondentes às frequências Mel são então isolados (Nasim et al., 2021; Gupta et al., 2022). Neste estudo, esses valores foram extraídos dos conjuntos de dados utilizando a biblioteca Librosa.

Root Mean Square (RMS)

O valor RMS é calculado para cada quadro de amostras de áudio de fala, oferecendo uma amplitude média do sinal, independentemente dos níveis de amplitude positivos ou negativos. Para um determinado sinal $x = x_1, x_2, x_3, \ldots, x_n$, o valor RMS x_{RMS} pode ser determinado usando a Equação (4.3) (Ahmed et al., 2023). Os valores RMS foram extraídos dos conjuntos de dados usando a biblioteca Librosa.

$$x_{RMS} = \sqrt{\frac{x^2}{n}} = \sqrt{\frac{1}{n}(x_1^2, x_2^2, x_3^2, \ldots, x_n^2)} \tag{4.3}$$

4.3.5 Passo 4: Investigação da Arquitetura Neural

A quarta etapa da metodologia proposta envolve investigar a melhor arquitetura. Nesse sentido, foram utilizados dois algoritmos bastante conhecidos na literatura: a Rede Neural Convolucional (*Convolutional Neural Network* - CNN) e a Memória Longa e de Curto Prazo (*Long Short Term Memory* - LSTM). A seguir, forneceremos mais informações sobre esses dois algoritmos de aprendizado profundo e como eles serão usados neste estudo.

Redes Neurais Convolucionais

Neste estudo, utilizou-se as Redes Neurais Convolucionais (CNNs) para classificar emoções com base em dados de fala. Serão investigadas as arquiteturas mais eficazes para cada conjunto de dados para atingir esse objetivo. As arquiteturas propostas estão ilustradas na Figura 4.3. Inicialmente, foram realizados testes utilizando a arquitetura da CNN de dois blocos. Na fase final da metodologia proposta, é investigado o número ideal de blocos para melhor desempenho do modelo, considerando 2 blocos, 4 blocos, 6 blocos e 8 blocos.

Cada bloco CNN inclui uma camada convolucional (1D) com ativação ReLu, *batch normalization*, *max pooling* 1D (tamanho = 5) e *dropout* (taxa = 0,2). O *batch normalization* é uma camada que normaliza as entradas aplicando uma transformação que mantém a saída média próxima de 0 e o desvio padrão da saída próximo de 1. O *max pooling* é uma técnica usada para reduzir a dimensionalidade espacial da representação do recurso e para manter as características mais relevantes. Além disso, o *dropout* é uma técnica de regularização para evitar *overfitting* em redes neurais.

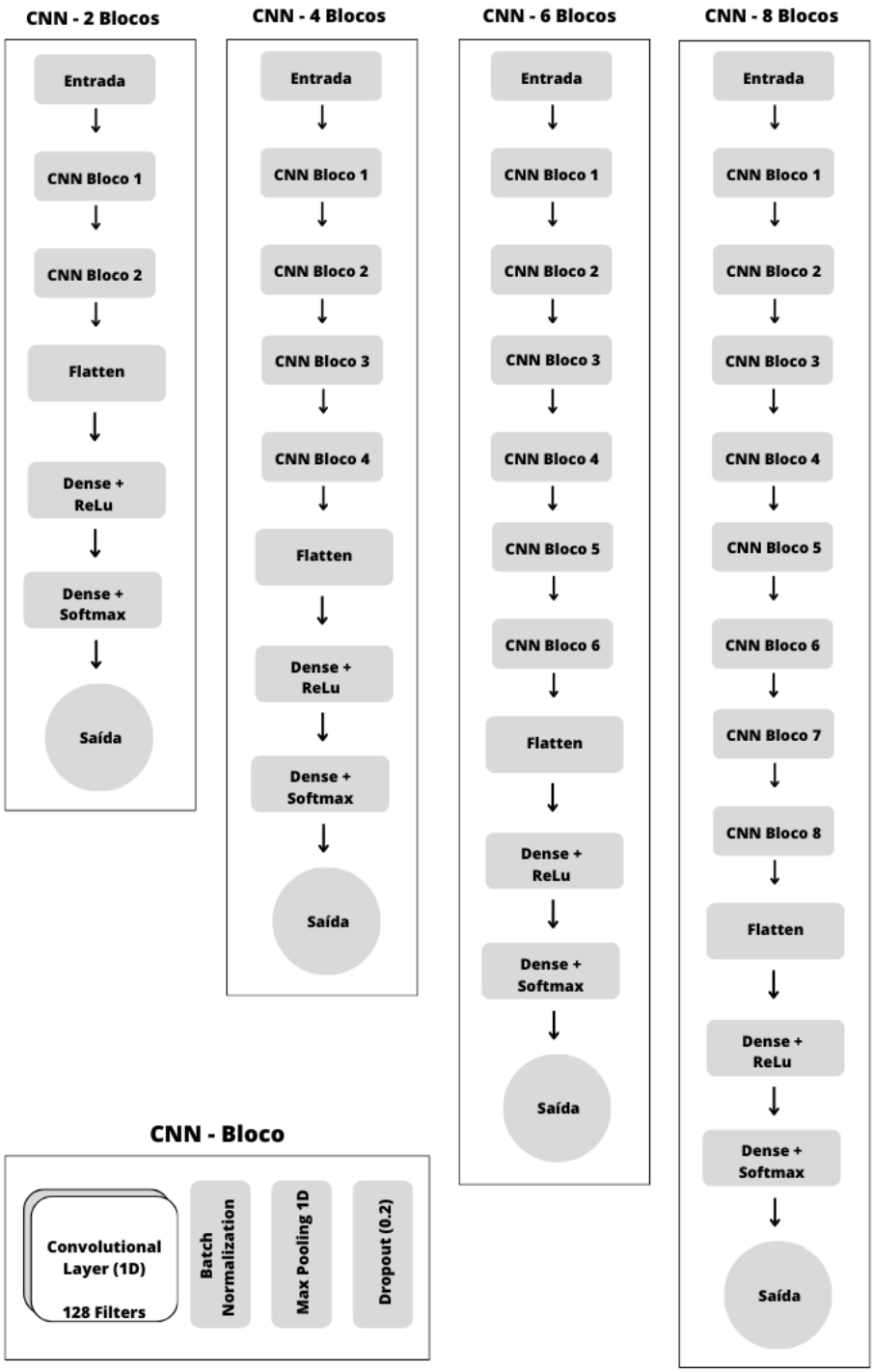


Figura 4.3: Arquiteturas de CNNs para comparação.

Ao final dos blocos CNN, uma camada *flatten* e duas camadas densas foram adicionadas para realizar a classificação final. A primeira camada densa contém a função de ativação ReLu e a camada densa de saída contém a função de ativação softmax.

Neste trabalho foram utilizados hiperparâmetros fixos, incluindo 100 épocas e um *batch size* de 64. Para a função de perda, empregamos entropia cruzada categórica, que pode ser

representada pela Equação (4.4) (Elgendy, 2020), onde $L(y, \hat{y})$ é o valor da função de perda, y_i representa a probabilidade real da classe i (um valor binário, 0 ou 1, indicando a classe correta), e $\hat{y}_i$ representa a probabilidade prevista para a classe i pelo modelo.

$$L(y, \hat{y}) = -\sum_{i} y_i \log(\hat{y}_i) \tag{4.4}$$

Memória Longa e de Curto Prazo

O Memória Longa e de Curto Prazo (*Long Short-Term Memory* -LSTM) é uma arquitetura de rede neural recorrente (RNN) que lida com dados sequenciais e é amplamente usada em processamento de linguagem natural, reconhecimento de voz e tarefas de previsão de séries temporais (Hazra et al., 2022). Após selecionar a melhor arquitetura CNN, é investigado o desempenho do modelo adicionando duas camadas LSTM. A figura 4.4 mostra que a primeira camada tem 258 unidades e a segunda camada LSTM tem 128 unidades.

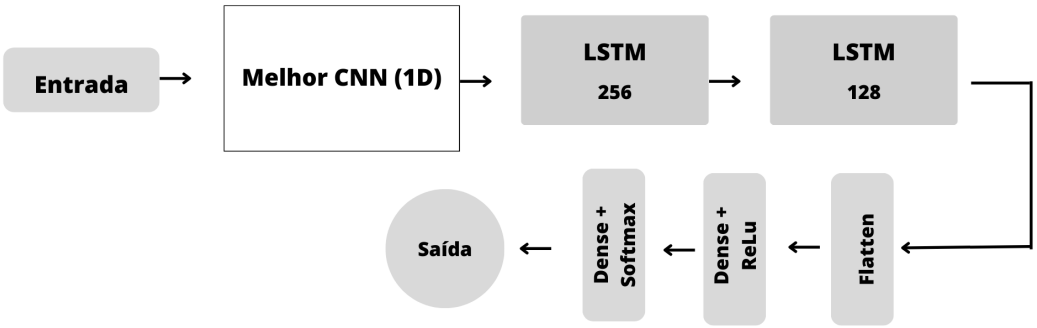


Figura 4.4: Arquitetura híbrida com Rede Neural Convolucional (CNN) e Long Short-Term Memory (LSTM).

4.3.6 Meta-Learning

O *Meta-learning* (MtL) é um processo de usar o conhecimento obtido em determinado conjunto de dados e transferi-los para novas bases de dados (Brazdil et al., 2022). O conhecimento transferido, também conhecido como meta-conhecimento, podem ser informações como arquiteturas neurais, modelos resultantes, e inclusive as configurações para se obter melhores modelos anteriores (Lemke et al., 2015). Desta forma é possível investigar como aprender com as experiências anteriores e reduzir o tempo e custo computacional gasto para ajustar o modelo (Aguiar et al., 2022).

A Figura 4.5 ilustra como foi realizado o *meta-learning* neste trabalho. Inicialmente, foram selecionados as bases de dados RAVDESS, TESS, SAVEE e R+T+S para a otimização das configurações de otimizador e taxa de aprendizado, *data augmentation*, extração de características e arquitetura neural, conforme descrito nas seções anteriores. Após encontrar a melhor configuração, i.e, o conjunto das configurações que encontraram um melhor valor de acurácia para a maioria das bases de dados, essa melhor configuração é então armazenada em uma base de conhecimento.

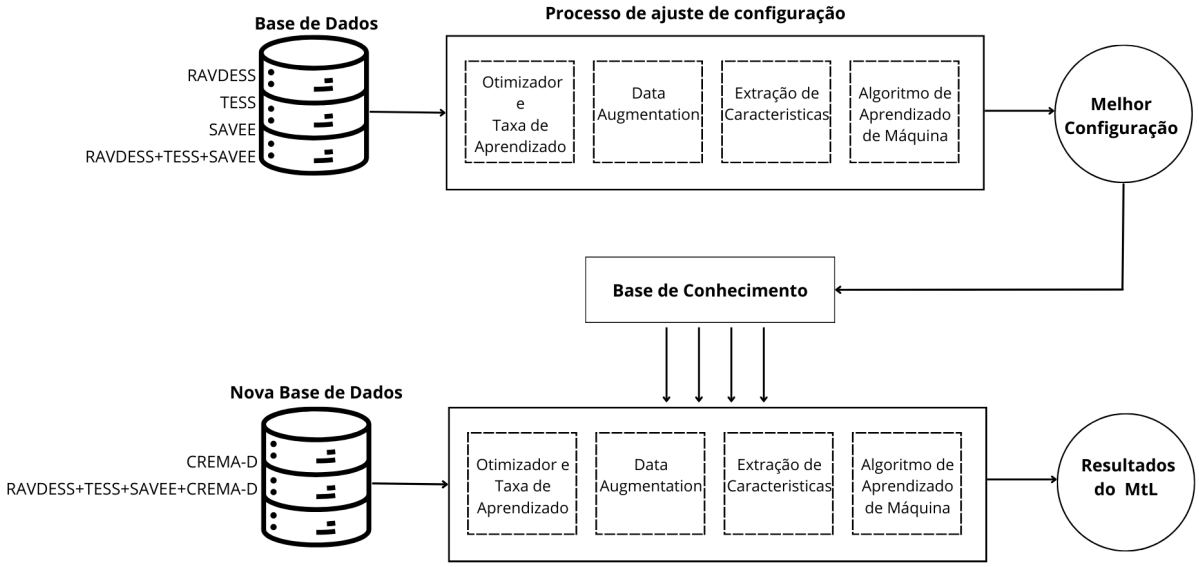


Figura 4.5: Sistema de MtL para transferir configurações SER (otimizador, taxa de aprendizagem, aumento de dados, extração de recursos e arquitetura neural) entre diferentes bancos de dados.

Após a melhor configuração ser encontrada e armazenada na base de conhecimento, é então realizado o *meta-learning*. Na qual, novos datasets CREMA-D e R+T+S+C (RAVDESS+TESS + SAVEE+CREMA-D) vão acessar as melhores configurações da base de conhecimento e aplicar as mesmas escolhas de otimizador, taxa de aprendizado, DA, extração de características e arquitetura neural. O resultado obtido por meio do MtL é então comparado com o não uso do *meta-learning*, ou seja, realizar toda a abordagem proposta para as bases CREMA-D e R+T+S+C.

4.4 Resultados

Nesta seção serão apresentados os resultados obtidos pela metodologia proposta, descrita na Seção 4.3. Para realizar os experimentos utilizou-se um notebook com sistema operacional

Windows 11, processador Intel i5 1135G7 2,40GHz, memória RAM DE 8 GB e uma GPU Nvidia GeForce MX350 com 2GB VRAM. O ambiente de desenvolvimento utilizado neste trabalho é composto pela IDE Jupyter, associada a linguagem Python.

Conforme descrito na Seção 4.3, as bases de dados utilizadas para avaliar a performance da metodologia foram as RAVDESS, TESS, SAVEE e RAVDESS+TESS+ SAVEE (R+T+S). Posteriormente, as combinações de melhor desempenho em termos da acurácia, foram transferidas para as bases CREMA-D e RAVDESS+TESS+ SAVEE+CREMA-D (R+T+S+C). Todos as bases utilizadas foram divididas em 80% dos dados para treinamento, 10% para validação e 10% para teste.

O primeiro passo, foi ajustar o otimizador e a taxa de aprendizado para cada base de dados. Neste sentido, foi utilizado os otimizadores Adam, SGD e Adagrad, combinados com as taxas de aprendizado de 0,01, 0,001 e taxa ajustável. O resultado para essa primeira investigação pode ser visto na Tabela 4.3, na qual é apresentado os valores de acurácia (%) de teste para cada base de dados avaliada.

Na Tabela 4.3, pode-se observar como a variação dos hiperparâmetros influenciam no desempenho do classificador. Ao analisar, por exemplo, a base RAVDESS percebe-se que a variação do otimizador e da taxa de aprendizado resulta em uma ampla melhoria no valor de acurácia, variando de 30,76% (com Adam 0,01) a 80,01% (com Adam variável). Esse comportamento se estende a outras bases de dados.

Tabela 4.3: Resultados de acurácia (%) obtidos ajustando o otimizador e a taxa de aprendizagem.

Dataset	Adam			SDG			Adagrad		
	0,01	0,001	variável	0,01	0,001	variável	0,01	0,001	variável
RAVDESS	30,76	78,84	**80,01**	75,07	52,88	60,09	77,88	50,00	42,78
TESS	98,21	98,57	**100,00**	99,64	98,93	98,02	99,64	99,29	99,29
SAVEE	64,58	**70,83**	68,75	66,67	43,75	62,50	58,33	54,17	56,25
R+T+S	67,54	**87,31**	87,13	86,57	77,05	84,51	86,94	76,49	74,63

Para as próximas etapas, foram utilizados as melhores combinação de otimizador e taxa de aprendizado em cada base de dados. Desta forma para base RAVDESS e TESS utilizou-se Adam variável e para SAVEE e R+T+S utilizou-se Adam 0,001.

A segunda etapa da metodologia diz respeito ao uso do *data augmentation*, nesse sentido, foram investigados três técnicas: ruído, mudança de tom e alongamento temporal, além a combinação de todas juntas. Na Tabela 4.4, são apresentados as acurácias (%) da investigação

anterior, ou seja, sem data (sem DA). Também são apresentados as acurácias (%) para cada técnica investigada. Conforme pode ser visto na Tabela 4.4, a técnica de alongamento temporal obteve um bom desempenho na maioria das bases, somente a base TESS se manteve com 100% de acurácia em todas as combinações desta etapa.

A Tabela 4.4 revela o impacto do DA, evidenciando que para o conjunto RAVDESS, a variação da acurácia vai de 77,88% (usando apenas ruído) a 96,63% (usando apenas alongamento), enquanto na base SAVEE varia de 67,50% (ruído) a 85,83%, indicando uma melhoria na acurácia. A base R+T+S tem uma melhora de aproximadamente 10% ao aplicar a operação de alongamento.

Tabela 4.4: Acurácia (%) alcançada para diferentes técnicas de DA: ruído, variação de tom e alongamento, em quatro conjuntos de dados distintos.

Dataset	sem DA	Ruído	Alongamento	Mudança de Tom	Todos DA
RAVDESS	80,01	77,88	**96,63**	86,30	85,23
TESS	100,00	100,00	100,00	100,00	100,00
SAVEE	70,83	67,50	**85,83**	80,00	81,94
R+T+S	87,31	88,99	**96,64**	93,47	92,21

Para a terceira etapa, foram utilizadas DA do tipo alongamento para todas as bases de dados. A técnica padrão para extrair as características foi o MFCC, nesta etapa, serão investigadas outras técnicas como a Chroma, ZCR, RMS, Mel e a aplicação de todas juntas.

Na Tabela 4.5 pode-se visualizar os valores de acurácia (%) de teste obtidas nesta terceira etapa. Como pode ser visto, a técnica MFCC obteve um melhor desempenho na maioria das bases de dados, com exceção para a base SAVEE que obteve maior acurácia com todas as técnicas aplicadas juntas. Além disso é possível observar que a escolha do método de extração de características causa um grande impacto da classificação SER. Na Tabela é possível observar que, por exemplo, a base RAVDESS tem uma acurácia de 20,91% ao utilizar o ZCR e de 96,63% ao utilizar o MFCC. Da mesma forma, as demais bases de dados variam dependendo da técnica de extração dos recursos.

Tabela 4.5: Acurácia (%) obtida para diferentes técnicas de extração de características, incluindo MFCC, Chroma, ZCR, RMS e Mel, em quatro conjuntos de dados distintos.

Base de Dados	MFCC	Chroma	ZCR	RMS	Mel	Todas
RAVDESS	**96,63**	63,70	20,91	30,70	67,79	94,23
TESS	**100,00**	89,82	19,82	32,14	98,57	99,08
SAVEE	85,83	52,50	32,08	32,50	70,42	**90,62**
R+T+S	**96,64**	73,41	19,96	24,63	83,68	96,36

A quarta etapa da metodologia busca investigar a melhor arquitetura da rede neural convolucional para o problema de reconhecimento de emoção da fala. Nesse sentido são testados 4 arquiteturas: com 2 blocos, 4 blocos, 6 blocos e 8 blocos, conforme descrito na Seção 4.3.5.

Na Tabela 4.6, pode ser visto os valores de acurácia (%) de teste para cada arquitetura. As bases RAVDESS e R+T+S obtiveram melhores acurácias com a CNN de 4 blocos, com acurácias de 96,88% e 97,11% respectivamente. A SAVEE, obteve uma acurácia de 90,62% com a CNN de 2 blocos. Já a base TESS não mostrou variação no valor de acurácia nesta etapa da metodologia.

Tabela 4.6: Acurácia (%) obtida para diferentes arquiteturas CNN, dois blocos, quatro blocos, seis blocos e oito blocos, em quatro conjuntos de dados diferentes.

Base de Dados	CNN (2 Blocos)	CNN (4 Blocos)	CNN (6 Blocos)	CNN (8 Blocos)
RAVESS	96,63	**96,88**	95,29	95,19
TESS	100,00	100,00	100,00	100,00
SAVEE	**90,62**	87,50	76,04	83,33
R+T+S	96,64	**97,11**	96,92	95,18

Após definir a melhor CNN para cada base de dados, ou seja, a melhor arquitetura da CNN investigada pela etapa anterior, é então adicionado duas camadas de LSTM para verificar se um algoritmo de aprendizado profundo híbrido melhora o valor da acurácia de teste. Como pode ser visto na Tabela 4.7, em que é apresentado para cada base de dados a acurácia da melhor arquitetura da CNN e a arquitetura híbrida CNN+LSTM. Para as bases RAVDESS e R+T+S a arquitetura híbrida (CNN - 4 blocos + LSTM) apresentou melhor desempenho. Para a base SAVEE a CNN de 2 blocos obteve melhor desempenho. E a base de dados TESS não apresentou nenhuma modificação.

Tabela 4.7: Acurácia (%) obtida para a melhor arquitetura CNN e arquitetura híbrida (CNN+LSTM) em quatro conjuntos de dados distintos.

Base de Dados	Melhor CNN	Melhor CNN + LSTM
RAVDESS	96.88	**97.01**
TESS	100,00	100,00
SAVEE	**90.62**	85.42
R+T+S	97.11	**97.37**

Ao completar as etapas da metodologia, foi possível descrever a melhor combinação para cada conjunto de dados, considerando o otimizador, a taxa de aprendizado, as técnicas de DA, os métodos de extração de características e a arquitetura neural que produziram o melhor desempenho. Para o conjunto de dados RAVDESS, a melhor combinação envolveu o uso do otimizador

Adam com uma taxa de aprendizado variável, aplicação de alongamento para aumento dos dados, MFCCs para extração de recursos de áudio e uma arquitetura híbrida que consiste em uma CNN de 4 blocos combinada com LSTM.

Na Figura 4.6, podemos observar os gráficos que ilustram as tendências de acurácia e perda ao longo do treinamento e validação do conjunto de dados RAVDESS. Além disso, Tabela 4.8 exibe as métricas de precisão, recall e F-score derivadas da configuração otimizada aplicada ao conjunto de dados RAVDESS.

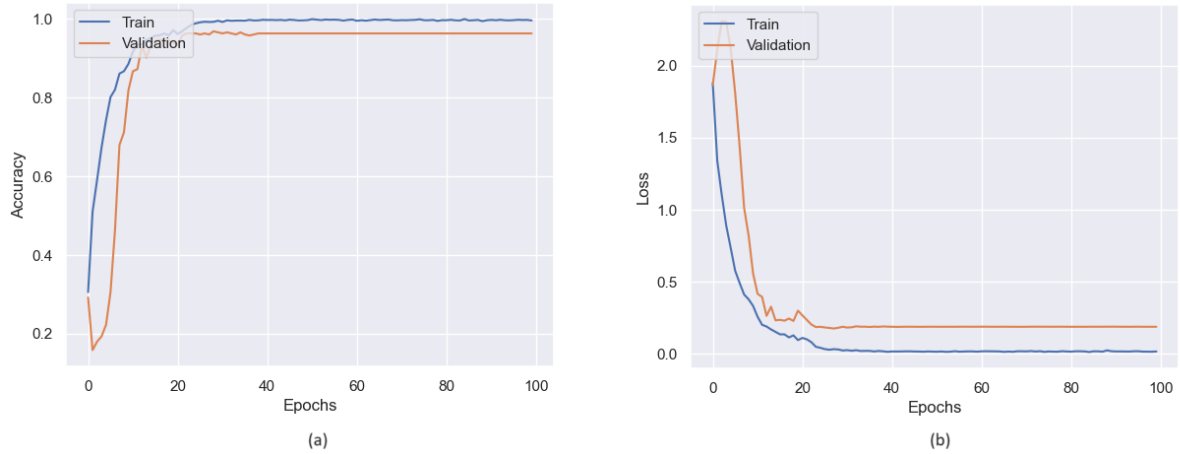


Figura 4.6: Gráfico da (**a**) acurácia e (**b**) valores de perda durante o treinamento e validação do conjunto de dados RAVDESS.

Tabela 4.8: Valores em porcentagem da Precisão, recall e F-score para a base RAVDESS.

Emoção	Precisão (%)	Recall (%)	F-Score (%)
Desgosto	95,0	93,0	94,0
Feliz	95,0	99,0	97,0
Medo	100,0	93,0	96,0
Neutro	100,0	100,0	100,0
Raiva	94,0	99,0	96,0
Surpresa	97,0	97,0	97,0
Triste	98,0	98,0	98,0
Acurácia (%)			97,0

O conjunto de dados TESS alcançou uma acurácia de 100% usando a combinação com o otimizador Adam com taxa de aprendizado variável, extração de característica MFCC, sem DA e uma arquitetura CNN de 2 blocos. Em diversas etapas, a acurácia permaneceu inalterada, levando a considerar a combinação inicial, que alcançou 100% de precisão, como a melhor combinação para a base TESS. Na Figura 4.7, pode-se observar a acurácia do treinamento e validação e as curvas de perda para o conjunto de dados TESS. Na Tabela 4.9, pode-se visualizar

as métricas de precisão, recall e F-score calculadas usando a combinação otimizada para o conjunto de dados TESS.

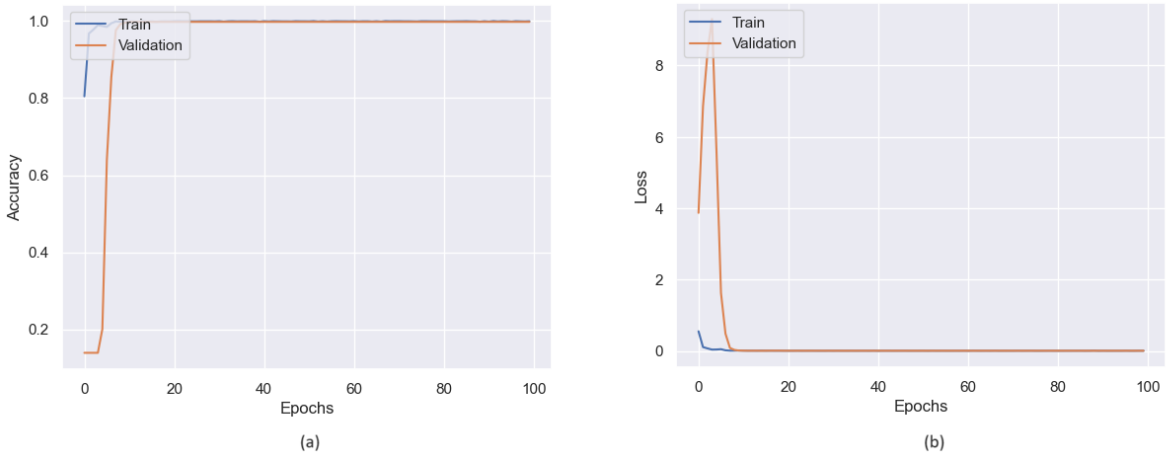


Figura 4.7: Gráfico da (**a**) acurácia e (**b**) valores de perda durante o treinamento e validação do conjunto de dados TESS.

Tabela 4.9: Valores em porcentagem da Precisão, recall e F-score para a base TESS.

Emoção	Precisão (%)	Recall (%)	F-Score (%)
Desgosto	100,0	100,0	100,0
Feliz	100,0	100,0	100,0
Medo	100,0	100,0	100,0
Neutro	100,0	100,0	100,0
Raiva	100,0	100,0	100,0
Surpresa	100,0	100,0	100,0
Triste	100,0	100,0	100,0
Acurácia (%)	-	-	100,0

A base de dados SAVEE obteve uma acurácia máxima de 90,62% com a combinação do otimizador Adam com taxa de aprendizado de 0,001, *data augmentation* do tipo alongamento, extração de características utilizando todas as técnicas em conjunto (MFCC, ZCR, RMS, Chroma e Mel), e para a arquitetura neural foi utilizado a CNN com 2 blocos. Na Figura 4.8, é apresentado os gráficos do histórico de treinamento e validação da acurácia e perda. Além disso, na Tabela 4.10 são apresentados os valores obtidos pela precisão, recall e F-score.

O conjunto de dados que combina RAVDESS + TESS + SAVEE alcançou uma precisão máxima de 97,37% com uma combinação do otimizador Adam, uma taxa de aprendizado de 0,001, DA do tipo alongamento, extração de recursos MFCC e uma arquitetura neural que consiste em um CNN híbrida com 4 blocos e LSTM. A Figura 4.9 apresenta a acurácia do treinamento e validação e o histórico de perdas, e a Tabela 4.11 exibe os valores de precisão,

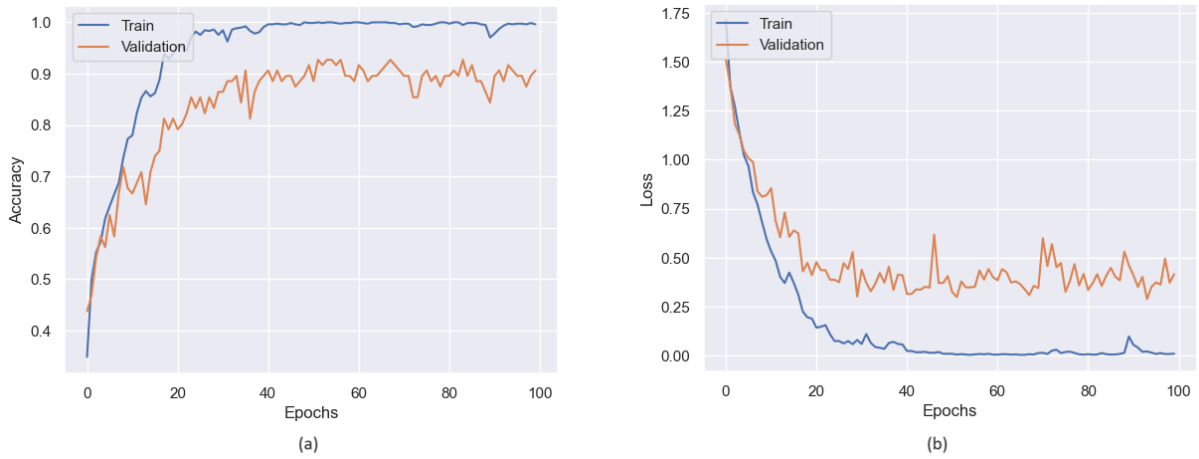


Figura 4.8: Gráfico da (**a**) acurácia e (**b**) valores de perda durante o treinamento e validação do conjunto de dados SAVEE.

Tabela 4.10: Valores em porcentagem da Precisão, Recall e F-score para a base SAVEE.

Emoção	Precisão (%)	Recall (%)	F-Score (%)
Desgosto	100,0	91,0	95,0
Feliz	86,0	86,0	86,0
Medo	92,0	100,0	96,0
Neutro	100,0	86,0	92,0
Raiva	92,0	92,0	92,0
Surpresa	93,0	87,0	90,0
Triste	73,0	100,0	85,0
Acurácia (%)	-	-	90,6

recall e F-score obtidos.

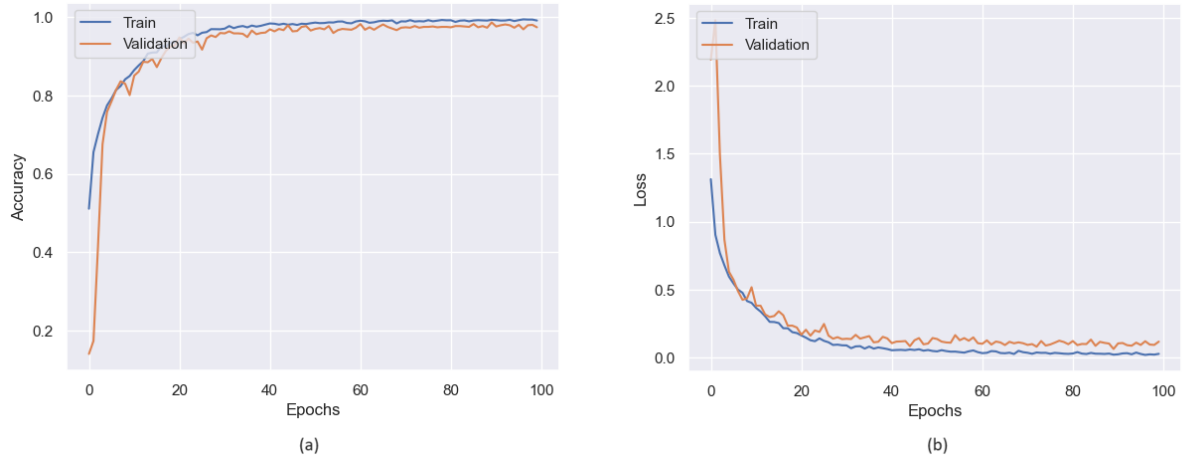


Figura 4.9: Gráfico da (**a**) acurácia e (**b**) valores de perda durante o treinamento e validação do conjunto de dados R+T+S.

Em relação ao tempo computacional gasto, o código foi executado 25 vezes para cada banco de dados, totalizando 150 execuções devido ao trabalho com seis bases de dados. A duração

Tabela 4.11: Valores em porcentagem da Precisão, Recall e F-score para a base RAVDESS+ TESS+ SAVEE.

Emoção	Precisão (%)	Recall (%)	F-Score (%)
Desgosto	98.0	96.0	97.0
Feliz	96.0	96.0	96.0
Medo	95.0	95.0	95.0
Neutro	95.0	100.0	97.0
Raiva	97.0	95.0	96.0
Surpresa	96.0	99.0	97.0
Triste	98.0	93.0	95.0
Acurácia (%)	-	-	97.4

de cada execução varia principalmente de acordo com a base de dados (número de arquivos de áudio) e a arquitetura neural utilizada. Portanto, pode-se afirmar que houve um custo computacional médio de 25 horas.

4.4.1 Resultados do Meta-Learning

Essa seção tem como objetivo apresentar os resultados para o *meta-learning*. Conforme descrito na na Figura 4.5, após ser encontrado as configurações de otimizador, taxa de aprendizado, *data augmentation*, extração de características e arquitetura neural para as bases de dados RAVDESS, TESS, SAVEE e R+T+S que obtiveram maior valor de acurácia, essas configurações são então armazenadas em uma base de conhecimento.

Conforme demonstrado a melhor configuração incluiu o uso do otimizador Adam com taxa de aprendizado variável, aplicação de DA na forma de alongamento temporal, a utilização do método MFCC para extração das características, e a adoção de uma arquitetura neural híbrida composta por uma CNN com 4 blocos seguida por camadas LSTM. Desta forma, é realizado a comparação entre utilizar as configurações contidas na base de conhecimento, realizando a transferência das configurações para as novas bases de dados CREMA-D e R+T+S+C, ou seja, usar ou não o *meta-learning*. A comparação será em termos do valor da acurácia e do tempo computacional gasto.

Na Tabela 4.12, é possível observar a comparação entre as configurações “sem MtL"e aquelas derivadas do processo de *meta-learning* “com MtL", juntamente com os valores de acurácia correspondentes e o tempo computacional. Nesse contexto, podemos observar que as configurações são bastante semelhantes. Para a base de dados CREMA-D, a diferença entre a configuração "sem MtL"para a "com MtL"foi apenas com o otimizador e a taxa de aprendizado,

essa diferença implicou em um aumento no valor da acurácia de 2,16%. Além disso, o uso do *meta-learning* fez muita diferença com relação ao tempo computacional gasto, conforme pode ser observado, de 218:51 minutos para 15:19 minutos.

Tabela 4.12: Comparação dos valores de acurácia (%) e tempo computacional (min) para os bancos de dados CREMA-D e R+T+S+C usando as configurações "sem meta-learning" e "com meta-learning". Valores em negrito indicam o melhor resultado em acurácia e tempo computacional.

Base	MtL	Otimiz.	L.R.	D.A.	Características	Arquit.	Acc. (%)	Tempo (min)
CREMAD	Sem	Adagrad	0.01	Stretch	MFCC	CNN4b+ LSTM	81,12	218:51
	Com	Adam	variável	Stretch	MFCC	CNN4b+ LSTM	**83,28**	**15:19**
R+T+S+C	Sem	Adagrad	0.01	Stretch	MFCC	CNN6b	86,87	388:32
	Com	Adam	variável	Stretch	MFCC	CNN4b+ LSTM	**90,94**	**25:50**

Para o conjunto de dados R+T+S+C, a diferença entre as configurações "sem MtL" e "com MtL" envolveu mudanças no otimizador, na taxa de aprendizado e na arquitetura neural. Essa diferença aumentou a precisão em 4,07%. Além disso, o *meta-learning* reduziu significativamente o tempo computacional necessário, como observado, de 388:32 minutos para 25:50 minutos.

Na Figura 4.10, pode-se examinar o histórico de treinamento e validação da acurácia e perda para o conjunto de dados CREMA-D. Como pode ser visto, ao longo de 100 épocas, o modelo conseguiu generalizar as emoções do conjunto de dados. E na Tabela 4.13, pode-se verificar os dados de precisão, recall e F-score obtidos por meio do processo de *meta-learning*.

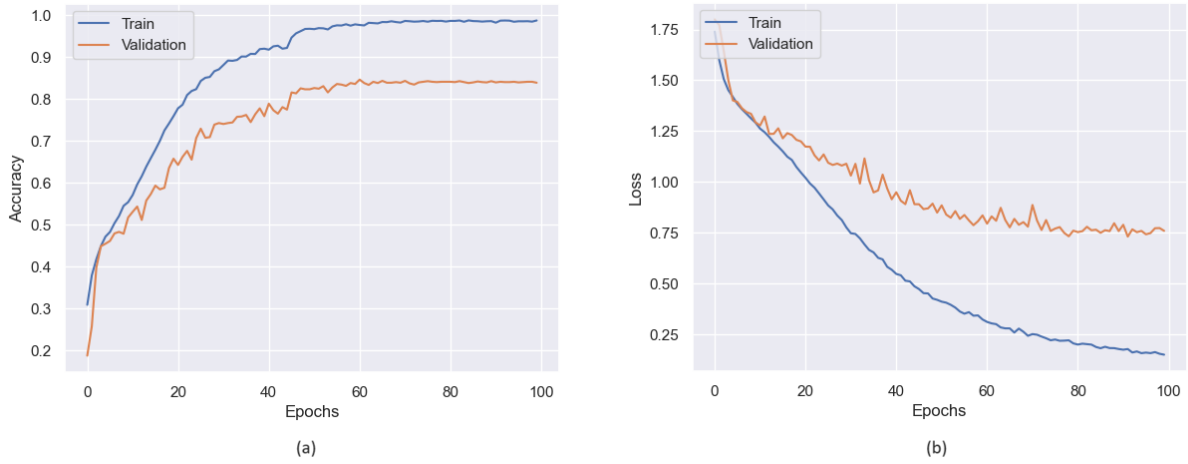


Figura 4.10: Gráfico da (**a**) acurácia e (**b**) valores de perda durante o treinamento e validação do conjunto de dados CREMA-D.

Tabela 4.13: Valores em porcentagem da Precisão, Recall e F-score para a base CREMA-D.

Emoção	Precisão (%)	Recall (%)	F-Score (%)
Desgosto	81,0	84,0	82,0
Feliz	80,0	83,0	82,0
Medo	85,0	80,0	82,0
Neutro	76,0	80,0	78,0
Raiva	91,0	87,0	89,0
Triste	86,0	85,0	85,0
Acurácia (%)	-	-	83,3

Na Figura 4.11, pode-se visualizar o histórico de treinamento e validação da acurácia e perda para o conjunto de dados R+T+S+C. Na Tabela 4.14, pode-se verificar os dados de precisão, recall e F-score obtidos para a base R+T+S+C por meio do uso do MtL.

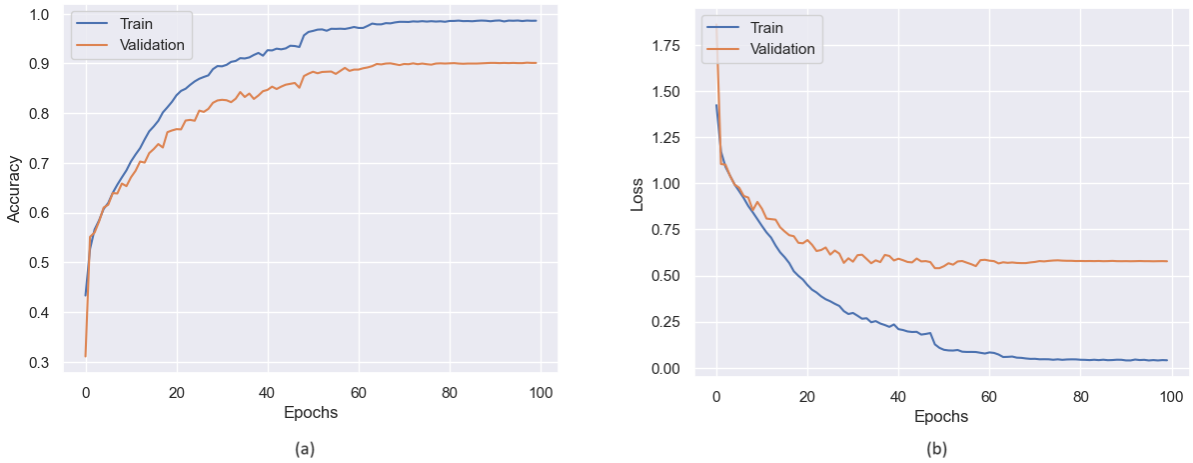


Figura 4.11: Gráfico da (**a**) acurácia e (**b**) valores de perda durante o treinamento e validação do conjunto de dados R+T+S+C.

Tabela 4.14: Valores em porcentagem da Precisão, Recall e F-score para a base R+T+S+C.

Emoção	Precisão (%)	Recall (%)	F-Score (%)
Desgosto	95,0	93,0	94,0
Feliz	95,0	99,0	97,0
Medo	100,0	93,0	96,0
Neutro	100,0	100,0	100,0
Raiva	94,0	99,0	96,0
Surpresa	97,0	97,0	97,0
Triste	98,0	98,0	98,0
Acurácia (%)	-	-	97,0

4.4.2 Comparação dos Resultados

Esta seção tem como objetivo realizar uma análise comparativa dos resultados obtidos em relação a pesquisas anteriores documentadas na literatura. A Tabela 4.15 fornece uma visão abrangente dos estudos em análise, dos conjuntos de dados utilizados e dos valores de acurácia correspondentes relatados em cada artigo.

Tabela 4.15: Comparação dos valores de acurácia (%) da abordagem proposta com trabalhos relacionados. Valores em negrito indicam o melhor resultado de acurácia.

Artigo	RAVDESS	TESS	SAVEE	CREMA-D	R+T+S	R+T+S+C
Pan e Wu (2023)	95,52	-	-	-	-	-
Ahmed et al. (2023)	95,22	99,46	**93,22**	**90,47**	-	-
Asiya e Kiran (2021)	68,00	-	-	-	89,00	-
Bautista et al. (2023)	89,33	-	-	-	-	-
Bhangale e Kothandaraman (2023)	94,18	-	-	-	-	-
Chatterjee et al. (2021)	90,48	95,79	-	-	-	-
Chitre et al. (2022)	77,31	-	-	-	89,93	72,94
Dolka et al. (2021)	88,76	99,52	86,80	71,69	-	-
Guizzo et al. (2023)	55,15	99,76	-	-	-	-
Gupta et al. (2022)	-	-	-	-	-	92,73
Hazra et al. (2022)	-	-	-	-	72,54	-
Jothimani e Premalatha (2022)	92,60	99,60	84,90	89,90	-	**94,50**
Kumar et al. (2021)	49,88	-	-	-	77,23	-
Mittal et al. (2022)	-	-	-	-	-	70,00
Nasim et al. (2021)	54,40	99,60	-	-	-	-
Shanthi et al. (2022)	-	-	-	-	-	82,42
Singh et al. (2023)	74,44	99,81	57,50	-	90,19	-
Zielonka et al. (2022)	-	-	-	-	-	57,42
Proposto	**97,01**	**100,00**	90,62	83,28	**97,37**	90,94

A maioria dos artigos mencionados na Tabela 4.15 utiliza a base de dados RAVDESS. Ao comparar os valores de acurácia obtidos, é possível observar que o método proposto atinge uma acurácia maior, de 97,01%. Os trabalhos de Pan e Wu (2023); Ahmed et al. (2023) são os que mais se aproximam do nosso resultado, com acurácia de 95,52% e 95,22%, respectivamente. O mesmo acontece quando a base de dados TESS é analisada, na qual o trabalho proposto alcançou 100% de acurácia.

As bases de dados SAVEE e CREMA-D tiveram melhor desempenho no trabalho em Ahmed et al. (2023), que utiliza uma arquitetura híbrida com CNN, LSTM e GRU como classificador. O modelo de Ahmed et al. (2023) alcançou uma precisão de 93,22% para SAVEE e 90,47% para CREMA-D. Nosso trabalho obteve uma precisão semelhante para SAVEE, em 90,62%, enquanto nosso trabalho alcançou 83,28% para o banco de dados CREMA-D.

Em relação às bases de dados que combinaram outros conjuntos de dados, como RAVDESS+TESS+SAVEE (R+T+S) e RAVDESS+TESS+SAVEE+CREMA-D (R+T+S+C), nosso

trabalho alcançou uma acurácia de 97,37% para o conjunto de dados R+T+S, destacando-se em relação aos trabalhos citados. Na base de dados R+T+S+C, o artigo de Jothimani e Premalatha (2022) obteve uma acurácia de 94,50% usando uma rede híbrida composta por uma CNN e LSTM.

4.5 Testes do Módulo de Reconhecimento de Emoção da Fala

Esta Seção tem como objetivo realizar testes do Módulo de Reconhecimento de Emoção da Fala. Esses testes serão, posteriormente, comparados com o Módulo de Reconhecimento de Emoção da Expressão Facial e a Fusão das Emoções. Neste sentido, para realizar as devidas comparações, foi selecionado uma base de dados áudio-visual, i. e., contem informações de áudio e vídeo.

Para realizar os testes, foram selecionados os modelos otimizados para cada base de dados. Neste sentido, o modelo obtido pela base de dados RAVDESS será denominado de Modelo 01, o da base de dados TESS será o Modelo 02, o da base SAVEE será o Modelo 03, a base CREMA-D será o Modelo 04, a base R+T+S será o Modelo 05 e a R+T+S+C será o Modelo 06.

Os Modelos de 01 a 06 foram testados utilizando uma nova base de dados. Para o teste a base de dados selecionada foi a MELD (*Multimodal Emotion Lines Dataset*) (Poria et al., 2018). O conjunto de dados Multimodal MELD abrange as modalidades de áudio, vídeo e texto. A MELD possui mais de 1400 vídeos da série de TV *Friends* em que vários atores participaram das cenas. As emoções foram rotulada dentro das sete emoções básicas: raiva, desgosto, triste, feliz, neutro, surpresa e medo.

Conforme mostrado na Tabela 4.16, o Modelo 01 foi o que conseguiu melhor generalizar os dados da nova base de dados MELD. Sendo portanto, o modelo a ser utilizado no Módulo de Reconhecimento de Emoção da Fala.

Tabela 4.16: Valores da acurácia (%) do teste realizado com a base MELD.

Modelo 01	Modelo 02	Modelo 03	Modelo 04	Modelo 05	Modelo 06
73.0	56.0	62.0	22.0	58.0	23.0

CAPÍTULO 5

Módulo de Reconhecimento de Emoção da Expressão Facial

5.1 Introdução

As expressões faciais desempenham um papel crucial na comunicação, servindo como meio não verbal de demonstrar emoções. A capacidade de reconhecer essas emoções de forma precisa desempenha um papel importante na manutenção de conexões interpessoais e na colaboração em sociedade. O rosto humano é composto por inúmeras dinâmicas de estados emocionais, incluindo desde felicidade e tristeza até raiva, medo, surpresa e desgosto (Dewi et al., 2023).

Essas expressões faciais são o resultado de complexas interações entre os músculos faciais, resultando em variações perceptíveis na forma e na aparência geral do rosto. O Reconhecimento de Emoções Faciais (*Facial Emotion Recognition* - FER), é uma área de estudo que reúne especialistas em visão computacional, computação afetiva, interação humano-computador e comportamento humano. Seu foco está na previsão de emoções por meio da análise de expressões faciais em imagens ou vídeos (Pham et al., 2023).

Nos últimos anos, avanços significativos têm sido feitos no desenvolvimento de técnicas para o FER. Com o surgimento das técnicas de aprendizado profundo, especialmente as redes neurais convolucionais, o campo experimentou uma transformação significativa. Os modelos de aprendizado profundo demonstraram uma capacidade excepcional de aprender automaticamente características diretamente das imagens faciais, resultando em uma notável melhoria na precisão das tarefas de reconhecimento de emoções (Wu, 2023).

Nos modelos de aprendizado profundo, a importância de ajustar os hiperparâmetros, como

otimizador e taxa de aprendizado, não pode ser subestimada. Esses hiperparâmetros desempenham um papel relevante na eficiência do treinamento do modelo, afetando diretamente sua capacidade de convergir para uma solução ótima. Além disso, o uso de técnicas de *data augmentation* (DA) é fundamental para enriquecer o conjunto de dados disponível para o modelo. Ao gerar variações sintéticas dos dados, o DA não apenas aumenta a quantidade de dados disponíveis para treinamento, mas também auxilia na criação de um modelo mais robusto e em uma maior capacidade de generalização. Por fim, a investigação da arquitetura neural é crucial para encontrar a estrutura mais adequada para o problema em questão.

Nesse sentido, este Capítulo tem como objetivo elaborar uma metodologia que busca otimizar o Módulo de Reconhecimento de Emoções da Expressão Facial (MREEF). Neste sentido, serão investigadas as melhores combinações entre otimizador, taxa de aprendizado, *data augmentation* e arquitetura neural. Ao final, o modelo otimizado será aplicado ao MREEF para ser testados em dados de vídeos.

Este capítulo está estruturado da seguinte forma: A seção 5.2 apresenta uma revisão dos trabalhos na área FER que realizam investigações para encontrar as melhores configurações. A seção 5.3 define a abordagem completa para obtenção dos resultados para base de dados FER2013. Os resultados são apresentados e discutidos na seção 5.4. E por fim, na seção 5.5 o MREEF é testado utilizando a base de dados MELD.

5.2 Trabalhos Relacionados

O domínio da pesquisa em reconhecimento de emoções faciais testemunhou numerosas inovações e avanços significativos ao longo do tempo. Em particular, as técnicas avançadas de aprendizado profundo têm impulsionado progressos substanciais nesse campo. A aplicação desses algoritmos tem suscitado a demanda por investigar e otimizar configurações visando o melhor desempenho do modelo.

Esta seção fornece uma análise dos estudos no âmbito do reconhecimento de emoções faciais. Neste contexto, foi examinado uma variedade de trabalhos recentes que se concentram em explorar a melhor combinação das configurações do FER. Esses artigos realizam comparações de pelo menos um parâmetro, incluindo o otimizador, a taxa de aprendizado, *data augmentation* e a arquitetura neural. Na Tabela 5.1, são apresentados os artigos e as configurações

investigadas em seus estudos.

Tabela 5.1: Trabalho relacionado na área FER que investiga as melhores combinações de otimizadores, taxas de aprendizado e arquiteturas neurais. Os artigos marcados com um check indicam que realizaram uma comparação com a respectiva configuração.

Artigo	Otimizador	Taxa de Aprendizado	Data Augmentation	Arquitetura Neural
Lawpanom et al. (2024)	✓	✓	-	✓
Sahoo et al. (2023)	-	-	-	✓
Pham et al. (2023)	-	-	-	✓
Helaly et al. (2023)	-	-	-	✓
Kim e Lee (2023a)	-	-	-	✓
Li e Xu (2020)	-	-	-	✓
Song (2021)	-	-	-	✓
Wu (2023)	-	-	-	✓
Nguyen et al. (2019)	-	-	-	✓
Bodapati et al. (2022)	-	-	✓	✓
Proposta	✓	✓	✓	✓

No artigo de Lawpanom et al. (2024), uma comparação entre otimizadores é realizada, abrangendo tanto o Adam quanto o SGD. Além disso, foram analisadas as diferentes taxas de aprendizado, variando de 0,01, 0,001 a 0,005. Adicionalmente, houve uma avaliação comparativa das arquiteturas neurais da CNN. Os autores empregaram técnicas de *data augmentation*, como espelhamento horizontal, zoom, rotação e deslocamento em largura e altura. Contudo, é importante destacar que o efeito do DA não foi objeto de comparação.

No estudo conduzido por Sahoo et al. (2023) e Pham et al. (2023), são comparadas exclusivamente arquiteturas de redes neurais convolucionais. Em Sahoo et al. (2023), o otimizador empregado foi o SGDM, utilizando uma taxa de aprendizado variável, juntamente com técnicas DA, como deslocamento em largura e altura, além de espelhamento horizontal e rotação. Enquanto isso, no estudo de Pham et al. (2023), foi adotado o SGD com taxa de aprendizado variável, combinado com o DA de espelhamento e rotação.

Já no artigo de Helaly et al. (2023), o otimizador Adam foi empregado com uma taxa de aprendizado variável. Quanto ao DA, embora os autores mencionem várias técnicas, elas não são especificadas. O estudo se concentra exclusivamente na comparação de diferentes arquiteturas neurais.

Nos estudos conduzidos por Kim e Lee (2023a) e Li e Xu (2020), ambos abordam a comparação de arquiteturas neurais. Entretanto, nenhum dos dois artigos faz uso da técnica de *data augmentation*. No trabalho de Kim e Lee (2023a), os autores optaram pelo otimizador Adam

com uma taxa de aprendizado fixa de 0,001, enquanto no estudo de Li e Xu (2020), não há descrição do otimizador e da taxa de aprendizado utilizados.

No artigo de Song (2021), foi empregado o otimizador Adam com uma taxa de aprendizado variável. Para o *data augmentation*, foram aplicadas técnicas de espelhamento, rotação e corte aleatório. O estudo também compara arquiteturas que empregam o algoritmo CNN com mecanismo de atenção. Já o artigo de Wu (2023), não apresenta informações sobre o otimizador, taxa de aprendizado e DA, mas também realiza uma comparação entre arquiteturas neurais utilizando CNN com mecanismo de atenção.

No estudo conduzido por Nguyen et al. (2019), foi empregado o otimizador SGD com uma taxa de aprendizado variável. Quanto ao DA, foram empregadas transformações que incluem translação, rotação, espelhamento e zoom. O estudo realizou comparações apenas com relação à arquitetura neural da CNN.

E por fim, artigo realizado por Bodapati et al. (2022) adota o otimizador Adam com uma taxa de aprendizado constante de 0,001. Nesse trabalho, os autores exploram técnicas aprimoradas de DA, incluindo espelhamento horizontal e rotação; a investigação abrange a variação das taxas dessas técnicas em quatro modelos distintos. Além do *data augmentation*, os autores também comparam quatro arquiteturas neurais diferentes.

Com o levantamento realizado é possível observar uma forte tendencia dos trabalhos amostrados em investigar arquiteturas neurais. Sendo que 80% dos trabalhos realizam apenas a análise da arquitetura. Além disso, é possível observar que os trabalhos apontados pela Tabela 5.1 demonstram uma lacuna no que se refere a análise comparativa entre otimizador, taxa de aprendizado e *data augmentation*.

Diante disso, este capítulo tem como motivação atender a lacuna apontada pelos trabalhos citados anteriormente. Desta forma, é uma contribuição desta tese a elaboração de uma metodologia que investiga a combinação entre otimizador, taxa de aprendizado, técnicas de *data augmentation* e arquitetura neural para o reconhecimento de emoção facial.

5.3 Metodologia Proposta

Diante das considerações apresentadas na Seção 5.2 e das lacunas identificadas, foi desenvolvida uma metodologia para aprimorar as combinações do módulo de reconhecimento de

emoções da expressão facial. Nesse contexto, serão avaliados otimizadores, taxas de aprendizado, técnicas de *data augmentation* e arquiteturas neurais.

A sequência das etapas da metodologia proposta é ilustrada na Figura 5.1. Para conduzir esta análise, será utilizada a base de dados FER2013, amplamente empregada na literatura para o reconhecimento de emoções na expressão facial.

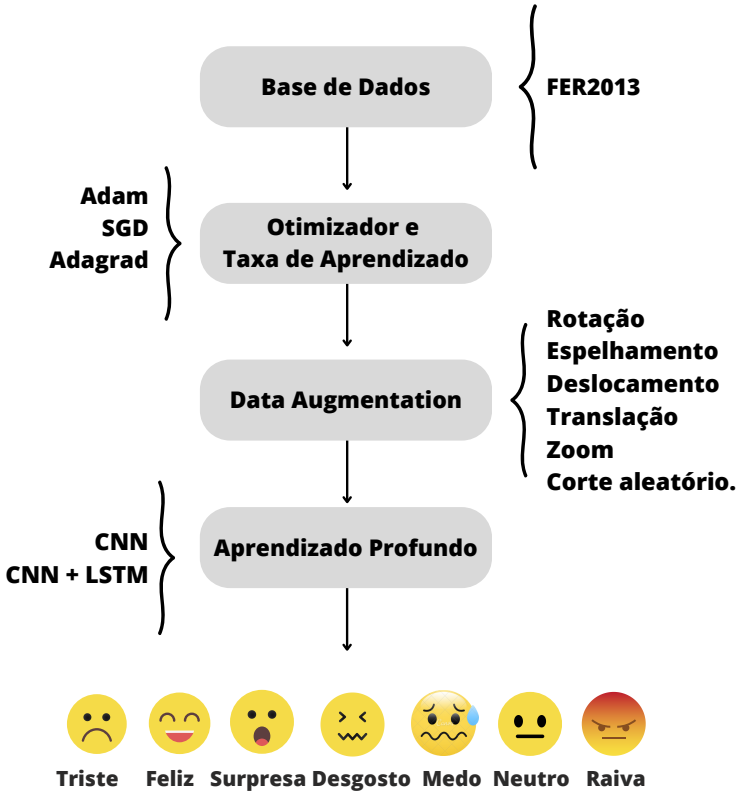


Figura 5.1: Metodologia proposta para otimizar o reconhecimento de emoções da expressão facial

A primeira etapa da metodologia consiste na investigação do otimizador e da taxa de aprendizado. Na segunda etapa, será realizada uma análise sobre a eficácia das técnicas de aumento de dados (*data augmentation*). Por fim, serão avaliados dois algoritmos de aprendizado profundo: uma Rede Neural Convolucional (CNN) e um algoritmo híbrido que combina CNN e LSTM. A seguir, serão fornecidas informações mais detalhadas sobre cada uma dessas etapas.

5.3.1 Bases de Dados

A base de dados FER2013 é um conjunto de dados de código aberto carregado publicamente para uma competição *Kaggle*. Este conjunto de dados consiste em 35887 fotos de rosto em escala de cinza, 48×48 pixels com as sete emoções diferentes. O conjunto de dados foi definido como acessível a todos após a conclusão da competição. Na Figura 5.2, pode-se visualizar

algumas das imagens contidas no conjunto FER2013 para as sete emoções: feliz, surpresa, desgosto, raiva, triste, neutro e medo. Com o total de 35887, o conjunto de dados FER2013 é desbalanceado, possuindo 4953 imagens de raiva, 547 de desgosto, 5121 de medo, 8989 de feliz, 6077 de neutro, 4002 de triste e 6198 de surpresa.

Figura 5.2: Exemplo das imagens contidas na base de dados FER2013.

5.3.2 Passo 1: Ajuste dos otimizadores e taxa de aprendizado

Dada a importância desses parâmetros para o modelo CNN, a primeira investigação do modelo proposto visa determinar o melhor otimizador e valor de aprendizado para cada conjunto de dados utilizado. Para isso, serão considerados os principais otimizadores na literatura, como Adam, SGD e Adagrad, juntamente com valores de aprendizado comuns em estudos de reco-

nhecimento de emoções através de expressões faciais (0,01, 0,001, e um esquema adaptativo com fator de 0,4 e um mínimo de 0,000001).

5.3.3 Passo 2: Otimização do *data augmentation*

A aplicação de técnicas de *data augmentation* é crucial no campo da visão computacional, especialmente em tarefas de reconhecimento de imagem, onde a disponibilidade de dados e a diversidade é essencial para treinar modelos eficazes (Lawpanom et al., 2024). Na Figura 5.3, é possível observar algumas das técnicas de *data augmentation* utilizadas.

Uma das técnicas mais simples e frequentemente empregadas é a rotação de imagens, Figura 5.3(a) na qual as amostras são giradas em ângulos específicos, permitindo ao modelo aprender a reconhecer objetos em diferentes orientações. Essa variação é essencial para garantir que o modelo seja robusto e capaz de lidar com objetos apresentados de maneiras diferentes nas imagens (Chollet, 2021).

Outras técnicas importantes são o deslocamento em largura e altura (Figura 5.3(c) e (d)) e o espelhamento horizontal ou vertical (Figura5.3(e) e (f)) que reflete a imagem em relação a um eixo, introduzindo variações na posição dos objetos. Isso é fundamental para ensinar o modelo a reconhecer objetos independentemente da sua orientação espacial, o que é especialmente relevante em tarefas como detecção de objetos em que a posição do objeto pode variar consideravelmente (Chollet, 2021).

Além disso, técnicas como cisalhamento (Figura 5.3(g)) e zoom (Figura 5.3(h)) também são amplamente utilizadas para introduzir variações na posição e escala dos objetos nas imagens, ajudando a melhorar a capacidade do modelo de generalizar para diferentes cenários. Técnicas como o ajustes de brilho e contraste, bem como a adição de ruído, são outras técnicas de DA utilizadas com menos frequência quando comparadas com as demais e que tornar o modelo mais robusto a variações na iluminação e a interferências externas (Chollet, 2021).

Com isso, a segunda etapa da metodologia é realizado a análise do uso do *data augmentation* para o reconhecimento de emoções pela expressão facial. Neste sentido, foram testadas separadamente e em conjunto, como mostrado na Figura 5.3(i), as operações de rotação, deslocamento em largura e altura, espelhamento vertical e horizontal, cisalhamento e zoom.

(a) Original (b) Rotação (c) Deslocamento Largura

(d) Deslocamento Altura (e) Espelhamento Horizontal (f) Espelhamento Vertical

(g) Cisalhamento (h) Zoom (i) Todas as operações juntas

Figura 5.3: Exemplo de operações de *data augementation*. Fonte da imagem Original: banco de dados WSEFEP (Olszanowski et al., 2015).

5.3.4 Passo 3: Investigação da Arquitetura Neural

O passo 3 da metodologia proposta consiste em investigar a melhor arquitetura neural, para isso, foram utilizados dois algoritmos: a Rede Neural Convolucional (CNN) e a Memória Longa e de Curto Prazo (LSTM). A seguir serão apresentadas mais informações sobre os dois algoritmos utilizados.

Redes Neurais Convolucionais

Neste estudo, utilizou-se a CNN para realizar a classificação multiclasse das emoções da expressão facial. A arquitetura utilizada, mostrada na Figura 5.4, é baseada na arquitetura apresentada no livro de Chollet (2021).

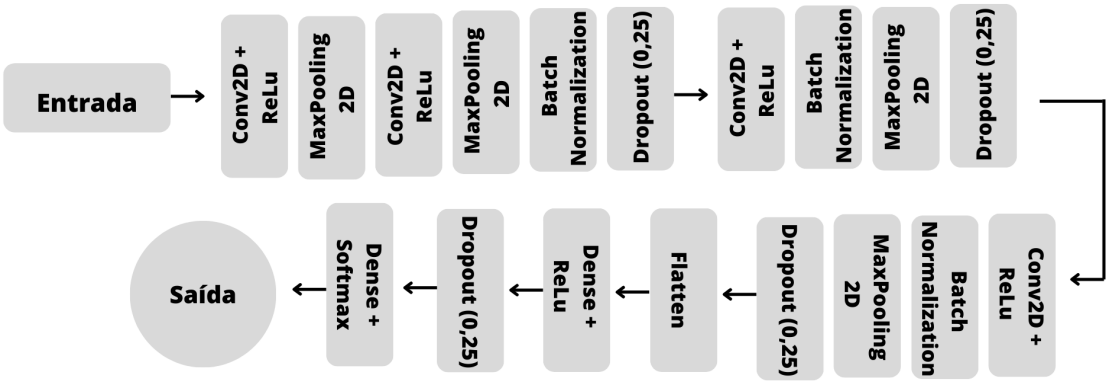


Figura 5.4: Arquitetura da rede neural convolucional utilizada para o reconhecimento de emoções pela expressão facial FER2013.

Na Figura 5.4 é possível observar as camadas da CNN. Foram utilizados a camada de convolução 2D (Conv2D) com ativação a função ReLu, e um *kernel size* de 3x3. Também foram utilizados a operação *MaxPooling2D* (tamanho=5), *BatchNormalization* e *Dropout*(taxa=0,25). Por fim, utilizou-se duas camadas densas uma com ativação ReLu e a última com ativação Softmax. Observa-se que as camadas que possuem número de parâmetros iguais a 0 são camadas de operação, como é o caso da *MaxPooling*, *Dropout* e *Flatten*. Foram executados 100 épocas com um *batch size* de 64. Para função perda, foi utilizado a “entropia cruzada categórica".

Memória Longa e de Curto Prazo

Memória Longa e de Curto Prazo (*Long Short-Term Memory* - LSTM) é uma arquitetura de Rede Neural Recorrente (RNN). Foi proposta uma arquitetura híbrida, unindo a CNN com LSTM. Desta forma, foram acrescentados duas camadas de LSTM a arquitetura da Figura 5.4, a primeira camada e a segunda com 128 unidades. A Figura 5.5 mostra a arquitetura híbrida CNN+LSTM.

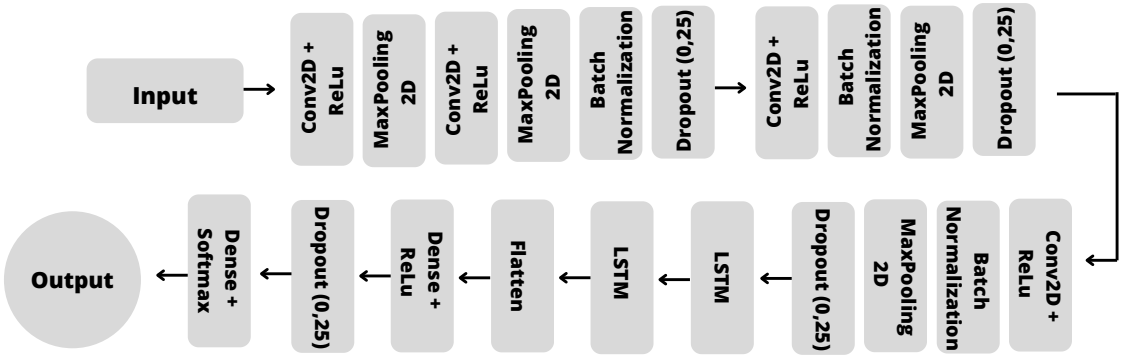


Figura 5.5: Arquitetura híbrida da CNN + LSTM utilizada para o reconhecimento de emoções pela expressão facial FER2013.

5.4 Resultados

Neste capítulo serão apresentados os resultados obtidos seguindo a metodologia descrita na Seção 5.3. Para realizar os experimentos utiliza-se um notebook com sistema operacional Windows 11, processador Intel i5 1135G7 2.40GHz, memória RAM DE 8 GB e uma GPU Nvidia GeForce MX350 com 2GB VRAM. O ambiente de desenvolvimento utilizado neste trabalho é composto pela IDE Jupyter, associada a linguagem Python.

Conforme descrito na Seção 5.3 a primeira etapa da investigação metodológica consiste em encontrar o melhor conjunto de otimizador e taxa de aprendizado para o modelo da rede neural convolucional que classifica as emoções da FER2013. Neste sentido, foram testados os otimizadores Adam, SGD e Adagrad juntamente com as taxas de aprendizado 0,01, 0,001 e taxa variável. Os resultados da etapa 1 estão descrito na Tabela 5.2.

Tabela 5.2: Valores em porcentagem da acurácia para os otimizadores Adam, SGD e Adagrad com a taxa de aprendizado de 0,01, 0,001 e taxa variável.

Taxa de Aprendizado	Adam	SGD	Adagrad
0,01	60,98	52,35	63,22
0,001	63,36	46,01	52,44
variável	**64,89**	53,43	64,35

Como pode ser visto na Tabela 5.2, o otimizador Adam em conjunto com a taxa de aprendizado variável obtiveram o melhor valor de acurácia. Desta forma, nas próximas etapas da metodologia serão utilizados o conjunto Adam variável.

A segunda etapa da metodologia, consiste em otimizar o uso do *data augmentantion*. Na Tabela 5.3 é possível observar a influência da escolha da técnica do DA aplicado ao *deep learning*.

Tabela 5.3: Valores em porcentagem da acurácia para o uso de data augmentation.

Data Augmentation	Acurácia (%)
Sem DA	64,89
Rotação	53,96
Espelhamento Vertical	67,21
Espelhamento Horizontal	**75,05**
Deslocamento Vertical	65,78
Deslocamento Horizontal	62,42
Translação	70,84
Zoom	62,35
Corte	68,08
Todas	70,09

E por fim, a última etapa da metodologia proposta é a investigação da arquitetura neural. Nesse trabalho utilizou-se duas arquitetura, conforme mostrado na Seção 5.3, uma arquitetura CNN e uma híbrida com CNN e LSTM. Na Tabela 5.4 é possível verificar os resultados de acurácia obtidos em cada a arquitetura neural.

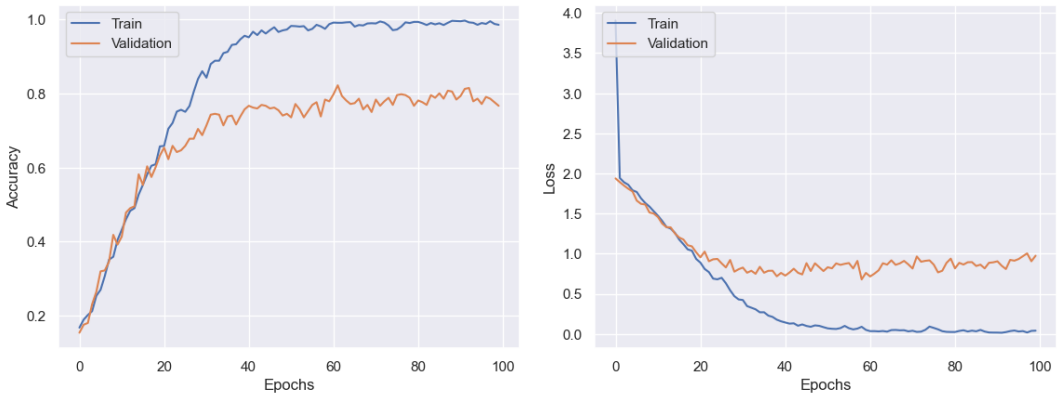
Tabela 5.4: Valores em porcentagem da acurácia para investigação da arquitetura neural.

Arquitetura	Acurácia (%)
CNN	**75,05**
CNN + LSTM	68,47

Desta forma, segundo a metodologia proposta, a melhor combinação para o reconhecimento de emoções pela expressão facial utilizando a base de dados FER2013 é com o otimizador Adam, taxa de aprendizado variável, usando a técnica de *data augmentation* de espelhamento horizontal com a arquitetura neural CNN. Essa combinação gerou um histórico de acurácia e perda mostrado na Figura 5.6 e valores de precisão, recall e f-score mostrados na Tabela 5.5.

Tabela 5.5: Valores em porcentagem da Precisão, Recall e F-score para a base FER2013.

Emoção	Precisão (%)	Recall (%)	F-Score (%)
Desgosto	81,00	65,0	72,0
Feliz	81,0	82,0	82,0
Medo	75,0	63,0	69,0
Neutro	68,0	88,0	77,0
Raiva	88,0	85,0	87,0
Surpresa	62,0	74,0	68,0
Triste	75,0	81,0	77,0
Acurácia (%)	-	-	75,0



(a) Acurácia (b) Perda

Figura 5.6: Histórico da (a) acurácia e (b) perda da base de dados FER2013.

5.4.1 Comparação dos Resultados

Esta seção se destina a realizar uma análise comparativa dos resultados obtidos em relação aos estudos apresentados na Tabela 5.1. Com esse propósito, a Tabela 5.6 é disponibilizada, oferecendo uma visão das métricas de acurácia encontradas em cada um dos artigos selecionados.

Tabela 5.6: Comparação entre os valores de acurácia para o reconhecimento de emoção pela expressão facial que utilizam a base de dados FER2013.

Trabalhos	Acurácia (%)
Proposto	75,05
Lawpanom et al. (2024)	75,15
Sahoo et al. (2023)	66,60
Pham et al. (2023)	61,05
Helaly et al. (2023)	83,00
Kim e Lee (2023a)	73,31
Li e Xu (2020)	72,35
Song (2021)	74,00
Wu (2023)	71,38
Nguyen et al. (2019)	74,09
Bodapati et al. (2022)	69,57

Ao analisar os dados apresentados na Tabela 5.6, é possível observar que os artigos obtém uma acurácia que varia aproximadamente de 61% a 83%. Sendo amplamente reconhecido na literatura que essa base de dados (FER2013) é desafiadora devido à sua complexidade. Portanto, ao comparar o desempenho da metodologia proposta com os resultados obtidos na literatura, constata-se que a acurácia alcançada, de 75,05%, está dentro da faixa observada na pesquisa, reforçando a consistência e relevância dos resultados obtidos.

5.5 Testes do Módulo de Reconhecimento de Emoção da Expressão Facial

Esta Seção tem como objetivo realizar testes com o modelo do classificador já otimizado, utilizando a base de dados multimodal MELD (*Multimodal Emotion Lines Dataset*) (Poria et al., 2018). O conjunto de dados Multimodal MELD abrange as modalidades de áudio, vídeo e texto. A MELD possui mais de 1400 vídeos da série de TV *Friends* em que vários atores participaram das cenas. As emoções são rotulada dentro das sete emoções básicas: raiva, desgosto, triste, feliz, neutro, surpresa e medo.

Para conduzir os testes, foi desenvolvido um algoritmo que segue o fluxograma descrito na Figura 5.7. Inicialmente, é selecionado o vídeo da base de dados, com duração aproximada de 3 a 5 minutos. A partir desse vídeo, é extraído um frame por segundo, e estas imagens são então encaminhadas para a próxima etapa do processo, que consiste em localizar o rosto da pessoa na imagem. Para essa finalidade, foi empregada a biblioteca *OpenCV*, possibilitando a detecção facial. Após a etapa de detecção, a imagem facial é submetida à predição por meio de uma rede neural convolucional. Por fim, são apresentadas as probabilidades associadas às emoções de alegria, medo, raiva, neutralidade, surpresa, desgosto e tristeza.

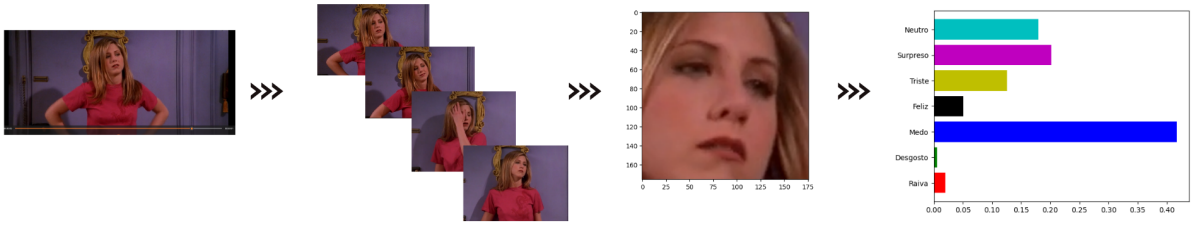


Figura 5.7: Fluxograma dos testes do MREEF em vídeos da base de dados MELD.

Após realizar o teste nos vídeos da base de dados MELD, observou-se uma acurácia de 78,06% no reconhecimento de emoções da expressão facial. Isso significa que o modelo treinado e testado pela base de dados FER2013, conseguiu uma boa generalização dos dados emocionais em faces humanas. Nesse aspecto, ao testá-lo em uma base nova, como a MELD, o modelo teve uma boa eficiência na classificação.

CAPÍTULO 6

Fusão das Emoções

6.1 Introdução

A fusão das emoções em sistemas multimodais tem sido um desafio abordado na literatura. O principal obstáculo reside na capacidade de fundir as emoções identificadas pelos variados módulos, interpretando eficientemente o estado emocional resultante do ser humano que está interagindo com o robô, sem comprometer a integridade das informações.

Os métodos de fusão na classificação multimodal de emoções podem ser divididos de duas formas: fusão precoce e fusão tardia (Zhang et al., 2019). A fusão precoce consiste em unir as características multimodais antes do processamento final das informações. Isto é, os métodos que realizam a fusão precoce realizam a união dos dados de entrada e, em seguida, processam o resultado da fusão como um único dado (Zadeh et al., 2017). Segundo Heredia et al. (2022), esses métodos estão em desuso, pois não permitem explorar corretamente as informações de entrada dos dados. Já os métodos de fusão tardia, realizam o processamento dos dados de entrada separadamente, e em seguida, seus dados são mesclados (Zhang et al., 2019). Desta forma, a abordagem de fusão tardia é considerada mais eficiente.

Após a elaboração do Módulo de Reconhecimento de Emoção da Fala (MREF) e Módulo de Reconhecimento de Emoção da Expressão Facial (MREEF) é proposto a fusão multimodal das emoções. O sistema multimodal de emoções terá como objetivo auxiliar na tomada de decisão, pois acredita-se que quanto mais informações o sistema robótico tiver melhor será a escolha da ação a ser realizada. Um exemplo seria, se um dos módulos de reconhecimento falhar ou classificar de forma errada as emoções do usuário o outro módulo poderá indicar a emoção

correta. Fazendo com que o robô seja mais assertivo em suas escolhas.

Neste contexto, neste capítulo, serão detalhados os procedimentos adotados para a realização da fusão tardia dos módulos de reconhecimento de emoção (MREF e MREEF). Para essa finalidade, foi analisado o sistema de inferência Fuzzy. As abordagens foram testadas utilizando os dados audiovisuais da base de dados MELD.

O Capítulo está dividido da seguinte forma: na Seção 6.2 é realizado uma revisão dos trabalhos relacionados na área de fusão tardia do reconhecimento das emoções. Já na Seção 6.3, é descrito a metodologia para a fusão tardia utilizando Sistema de Inferência *Fuzzy*. E por fim, a seção 6.4 mostra os resultados obtidos.

6.2 Trabalhos Relacionados

Para avaliar os trabalhos que realizam a fusão multimodal para o reconhecimento de emoções na literatura, realizou-se uma revisão de literatura. O levantamento dos artigos foi através da base de dados *SCOPUS*, como critério de exclusão foram abordados apenas os trabalhos relacionados com esta proposta, ou seja, a fusão tardia multimodal de emoções na interação humano-robô. Com isso, a busca retornou 10 artigos. Na Tabela 6.1, é possível observar o artigo, o nome do método de fusão, a categoria dos módulos e as emoções utilizadas em cada trabalho.

O artigo de Heredia et al. (2022) realiza o reconhecimento de emoções pelos módulos de fala, texto e expressão facial, possuindo quatro classes de emoções: alegria, neutro, raiva e tristeza. Os autores utilizam de redes neurais convolucionais para classificar as modalidades e realiza a fusão das emoções pelo método descrito no trabalho, o EmbraceNet+ que utiliza duas MLP.

O trabalho de Tzirakis et al. (2021) utiliza módulos de reconhecimento de emoções de fala, texto e expressão facial. As emoções avaliadas são a animação e a valência. Os autores utilizam de uma CNN para classificar de forma binária das emoções. A fusão ocorre unindo os módulos com um modelo que possui uma camada de atenção e uma Rede Neural LSTM (*Long Short Term Memory*) de uma camada. A camada de atenção mescla as modalidades concatenando-as e usando uma abordagem residual para gerar a fusão.

O trabalho de Mittal et al. (2020) aborda o reconhecimento de emoções da expressão facial,

Tabela 6.1: Trabalhos relacionados da literatura que realizam fusão multimodal de emoções na Interação Humano-Robô.

Artigo	**Método**	**Módulos**	**Emoções**
Heredia et al. (2022)	EmbraceNet+	Texto, fala e Expressão Facial	Feliz, neutro, raiva e tristeza
Tzirakis et al. (2021)	Rede LSTM	Texto, fala e Expressão Facial	Animação e valência
Mittal et al. (2020)	MLP	Texto, fala e Expressão Facial	Feliz, neutro, raiva e tristeza
Lan et al. (2020)	DGCCA	EEG e Expressão Facial	Feliz, triste, medo, desgosto e neutro
Ortega et al. (2019)	MLP	Texto, fala e Expressão Facial	Animação, valência e gostar
Akhtar et al. (2019)	biGRU	Texto, fala e Expressão Facial	Feliz, surpresa, desgosto, medo, raiva e tristeza
Tripathi et al. (2018)	MLP	Texto, Expressão Corporal e Expressão Facial	Feliz, neutro, raiva e tristeza
Poria et al. (2016)	MKL	Texto, fala e Expressão Facial	Feliz, neutro, raiva e tristeza
Kahou et al. (2016)	SVM e MLP	Fala e Expressão Facial	Feliz, surpresa, neutro, desgosto, medo, raiva e tristeza
Sun et al. (2016)	Algoritmo de voto majoritário e uma rede de fusão compartilhada por peso	Fala e Expressão Facial	Raiva, desgosto, preocupado, triste, feliz, ansioso, surpresa e neutro
Proposto	Sistema Fuzzy	Fala e Expressão Facial	Feliz, surpresa, neutro, desgosto, medo, raiva e tristeza

fala e texto. As emoções trabalhadas pelos autores são a raiva, feliz, triste e neutro; realizando a classificação binária por classe. Os autores utilizam da rede neural LSTM para cada uma das modalidades e realiza a fusão das emoções utilizando MLP com fusão multiplicativa.

Já em Lan et al. (2020), é utilizado para reconhecimento de emoções as expressões faciais e sinal de eletroencefalograma (EEG). Classificando 5 emoções: alegria, tristeza, medo, desgosto e neutro. O método denominado DGCCA, utiliza Gradiente Descendente Estocástico e adotam retro-propagação para atualizar as matrizes de pesos.

Em Ortega et al. (2019) é realizado o reconhecimento das emoções da fala, expressão facial e texto. As classes binárias consideradas são: animação, valência e o gostar. Os módulos de reconhecimento de emoção utilizam MLP e a fusão acontece na última camada de cada módulo, na qual, são conectadas utilizando uma MLP com uma camada totalmente conectada seguida

por uma camada de ativação linear que faz a soma ponderada da camada anterior.

Em Akhtar et al. (2019), os autores utilizam módulos de texto, fala e expressões faciais para classificar as emoções alegria, surpresa, desgosto, medo, raiva e tristeza. Na implementação dos módulos é utilizado rede BiGRU (*Bidirectional Gated Recurrent Unit Neural Network*). Para realizar a fusão das emoções é utilizado rede biGRU com as saídas de um mecanismo de atenção par a par concatenado.

No trabalho dos autores Tripathi et al. (2018) são considerados os módulos de reconhecimento de emoção pela fala, expressão facial, expressão corporal e texto. É realizado a classificação binária das classes feliz, triste, raiva e neutro. O módulo de fala utiliza da MLP, o módulo de texto classifica utilizando LSTM, e os módulos de expressão corporal e facial utilizam uma combinação do LSTM e CNN. A fusão das emoções é realizada nas camadas finais da rede, adicionando uma RNP com uma camada totalmente conectada de 256 neurônios seguida por uma camada de ativação *softmax*.

Já em Poria et al. (2016) os dados do reconhecimento de emoções pela fala, texto e expressão facial são processado separadamente por uma rede neural convolucional e cada um dá uma pontuação, que são concatenadas e inseridas em um MKL (*Multiple Kernel Learning*). As emoções utilizadas são: alegria, neutro, raiva e tristeza.

Em Kahou et al. (2016) são utilizados módulos de reconhecimento de emoções pela fala, expressão facial. As classes de emoções são feliz, surpresa, neutro, desgosto, medo, raiva e tristeza. Para os classificadores são utilizados redes neurais convolucionais e as redes DBN (*Deep Belief Network*). A fusão de emoções é realizada utilizando a agregação das técnicas SVM (*Support Vector Machine*) e uma MLP

E por fim, em Sun et al. (2016) são implementados os módulos de fala e expressão facial. As classes consideradas são raiva, desgosto, preocupado, triste, feliz, ansioso, surpresa e neutro. A implementação dos módulos é realizada utilizando SVM, MLP, LSTM e CNN. Na fusão das emoções é utilizado um método que utiliza um algoritmo de voto majoritário e uma rede de fusão compartilhada por peso.

A revisão de literatura realizada mostra que os métodos existentes utilizam de técnicas de aprendizado de máquina para realizar a fusão multimodal das emoções, muitos dos trabalhos realizam a fusão das emoções de forma concatenada com os módulos de reconhecimento. Além disso, a maioria utilizam os mesmos métodos, principalmente a MLP. Também, pode-se obser-

var que não são considerados pelos autores dos trabalhos o estado emocional, e sim a emoção predominante das modalidades.

Desta forma, propõe-se a realização a fusão das emoções utilizando sistema *fuzzy*. No qual, acredita-se que a natureza das emoções humanas emoções não devem ser tratadas de forma binária e sim de forma subjetiva, assemelhando a uma distribuição de probabilidades sobre as emoções discretas. Assim, o sistema de fusão de emoções deve ser uma classificação multi-classe. Afinal, em uma imagem ou áudio podem conter mais de uma emoção e não necessariamente 100% de apenas um estado emocional. Com isso, espera-se que o sistema *fuzzy* seja capaz de realizar a fusão das emoções de forma eficiente.

Apesar de nenhum dos trabalhos apontados pela Tabela 6.1 utilizar o sistema de inferência *fuzzy* para realizar a fusão, existem precedentes na literatura que utilizaram para realizar fusão sensorial, também conhecida como fusão de dados (Sasiadek, 2002). Alguns exemplos são na fusão de imagens médicas utilizando *fuzzy* e neuro-*fuzzy* (Yang et al., 2016; Javed et al., 2014; Singh et al., 2004; Myna e Prakash, 2015; Hermessi et al., 2017). Também pode-se citar a fusão de dados para guiar um veículo autônomo, utilizando sensores de velocidade, visão computacional e radar a laser (Subramanian et al., 2009) e outros exemplos.

6.3 Metodologia

Diante do que foi abordado na Seção 6.2, esta Seção tem como objetivo propor um método para realizar a fusão tardia dos módulos apresentados no Capítulo 4 (MREF) e 5 (MREEF). Neste sentido, propõe-se a fusão descrita na Figura 6.1.

A partir de uma gravação áudio-visual, o método proposto capta o áudio e os segmentos de imagens dos vídeos da base de dados MELD. Em seguida, esses dados são pré-processados conforme descritos nos Capítulos 4 (MREF) e 5 (MREEF). Os módulos realizam a classificação da emoção de forma multiclasse por seus algoritmos de aprendizado profundo. Ao final, o resultado é armazenado em um vetor que passa por um filtro de média móvel. Esse filtro auxilia na sincronização das leituras dos módulos, além de filtrar pequenos erros de leitura ou de classificação. E então, é realizado a fusão das emoções multiclasse e multimodal utilizando um Sistema de Inferência *Fuzzy*. A seguir serão apresentadas mais informações sobre a metodologia de seleção das funções de pertinência das entradas e saída.

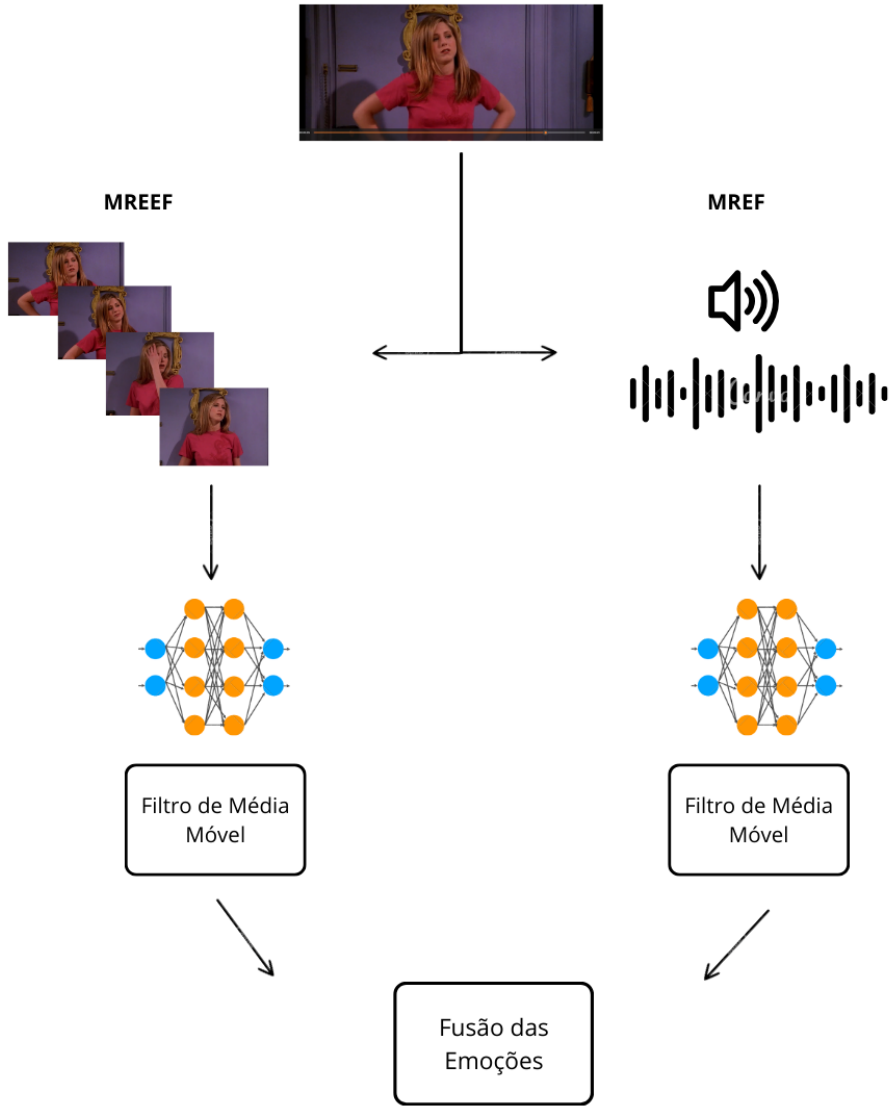


Figura 6.1: Fluxograma dos testes da fusão com vídeos da base de dados MELD.

6.3.1 Sistema de Inferência Fuzzy

O algoritmo a ser investigado será o sistema de inferência *fuzzy* tradicional, do tipo mamdani. Conforme explicado no Capítulo 3, uma característica importante dos sistemas *fuzzy* são as funções de pertinência (FP), tanto as FPs antecedente que são as entradas do sistema *fuzzy*, quanto a FP consequente que é a saída.

Para o sistema *fuzzy* proposto, são utilizadas duas entradas e uma saída. As entradas são os módulos MREF e MREEF, e a saída é a emoção final. Cada uma das sete emoções (feliz, triste, surpreso, medo, raiva, desgosto e neutro) passa pelo sistema *fuzzy*, por exemplo, a emoção feliz de cada módulo será a entrada do SF e a saída será a quantidade em porcentagem de emoção feliz ao final da fusão. Esse processo é realizado por todas as emoções, sendo que ao final

tem-se um vetor com a porcentagem das sete emoções finais.

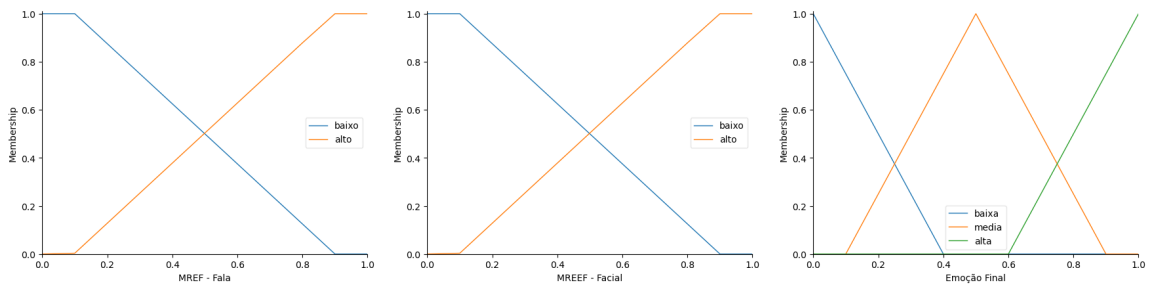
Para realizar a fusão utilizando o SF tradicional, propõe-se uma investigação das funções de pertinência antecedente e consequente. Foram selecionadas três funções de pertinências diferentes para as entradas, modificando a quantidade de variáveis linguísticas (baixo, médio, alto, etc.) e por sua vez, variando os limites de cada FP. Ao final, também foram realizados dois testes variando a FP da saída.

Funções de Pertinência Antecedente

A primeira investigação a ser realizada no sistema *fuzzy* tradicional, do tipo mamdani é variar as funções de pertinência de entrada, isto é, as antecedentes. Conforme explicado anteriormente, o SF possui duas entradas referentes a cada um dos módulos MREF e MREEF.

Em um sistema *fuzzy*, as variáveis deixam de ser numéricas e passam a ser linguísticas. Nesse sentido, foi proposto três variações de entrada, a primeira consiste em duas variáveis: baixo e alto (Figura 6.2). A segunda possui três variáveis: baixo, médio e alto (Figura 6.3). E por fim, a terceira contem cinco variáveis: muito baixo, baixo, médio, alto e muito alto (Figura 6.4). Todas as FP de entrada são do tipo trapezoidal e ambas utilizam a mesma saída padrão para a realização dos testes, que consistem em uma FP consequente do tipo triangular com três variáveis: baixo, médio e alto.

No primeiro teste, relativo a FP com duas variáveis: alto e baixa, demostradas na Figura 6.2, são utilizadas regras *fuzzy* do tipo mostradas abaixo. Em que para as emoções dos módulos iguais (alto ou baixa), a emoção final é considerada igual a dos módulos. Já para a situação em que os módulos apresentam divergência, isto é, quando um está alto e o outro está baixo, então a emoção final é considerada média.

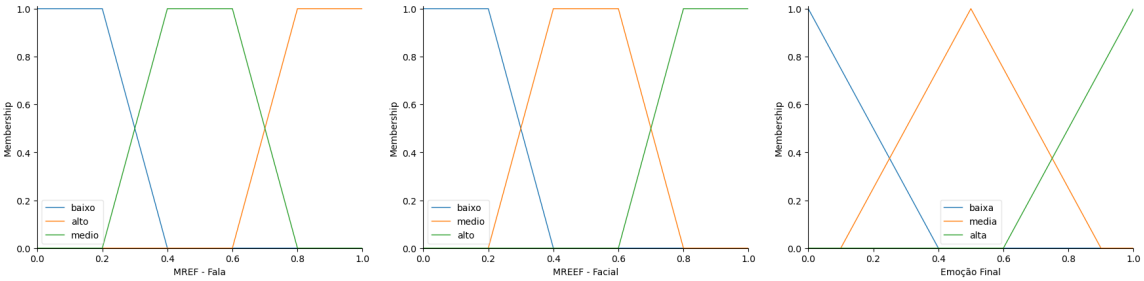


(a) Função de Pertinência do MREF (b) Função de Pertinência do MREEF (c) Função de Pertinência da Saída

Figura 6.2: Funções de pertinência antecedente (MREEF e MREF) com duas variáveis.

se MREF ALTO e MREEF ALTO então Emoção ALTA;
se MREF ALTO e MREEF BAIXO então Emoção MEDIA;
se MREF BAIXO e MREEF ALTO então Emoção MEDIA;
se MREF BAIXO e MREEF BAIXO então Emoção BAIXO.

Para o segundo teste, foram utilizadas três variáveis linguísticas: baixo, médio e alto, conforme pode ser visto na Figura 6.3. As regras *fuzzy* utilizadas foram:

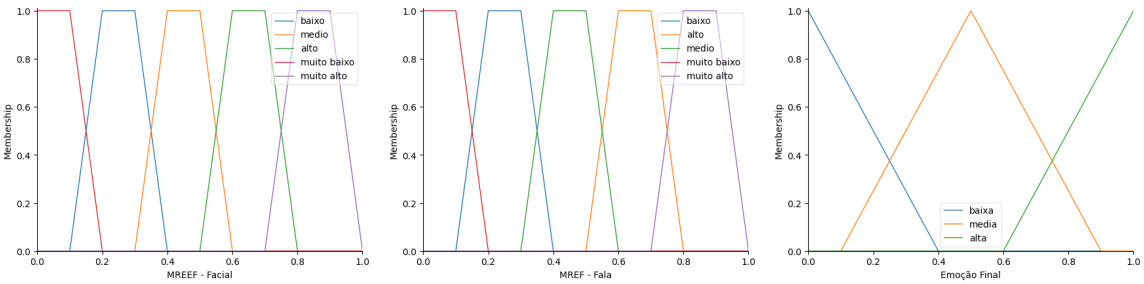


(a) Função de Pertinência do MREF (b) Função de Pertinência do MREEF (c) Função de Pertinência da Saída

Figura 6.3: Funções de pertinência antecedente (MREEF e MREF) com três variáveis.

se MREF ALTO e MREEF ALTO então Emoção ALTA;
se MREF ALTO e MREEF BAIXO então Emoção MEDIA;
se MREF BAIXO e MREEF ALTO então Emoção MEDIA;
se MREF MEDIO e MREEF ALTO então Emoção MEDIA;
se MREF MEDIO e MREEF BAIXO então Emoção MEDIA;
se MREF ALTO e MREEF MEDIO então Emoção MEDIA;
se MREF BAIXO e MREEF MEDIO então Emoção MEDIA;
se MREF MEDIO e MREEF MEDIO então Emoção MEDIA;
se MREF BAIXO e MREEF BAIXO então Emoção BAIXA.

E por fim, o terceiro teste consiste em cinco variáveis linguísticas para as FP de entrada, muito baixo, baixo, médio, alto e muito alto, conforme ilustrado na Figura 6.4. As regras *fuzzy* geraram uma combinação de 75 regras, alguma delas são:



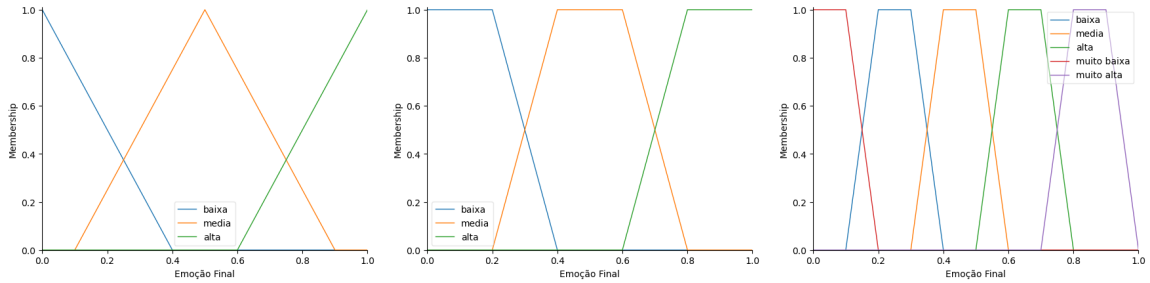
(a) Função de Pertinência do MREEF (b) Função de Pertinência do MREF (c) Função de Pertinência da Saída

Figura 6.4: Funções de pertinência antecedente (MREEF e MREF) com cinco variáveis.

se MREF MUITO ALTO e MREEF MUITO ALTO então Emoção ALTA;
se MREF ALTO e MREEF MUITO ALTO então Emoção ALTA;
se MREF BAIXO e MREEF ALTO então Emoção MEDIA;
se MREF MEDIO e MREEF ALTO então Emoção MEDIA;
se MREF MEDIO e MREEF MEDIO então Emoção MEDIA;
se MREF BAIXO e MREEF BAIXO então Emoção BAIXA.
se MREF MUITO BAIXO e MREEF BAIXO então Emoção BAIXA.
se MREF MUITO BAIXO e MREEF MUITO BAIXO então Emoção BAIXA.

Função de Pertinência Consequente

Após a seleção da melhor função de pertinência antecedente para as duas entradas (MREF e MREEF), foram investigadas a melhor FP consequente. Neste sentido foram testadas outras duas FP de saída, conforme mostrado na Figura 6.5 (b) e (c). As novas FP utilizadas são no formato trapezoidal, sendo que a FP da Figura 6.5(b) mantém o número de variáveis (alta, média e baixa) e FP da Figura 6.5(c) contém cinco variáveis linguísticas de saída (muito alta, alta, média, baixa e muito baixa). Método de defuzzificação para calcular valores de saída nítidos do conjunto difuso foi o centroide.



(a) FP Consequente 01 - padrão (b) FP Consequente 02 (c) FP Consequente 03

Figura 6.5: Funções de pertinência consequente (Saída).

6.4 Resultados

Neste capítulo serão apresentados os resultados obtidos seguindo a proposta a metodologia descrita na Seção 6.3. Para realizar os experimentos utiliza-se um notebook com sistema operacional Windows 11, processador Intel i5 1135G7 2.40GHz, memória RAM DE 8 GB e uma GPU Nvidia GeForce MX350 com 2GB VRAM. O ambiente de desenvolvimento utilizado neste trabalho é composto pela IDE Jupyter, associada a linguagem Python. Para a elaboração

do sistema *fuzzy* descrito, foi utilizado a biblioteca do Python scikit-fuzzy [1].

A primeira investigação diz respeito as funções de pertinência antecedente, alterando com 2, 3 e 5 variáveis linguísticas. O resultado obtido por essa abordagem é dado pela Tabela 6.2. No qual pode-se observar que a função de transferência que obteve um melhor desempenho foi com 5 variáveis (muito baixo, baixo, médio, alto e muito alto).

A segunda investigação é relativo a função de pertinência consequente, com 3 e 5 variáveis linguísticas. O resultado descrito na Tabela 6.2 informa o valor de acurácia obtido, sendo que a FP com 5 variáveis de saída obteve melhor desempenho, com 78,94%.

Tabela 6.2: Resultados dos experimentos da base de dados MELD.

Técnica	Variáveis	Acurácia (%)
FP Antecedente	Com 2	34,09
	Com 3	53,88
	Com 5	69,27
FP Consequente	Com 3	69,98
	Com 5	**78,94**

A Tabela 6.3, mostra os valores de acurácia para os testes realizados utilizando somente o módulo da expressão facial (MREEF), somente o módulo da fala (MREF) e a fusão dos módulos utilizando o *fuzzy*. A partir dos resultados apresentados na Tabela 6.3, é possível observar que a fusão das emoções aumenta a acurácia em quase 1%. Nesse sentido, é possível notar que o sistema nebuloso alcançou o objetivo esperado de unir as duas modalidades propostas.

Tabela 6.3: Resultados dos experimentos da base de dados MELD.

Técnica	Acurácia (%)
MREEF	78,06
MREF	73,00
MREEF + MREF + Fuzzy	78,94

[1] https://pypi.org/project/scikit-fuzzy/

CAPÍTULO 7

Considerações Finais

Este trabalho teve como objetivo propor um sistema de reconhecimento de emoção multimodal e multiclasse para a interação humano-robô. Neste sentido, foi proposto o Módulo de Reconhecimento de Emoção da Fala (MREF), o Módulo de Reconhecimento de Emoção da Expressão Facial (MREEF) e a fusão das emoções. Por tanto, nesta sessão, serão discutidos os resultados obtidos e apresentados alguns aspectos em aberto para discussão de trabalhos de continuidade.

A metodologia proposta para o Módulo de Reconhecimento de Emoções da Fala obteve resultados notáveis que se destacam na literatura. Os resultados de acurácia nas bases de dados foram: RAVDESS atingiu 97,01%, TESS 100%, SAVEE 90,62%, e R+T+S 97,37%. A melhor configuração obtida, de forma geral, envolveu o uso do otimizador Adam com taxa de aprendizado variável, aplicação de *data augmentation* no formato alongamento temporal, extração de características no estilo MFCC e a implementação de uma arquitetura neural híbrida composta por CNN com 4 blocos e LSTM. Essas configurações foram transferidas para as bases de dados, CREMA-D e R+T+S+C utilizando MtL. A base CREMA-D registrou um aumento na acurácia de 2.16%, enquanto a R+T+S+C obteve um incremento de 4.07%. Isso sugere que é viável aplicar o meta-aprendizado em diversas bases de dados de reconhecimento de emoções na fala (SER), considerando otimizador, taxa de aprendizado, D.A., extração de características e arquitetura neural.

Com relação à metodologia proposta para o Módulo de Reconhecimento de Emoções da Expressão Facial, os resultados demonstraram uma notável melhoria no valor da acurácia. A base de dados FER2013 apresentou uma acurácia de 75%, sendo amplamente reconhecido na

literatura que essa base de dados (FER2013) é desafiadora devido à sua complexidade. Portanto, ao comparar o desempenho da metodologia proposta com os resultados obtidos na literatura, constata-se que a acurácia alcançada, de 75,05%, está dentro da faixa observada na pesquisa, reforçando a consistência e relevância dos resultados obtidos. Esta melhoria significativa na acurácia valida a eficácia da abordagem proposta, destacando seu potencial para aplicações práticas no reconhecimento de emoções pela expressão facial.

E por fim, a metodologia para buscar uma melhor configuração do sistema nebuloso, apresentado no Capítulo 6, demonstrou eficacia em seus resultados. Ao compararmos as três abordagens: facial, fala e a fusão, foi possível observar que ao realizar a fusão o valor de acurácia para a base de dados MELD aumentou em quase 1%. Nesse sentido, a fusão proposta demonstra sua eficácia e eficiência.

Em trabalhos de continuidade deseja-se realizar a inferência da personalidade da pessoa com quem o robô estará interagindo. No entanto, até o momento, não conseguimos o acesso a uma base de dados para reconhecer os cinco traços de personalidade: abertura a experiência, responsabilidade, extroversão, cordialidade e neuroticismo. E com a inferência da personalidade seria desejável que o método proposto também realize o reconhecimento biométrico da pessoa com quem está interagindo. Desta forma, no futuro, equiparia o robô Hibot com mais informações para a tomada de decisão.

Referências Bibliográficas

Aguiar, G. J., Santana, E. J., de Carvalho, A. C., e Junior, S. B. (2022). Using meta-learning for multi-target regression. *Information Sciences*, 584:665–684.

Ahmed, M. R., Islam, S., Islam, A. M., e Shatabda, S. (2023). An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition. *Expert Systems with Applications*, 218:1–21.

Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., e Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*.

Andreasson, R., Alenljung, B., Billing, E., e Lowe, R. (2018). Affective touch in human–robot interaction: conveying emotion to the nao robot. *International Journal of Social Robotics*, 10(4):473–491.

Arkin, R. C., Fujita, M., Takagi, T., e Hasegawa, R. (2003). An ethological and emotional basis for human–robot interaction. *Robotics and Autonomous Systems*, 42(3-4):191–201.

Ashok, A., Pawlak, J., Paplu, S., Zafar, Z., e Berns, K. (2022). Paralinguistic cues in speech to adapt robot behavior in human-robot interaction. In *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pp. 01–06. IEEE.

Asiya, U. e Kiran, V. (2021). Speech emotion recognition-a deep learning approach. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 867–871. IEEE.

Baek, J.-Y. e Lee, S.-P. (2023). Enhanced speech emotion recognition using dcgan-based data augmentation. *Electronics*, 12(18):1–11.

Bagheri, E., Roesler, O., Cao, H.-L., e Vanderborght, B. (2020). A reinforcement learning based cognitive empathy framework for social robots. *International Journal of Social Robotics*, pp. 1–15.

Bajracharya, M., Maimone, M. W., e Helmick, D. (2008). Autonomy for mars rovers: Past, present, and future. *Computer*, 41(12):44–50.

Banik, S. C., Gupta, A. K. S., Habib, M., e Mousumi, R. (2013). Determination of active personal space based on emotion when interacting with a service robot. *International Journal of Advanced Robotic Systems*, 10(3):1–7.

Basori, A. H. (2013). Emotion walking for humanoid avatars using brain signals. *International Journal of Advanced Robotic Systems*, 10(1):1–11.

Bautista, J. L., Lee, Y. K., e Shin, H. S. (2023). Speech emotion recognition based on parallel cnn-attention networks with multi-fold data augmentation. *Electronics*, 11(23):1–14.

Belo, J. P. R., Romero, R. A., e Azevedo, H. (2017). Simulador para sistemas cognitivos voltado para robótica social. *XIII Simpósio Brasileiro de Automação Inteligente (SBAI)*.

Bennett, C. C. e Šabanović, S. (2014). Deriving minimal features for human-like facial expressions in robotic faces. *International Journal of Social Robotics*, 6(3):367–381.

Berry, C. A. (2015). Teaching a first course in human-robot interaction. *The ASEE Computers in Education (CoED) Journal*, 6(4):100.

Bhangale, K. e Kothandaraman, M. (2023). Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics*, 12(4):1–14.

Boada, J. P. (2021). Prolegómenos a una ética para la robótica social. *Dilemata*, (34):71–87.

Boada, J. P., Maestre, B. R., e Genís, C. T. (2021). The ethical issues of social assistive robotics: A critical literature review. *Technology in Society*, 67:101726.

Bodapati, J. D., Srilakshmi, U., e Veeranjaneyulu, N. (2022). Fernet: a deep cnn architecture for facial expression recognition in the wild. *Journal of The institution of engineers (India): series B*, 103(2):439–448.

Bove, M. S., Cerqueira, J. J., e Simas Filho, E. F. (2020). Novelty detection applied in recognition of facial expressions. *Anais da Sociedade Brasileira de Automática*, 2(1).

Boyette, L.-L., Korver-Nieberg, N., Meijer, C., de Haan, L., et al. (2014). Quality of life in patients with psychotic disorders: impact of symptoms, personality, and attachment. *The Journal of nervous and mental disease*, 202(1):64–69.

Brazdil, P., van Rijn, J. N., Soares, C., e Vanschoren, J. (2022). *Metalearning: applications to automated machine learning and data mining*. Springer Nature.

Broadbent, E., Lee, Y. I., Stafford, R. Q., Kuo, I. H., e MacDonald, B. A. (2011). Mental schemas of robots as more human-like are associated with higher blood pressure and negative emotions in a human-robot interaction. *International Journal of Social Robotics*, 3(3):291.

Burke, J. L., Murphy, R. R., Rogers, E., Lumelsky, V. J., e Scholtz, J. (2004). Final report for the darpa/nsf interdisciplinary study on human-robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):103–112.

Caiado, R. G. G., Scavarda, L. F., Gavião, L. O., Ivson, P., de Mattos Nascimento, D. L., e Garza-Reyes, J. A. (2021). A fuzzy rule-based industry 4.0 maturity model for operations and supply chain management. *International Journal of Production Economics*, 231:107883.

Camada, M. Y., Cerqueira, J. J. F., e Lima, A. M. (2021). Computational model for identifying stereotyped behaviors and determining the activation level of pseudo-autistic. *Applied Soft Computing*, 99:106877.

Cano, S., González, C. S., Gil-Iranzo, R. M., e Albiol-Pérez, S. (2021). Affective communication for socially assistive robots (sars) for children with autism spectrum disorder: A systematic review. *Sensors*, 21(15):5166.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., e Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Cattell, R. B. (1946). Description and measurement of personality.

Chatterjee, R., Mazumdar, S., Sherratt, R. S., Halder, R., Maitra, T., e Giri, D. (2021). Real-time speech emotion analysis for smart home assistants. *IEEE Transactions on Consumer Electronics*, 67(1):68–76.

Chitre, N., Bhorade, N., Topale, P., Ramteke, J., e Gajbhiye, C. (2022). Speech emotion recognition to assist autistic children. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 983–990. IEEE.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Claret, J.-A., Venture, G., e Basañez, L. (2017). Exploiting the robot kinematic redundancy for emotion conveyance to humans as a lower priority task. *International journal of social robotics*, 9(2):277–292.

Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D., e Vanderborght, B. (2016). A survey of expectations about the role of robots in robot-assisted therapy for children with asd: Ethical acceptability, trust, sociability, appearance, and attachment. *Science and engineering ethics*, 22(1):47–65.

Compton, M. T., Bakeman, R., Alolayan, Y., Balducci, P. M., Bernardini, F., Broussard, B., Crisafio, A., Cristofaro, S., Amar, P., Johnson, S., et al. (2015). Personality domains, duration of untreated psychosis, functioning, and symptom severity in first-episode psychosis. *Schizophrenia research*, 168(1-2):113–119.

Crumpton, J. e Bethel, C. L. (2016). A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics*, 8(2):271–285.

Dautenhahn, K. e Billard, A. (2002). Games children with autism can play with robota, a humanoid robotic doll. In *Universal access and assistive technology*, pp. 179–190. Springer.

De Beir, A., Cao, H.-L., Esteban, P. G., Van de Perre, G., Lefeber, D., e Vanderborght, B. (2016). Enhancing emotional facial expressiveness on nao. *International Journal of Social Robotics*, 8(4):513–521.

de Lope, J. e Graña, M. (2023). An ongoing review of speech emotion recognition. *Neurocomputing*, pp. 1–11.

Devillers, L., Tahon, M., Sehili, M. A., e Delaborde, A. (2015). Inference of human beings' emotional states from speech in human–robot interactions. *International Journal of Social Robotics*, 7(4):451–463.

Dewi, C., Gunawan, L. S., Hastoko, S. G., e Christanto, H. J. (2023). Real-time facial expression recognition: Advances, challenges, and future directions. *Vietnam Journal of Computer Science*, pp. 1–27.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Dolka, H., VM, A. X., e Juliet, S. (2021). Speech emotion recognition using ann on mfcc features. In *2021 3rd international conference on signal processing and communication (ICPSC)*, pp. 431–435. IEEE.

dos Reis Alves, S. F. (2016). *Arquitetura de Controle Inteligente para Interação Humano-Robô*. Tese, Universidade de São Paulo (USP-São Carlos).

dos Reis Alves, S. F. e Ferasoli Filho, H. (2016). Intelligent control architecture for assistive mobile robots. *Journal of Control, Automation and Electrical Systems*, 27(5):515–526.

Duchi, J., Hazan, E., e Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Duquette, A., Michaud, F., e Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonomous Robots*, 24(2):147–157.

Eid, M. A. e Al Osman, H. (2015). Affective haptics: Current research and future directions. *IEEE Access*, 4:26–40.

Ekman, P. e Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. *Motivation and emotion*, 10(2):159–168.

Elgendy, M. (2020). *Deep learning for vision systems*. Simon and Schuster.

Esterwood, C. e Robert, L. P. (2021). A systematic review of human and robot personality in health care human-robot interaction. *Frontiers in Robotics and AI*, pp. 306.

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3?—criteria for a taxonomic paradigm. *Personality and individual differences*, 12(8):773–790.

Filippini, C., Spadolini, E., Cardone, D., Bianchi, D., Preziuso, M., Sciarretta, C., del Cimmuto, V., Lisciani, D., e Merla, A. (2020). Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: The case of mio amico robot. *International Journal of Social Robotics*, pp. 1–13.

Gebhard, P. (2005). Alma: a layered model of affect. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 29–36.

Giritlioğlu, D., Mandira, B., Yilmaz, S. F., Ertenli, C. U., Akgür, B. F., Kınıklıoğlu, M., Kurt, A. G., Mutlu, E., Gürel, Ş. C., e Dibeklioğlu, H. (2021). Multimodal analysis of personality traits on videos of self-presentation and induced behavior. *Journal on Multimodal User Interfaces*, 15(4):337–358.

Gockley, R. e MatariĆ, M. J. (2006). Encouraging physical therapy compliance with a hands-off mobile robot. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 150–155.

Goetz, J. e Kiesler, S. (2002). Cooperation with a robotic assistant. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, pp. 578–579, New York, NY, USA. Association for Computing Machinery.

Goodfellow, I., Bengio, Y., e Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodrich, M. A. e Schultz, A. C. (2008). *Human-robot interaction: a survey*. Now Publishers Inc.

Goodwin, M. S., Intille, S. S., Albinali, F., e Velicer, W. F. (2011). Automated detection of stereotypical motor movements. *Journal of autism and developmental disorders*, 41(6):770–782.

Goulart, C., Castillo, J., Valadão, C., Bastos, T., e Caldeira, E. (2014). Eeg analysis and mobile robot as tools for emotion characterization in autism. In *BMC proceedings*, volume 8, pp. P85. Springer.

Guizzo, E., Weyde, T., Scardapane, S., e Comminiello, D. (2023). Learning speech emotion representations in the quaternion domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Gupta, M. e Chandra, S. (2021). Speech emotion recognition using mfcc and wide residual network. In *2021 Thirteenth International Conference on Contemporary Computing (IC3-2021)*, pp. 320–327.

Gupta, M., Patel, T., Mankad, S. H., e Vyas, T. (2022). Detecting emotions from human speech: role of gender information. In *2022 IEEE Region 10 Symposium (TENSYMP)*, pp. 1–6. IEEE.

Hazra, S. K., Ema, R. R., Galib, S. M., Kabir, S., e Adnan, N. (2022). Emotion recognition of human speech using deep learning method and mfcc features. *Radioelectronic and Computer Systems*, (4):161–172.

Helaly, R., Messaoud, S., Bouaafia, S., Hajjaji, M. A., e Mtibaa, A. (2023). Dtl-i-resnet18: Facial emotion recognition based on deep transfer learning and improved resnet18. *Signal, Image and Video Processing*, 17(6):2731–2744.

Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., Graterol, W., e Aguilera, A. (2022). Adaptive multimodal emotion detection architecture for social robots. *IEEE Access*, 10:20727–20744.

Hermessi, H., Mourali, O., e Zagrouba, E. (2017). Multimodal image fusion based on non-subsampled shearlet transform and neuro-fuzzy. In *Representations, Analysis and Recognition of Shape and Motion from Imaging Data: 6th International Workshop, RFMI 2016, Sidi Bou Said Village, Tunisia, October 27-29, 2016, Revised Selected Papers 6*, pp. 161–175. Springer.

Hoffman, M. W., Grimes, D. B., Shon, A. P., e Rao, R. P. (2006). A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 19(3):299–310.

Howard, A. M. e Cruz, G. (2006). Adapting human leadership approaches for role allocation in human-robot navigation scenarios. In *2006 World Automation Congress*, pp. 1–8. IEEE.

Iacono, I., Lehmann, H., Marti, P., Robins, B., e Dautenhahn, K. (2011). Robots as social mediators for children with autism-a preliminary analysis comparing two different robotic

platforms. In *2011 IEEE international conference on development and learning (ICDL)*, volume 2, pp. 1–6. IEEE.

Jackson, P. e Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

James, J., Balamurali, B., Watson, C. I., e MacDonald, B. (2020). Empathetic speech synthesis and testing for healthcare robots. *International Journal of Social Robotics*, pp. 1–19.

Javed, U., Riaz, M. M., Ghafoor, A., Ali, S. S., Cheema, T. A., et al. (2014). Mri and pet image fusion using fuzzy logic and image local features. *The Scientific World Journal*, 2014.

Jerčić, P., Wen, W., Hagelbäck, J., e Sundstedt, V. (2018a). The effect of emotions and social behavior on performance in a collaborative serious game between humans and autonomous robots. *International Journal of Social Robotics*, 10(1):115–129.

Jerčić, P., Wen, W., Hagelbäck, J., e Sundstedt, V. (2018b). The effect of emotions and social behavior on performance in a collaborative serious game between humans and autonomous robots. *International Journal of Social Robotics*, 10(1):115–129.

Jiang, Y., Xiao, N., e Han, J. (2013). Automatic control of contextual interaction integrated with affection and architectural attentional control. *International Journal of Advanced Robotic Systems*, 10(3):1–11.

Johnson, D. O., Cuijpers, R. H., Juola, J. F., Torta, E., Simonov, M., Frisiello, A., Bazzani, M., Yan, W., Weber, C., Wermter, S., et al. (2014). Socially assistive robots: a comprehensive approach to extending independent living. *International journal of social robotics*, 6(2):195–211.

Johnson, D. O., Cuijpers, R. H., Pollmann, K., e van de Ven, A. A. (2016). Exploring the entertainment value of playing games with a humanoid robot. *International Journal of Social Robotics*, 8(2):247–269.

Johnson, D. O., Cuijpers, R. H., e van der Pol, D. (2013). Imitating human emotions with artificial facial expressions. *International Journal of Social Robotics*, 5(4):503–513.

Jothimani, S. e Premalatha, K. (2022). Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. *Chaos, Solitons & Fractals*, 162:112–512.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., et al. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111.

Ke, X., Cao, B., Bai, J., Zhang, W., e Zhu, Y. (2020). An interactive system for humanoid robot shfr-iii. *International Journal of Advanced Robotic Systems*, 17(2):1–14.

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., e Acharya, U. R. (2023). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, pp. 1–36.

Khosla, R., Chu, M.-T., e Nguyen, K. (2013). Enhancing emotional well being of elderly using assistive social robots in australia. In *2013 International Conference on Biometrics and Kansei Engineering*, pp. 41–46. IEEE.

Kim, B., Joo, Y. H., Kim, S. Y., Lim, J.-H., e Kim, E. O. (2011). Personality traits and affective morbidity in patients with bipolar i disorder: the five-factor model perspective. *Psychiatry research*, 185(1-2):135–140.

Kim, B., Lim, J.-H., Kim, S. Y., e Joo, Y. H. (2012a). Comparative study of personality traits in patients with bipolar i and ii disorder from the five-factor model perspective. *Psychiatry investigation*, 9(4):347.

Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., e Scassellati, B. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *Journal of autism and developmental disorders*, 43(5):1038–1049.

Kim, E. S., Paul, R., Shic, F., e Scassellati, B. (2012b). Bridging the research gap: Making hri useful to individuals with autism. *Journal of Human-robot interaction*, 1(1):26–54.

Kim, J. e Lee, D. (2023a). Facial expression recognition robust to occlusion and to intra-similarity problem using relevant subsampling. *Sensors*, 23(5):2619.

Kim, S. e Lee, S.-P. (2023b). A bilstm–transformer and 2d cnn architecture for emotion recognition from speech. *Electronics*, 12(19):1–14.

Koay, C. G., Özarslan, E., e Basser, P. J. (2009). A signal transformational framework for breaking the noise floor and its applications in mri. *Journal of magnetic resonance*, 197(2):108–119.

Koch, S. A., Stevens, C. E., Clesi, C. D., Lebersfeld, J. B., Sellers, A. G., McNew, M. E., Biasini, F. J., Amthor, F. R., e Hopkins, M. I. (2017). A feasibility study evaluating the emotionally expressive robot sam. *International Journal of Social Robotics*, 9(4):601–613.

Kotov, R., Gamez, W., Schmidt, F., e Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological bulletin*, 136(5):768.

Kozima, H., Michalowski, M. P., e Nakagawa, C. (2009a). Keepon. *International Journal of Social Robotics*, 1(1):3–18.

Kozima, H., Michalowski, M. P., e Nakagawa, C. (2009b). Keepon. *International Journal of Social Robotics*, 1(1):3–18.

Kühnlenz, B., Sosnowski, S., Buß, M., Wollherr, D., Kühnlenz, K., e Buss, M. (2013). Increasing helpfulness towards a robot by emotional adaption to the user. *International Journal of Social Robotics*, 5(4):457–476.

Kumar, N., Kaushal, R., Agarwal, S., e Singh, Y. B. (2021). Cnn based approach for speech emotion recognition using mfcc, croma and stft hand-crafted features. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 981–985. IEEE.

Lan, Y.-T., Liu, W., e Lu, B.-L. (2020). Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE.

Landowska, A. (2013). Affect-awareness framework for intelligent tutoring systems. In *2013 6th International Conference on Human System Interactions (HSI)*, pp. 540–547. IEEE.

Law, T., de Leeuw, J., e Long, J. H. (2020). How movements of a non-humanoid robot affect emotional perceptions and trust. *International Journal of Social Robotics*, pp. 1–12.

Lawpanom, R., Songpan, W., e Kaewyotha, J. (2024). Advancing facial expression recognition in online learning education using a homogeneous ensemble convolutional neural network approach. *Applied Sciences*, 14(3):1156.

Lecomte, T., Spidel, A., Leclerc, C., MacEwan, G. W., Greaves, C., e Bentall, R. P. (2008). Predictors and profiles of treatment non-adherence and engagement in services problems in early psychosis. *Schizophrenia research*, 102(1-3):295–302.

Lee, J.-J., Kim, D.-W., e Kang, B.-Y. (2012). Exploiting child-robot aesthetic interaction for a social robot. *International Journal of Advanced Robotic Systems*, 9(3):1–9.

Lee, J.-J., Kim, D.-W., e Kang, B.-Y. (2017). Esthetic interaction model of robot with human to develop social affinity. *International Journal of Advanced Robotic Systems*, 14(4):1–16.

Lemke, C., Budka, M., e Gabrys, B. (2015). Metalearning: a survey of trends and technologies. *Artificial intelligence review*, 44:117–130.

Leuwerink, K. (2012). A robot with personality: Interacting with a group of humans. In *Proceedings of the 16th twente student conference on IT*, volume 4.

Li, H. e Xu, H. (2020). Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowledge-Based Systems*, 204:106172.

Li, J. e Chignell, M. (2011). Communication of emotion in social robots through simple head and arm movements. *International Journal of Social Robotics*, 3(2):125–142.

Liu, C., Ham, J., Postma, E., Midden, C., Joosten, B., e Goudbeek, M. (2013). Representing affective facial expressions for robots and embodied conversational agents by facial landmarks. *International Journal of Social Robotics*, 5(4):619–626.

Livingstone, S. R. e Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Lu, J. e Wan, X. (2023). Affective computing model for natural interaction based on large-scale self-built dataset. *SN Applied Sciences*, 5(2):53.

Lysaker, P. H. e Davis, L. W. (2004). Social function in schizophrenia and schizoaffective disorder: associations with personality, symptoms and neurocognition. *Health and quality of life outcomes*, 2(1):1–6.

Mamdani, E. e Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy. *International Journal of Man-Machine Studies*, 7(1):1–13.

Mantovani, R. G., Rossi, A. L., Alcobaça, E., Vanschoren, J., e de Carvalho, A. C. (2019). A meta-learning recommender system for hyperparameter tuning: Predicting when tuning improves svm classifiers. *Information Sciences*, 501:193–221.

Martín, F. A., Castro-González, Á., Salichs, M. Á., et al. (2017). Sound synthesis for communicating nonverbal expressive cues. *IEEE Access*, 5:1941–1957.

McColl, D. e Nejat, G. (2014). Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics*, 6(2):261–280.

Menne, I. M. e Schwab, F. (2018). Faces of emotion: investigating emotional facial expressions towards a robot. *International Journal of Social Robotics*, 10(2):199–209.

Michaud, F. e Clavet, A. (2001). Robotoy contest-designing mobile robotic toys for autistic children. *Proc. of the American Society for Engineering Education (ASEE'01)*.

Mittal, R., Vart, S., Shokeen, P., e Kumar, M. (2022). Speech emotion recognition. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pp. 1–6. IEEE.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., e Manocha, D. (2020). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1359–1367.

Moshkina, L. e Arkin, R. C. (2005). Human perspective on affective robotic behavior: A longitudinal study. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1444–1451. IEEE.

Mulder, R. T. (2002). Personality pathology and treatment outcome in major depression: a review. *American Journal of Psychiatry*, 159(3):359–371.

Myna, A. e Prakash, J. (2015). Fusion of ct and mri images based on fuzzy logic and discrete wavelet transform. *Int. J. Comput. Sci. Inf. Technol*, 6(5):4512–4519.

Nasim, A. S., Chowdory, R. H., Dey, A., e Das, A. (2021). Recognizing speech emotion based on acoustic features using machine learning. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 1–7. IEEE.

Neto, A. F. B. e da Silva, F. S. C. (2012). A computer architecture for intelligent agents with personality and emotions. In *Human-Computer Interaction: The Agency Perspective*, pp. 263–285. Springer.

Nguyen, H.-D., Yeom, S., Lee, G.-S., Yang, H.-J., Na, I.-S., e Kim, S.-H. (2019). Facial emotion recognition using an ensemble of multi-level convolutional neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(11):1940015.

Nieuwenhuisen, M. e Behnke, S. (2013). Human-like interaction skills for the mobile communication robot robotinho. *International Journal of Social Robotics*, 5(4):549–561.

Nomura, T. e Nakao, A. (2010). Comparison on identification of affective body motions by robots between elder people and university students: A case study in japan. *International Journal of Social Robotics*, 2(2):147–157.

Ohi, K., Shimada, T., Nitta, Y., Kihara, H., Okubo, H., Uehara, T., e Kawasaki, Y. (2016). The five-factor model personality traits in schizophrenia: a meta-analysis. *Psychiatry research*, 240:34–41.

Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., e Ohme, R. K. (2015). Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. *Frontiers in psychology*, 5:1516.

Ortega, J. D., Senoussaoui, M., Granger, E., Pedersoli, M., Cardinal, P., e Koerich, A. L. (2019). Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*.

Ortega, M. G. S., Rodríguez, L. F., e Gutierrez-Garcia, J. O. (2020). Towards emotion recognition from contextual information using machine learning. *Journal of Ambient Intelligence and Humanized Computing*, 11(8):3187–3207.

Ottoni, A. L. C., de Amorim, R. M., Novo, M. S., e Costa, D. B. (2023a). Tuning of data augmentation hyperparameters in deep learning to building construction image classification

with small datasets. *International Journal of Machine Learning and Cybernetics*, 14(1):171–186.

Ottoni, A. L. C., Souza, A. M., e Novo, M. S. (2023b). Automated hyperparameter tuning for crack image classification with deep learning. *Soft Computing*, 27(23):18383–18402.

Ottoni, L. T. C. e Cerqueira, J. J. F. (2021). A review of emotions in human-robot interaction. In *2021 Latin American Robotics Symposium (LARS)*, pp. 7–12. IEEE.

Pais, A. L., Argall, B. D., e Billard, A. G. (2013). Assessing interaction dynamics in the context of robot programming by demonstration. *International Journal of Social Robotics*, 5(4):477–490.

Pan, S.-T. e Wu, H.-J. (2023). Performance improvement of speech emotion recognition systems by combining 1d cnn and lstm with data augmentation. *Electronics*, 12(11):1–21.

Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., e Akagi, M. (2020). Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access*, 8:16560–16572.

Pham, T.-D., Duong, M.-T., Ho, Q.-T., Lee, S., e Hong, M.-C. (2023). Cnn-based facial expression recognition with simultaneous consideration of inter-class and intra-class variations. *Sensors*, 23(24):9658.

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64.

Pichora-Fuller, M. K. e Dupuis, K. (2020). Toronto emotional speech set (tess). *Scholars Portal Dataverse*, 1:2020.

Poria, S., Chaturvedi, I., Cambria, E., e Hussain, A. (2016). Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 439–448. IEEE.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., e Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Prado, J. A., Simplício, C., Lori, N. F., e Dias, J. (2012). Visuo-auditory multimodal emotional structure to improve human-robot-interaction. *International journal of social robotics*, 4(1):29–51.

Rad, N. M., Kia, S. M., Zarbo, C., van Laarhoven, T., Jurman, G., Venuti, P., Marchiori, E., e Furlanello, C. (2018). Deep learning for automatic stereotypical motor movement detection using wearable sensors in autism spectrum disorders. *Signal Processing*, 144:180–191.

Rairán, J. D. e Nino, L. F. (2017). Robot motion control based on anticipatory emotions. *International Journal of Advanced Robotic Systems*, 14(6):1–9.

Read, R. e Belpaeme, T. (2016). People interpret robotic non-linguistic utterances categorically. *International Journal of Social Robotics*, 8(1):31–50.

Reich-Stiebert, N., Eyssel, F., e Hohnemann, C. (2019). Exploring university students' preferences for educational robot design by means of a user-centered design approach. *International Journal of Social Robotics*, pp. 1–11.

Reif, M., Shafait, F., e Dengel, A. (2012). Meta-learning for evolutionary parameter optimization of classifiers. *Machine learning*, 87:357–380.

Remazeilles, A., Leroux, C., e Chalubert, G. (2008). Sam: a robotic butler for handicapped people. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 315–321. IEEE.

Reyes, M. E., Meza, I. V., e Pineda, L. A. (2019). Robotics facial expression of anger in collaborative human–robot interaction. *International Journal of Advanced Robotic Systems*, 16(1):1–11.

Richardson, R., Devereux, D., Burt, J., e Nutter, P. (2012). Humanoid upper torso complexity for displaying gestures. *International Journal of Advanced Robotic Systems*, 9(1):1–9.

Robert, L. (2018). Personality in the human robot interaction literature: A review and brief critique. In *Robert, LP (2018). Personality in the Human Robot Interaction Literature: A Review and Brief Critique, Proceedings of the 24th Americas Conference on Information Systems, Aug*, pp. 16–18.

Robins, B., Dautenhahn, K., Te Boekhorst, R., e Billard, A. (2004). Effects of repeated exposure to a humanoid robot on children with autism. In *Designing a more inclusive world*, pp. 225–236. Springer.

Robins, B., Dautenhahn, K., Te Boekhorst, R., e Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105–120.

Robins, B., Ferrari, E., Dautenhahn, K., Kronreif, G., Prazak-Aram, B., Gelderblom, G.-j., Tanja, B., Caprino, F., Laudanna, E., e Marti, P. (2010). Human-centred design methods: Developing scenarios for robot assisted play informed by user panels and field trials. *International Journal of Human-Computer Studies*, 68(12):873–898.

Romero, R. A. F., Prestes, E., Osório, F., e Wolf, D. (2014). *Robótica móvel.* LTC.

Rosenthal-von der Pütten, A. M., Krämer, N. C., e Herrmann, J. (2018). The effects of human-like and robot-specific affective nonverbal behavior on perception, emotion, and behavior. *International Journal of Social Robotics*, 10(5):569–582.

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., e Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1):17–34.

Rumelhart, D. E., Hinton, G. E., e Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Sadouk, L., Gadi, T., e Essoufi, E. H. (2018). A novel deep learning approach for recognizing stereotypical motor movements within and across subjects on the autism spectrum disorder. *Computational intelligence and neuroscience*, 2018.

Sahoo, G. K., Das, S. K., e Singh, P. (2023). Performance comparison of facial emotion recognition: a transfer learning-based driver assistance framework for in-vehicle applications. *Circuits, Systems, and Signal Processing*, 42(7):4292–4319.

Sajjad, M., Kwon, S., et al. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 8:79861–79875.

Saldien, J., Goris, K., Vanderborght, B., Vanderfaeillie, J., e Lefeber, D. (2010). Expressing emotions with the social robot probo. *International Journal of Social Robotics*, 2(4):377–389.

Salem, M., Lakatos, G., Amirabdollahian, F., e Dautenhahn, K. (2015). Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 1–8. IEEE.

Sasiadek, J. Z. (2002). Sensor fusion. *Annual Reviews in Control*, 26(2):203–228.

Scassellati, B., Admoni, H., e Matarić, M. (2012). Robots for use in autism research. *Annual review of biomedical engineering*, 14:275–294.

Shanthi, N., Stonier, A. A., Sherine, A., Devaraju, T., Abinash, S., Ajay, R., Arul Prasath, V., e Ganji, V. (2022). An integrated approach for mental health assessment using emotion analysis and scales. *Healthcare Technology Letters*.

Silva, I. N. d., Spatti, D. H., e Flauzino, R. A. (2010). Redes neurais artificiais para engenharia e ciências aplicadas.

Silva, I. N. d., Spatti, D. H., e Flauzino, R. A. (2016). *Redes neurais artificiais para engenharia e ciências aplicadas*. Artliber.

Silvera-Tawil, D., Rye, D., e Velonaki, M. (2014). Interpretation of social touch on an artificial arm covered with an eit-based sensitive skin. *International Journal of Social Robotics*, 6(4):489–505.

Simut, R. E., Vanderfaeillie, J., Peca, A., Van de Perre, G., e Vanderborght, B. (2016). Children with autism spectrum disorders make a fruit salad with probo, the social robot: an interaction study. *Journal of autism and developmental disorders*, 46(1):113–126.

Singh, H., Raj, J., Kaur, G., e Meitzler, T. (2004). Image fusion using fuzzy logic and applications. In *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542)*, volume 1, pp. 337–340. IEEE.

Singh, J., Saheer, L. B., e Faust, O. (2023). Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 20(6):1–21.

Song, Z. (2021). Facial expression emotion recognition model integrating philosophy and machine learning theory. *Frontiers in Psychology*, 12:759485.

Subramanian, V., Burks, T., e Dixon, W. (2009). Sensor fusion using fuzzy logic enhanced kalman filter for autonomous vehicle guidance in citrus groves. *Transactions of the ASABE*, 52(5):1411–1422.

Sun, B., Xu, Q., He, J., Yu, L., Li, L., e Wei, Q. (2016). Audio-video based multimodal emotion recognition using svms and deep learning. In *Chinese Conference on Pattern Recognition*, pp. 621–631. Springer.

Syrdal, D. S., Dautenhahn, K., Woods, S., Walters, M. L., e Koay, K. L. (2006). 'doing the right thing wrong'-personality and tolerance to uncomfortable robot approaches. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 183–188. IEEE.

Syrdal, D. S., Dautenhahn, K., Woods, S. N., Walters, M. L., e Koay, K. L. (2007). Looking good? appearance preferences and robot personality inferences at zero acquaintance. In *AAAI Spring symposium: multidisciplinary collaboration for socially assistive robotics*, volume 86.

Takagi, T. e Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(1):116–132.

Talpur, N., Abdulkadir, S. J., Alhussian, H., Aziz, N., Bamhdi, A., et al. (2022). A comprehensive review of deep neuro-fuzzy system architectures and their optimization methods. *Neural Computing and Applications*, pp. 1–39.

Talpur, N., Abdulkadir, S. J., Alhussian, H., Hasan, M. H., Aziz, N., e Bamhdi, A. (2023). Deep neuro-fuzzy system application trends, challenges, and future perspectives: A systematic survey. *Artificial intelligence review*, 56(2):865–913.

Tay, B., Jung, Y., e Park, T. (2014). When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38:75–84.

Tkalčič, M., De Carolis, B., De Gemmis, M., Odić, A., e Košir, A. (2016). Emotions and personality in personalized services. In *Human-Computer Interaction Series*. Springer.

Tripathi, S., Tripathi, S., e Beigi, H. (2018). Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*.

Trovato, G., Kishi, T., Endo, N., Zecca, M., Hashimoto, K., e Takanishi, A. (2013). Cross-cultural perspectives on emotion expressive humanoid robotic head: recognition of facial expressions and symbols. *International Journal of Social Robotics*, 5(4):515–527.

Tsiourti, C., Weiss, A., Wac, K., e Vincze, M. (2019). Multimodal integration of emotional signals from voice, body, and context: effects of (in) congruence on emotion recognition and attitudes towards robots. *International Journal of Social Robotics*, 11(4):555–573.

Tzirakis, P., Chen, J., Zafeiriou, S., e Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53.

Val-Calvo, M., Álvarez-Sánchez, J. R., Ferrández-Vicente, J. M., e Fernández, E. (2020). Affective robot story-telling human-robot interaction: Exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access*, 8:134051–134066.

Valadão, C. T., Goulart, C., Rivera, H., Caldeira, E., Bastos Filho, T. F., Frizera-Neto, A., e Carelli, R. (2016). Analysis of the use of a robot to improve social skills in children with autism spectrum disorder. *Research on Biomedical Engineering*, 32(2):161–175.

Valadão, C. T. (2016). *Sistema de Supervisão e Controle para Interação Assistiva Humano-Robô*. Tese, Universidade Federal do Espírito Santo (UFES).

Vu, H. A., Yamazaki, Y., Dong, F., e Hirota, K. (2011). Emotion recognition based on human gesture and speech information using rt middleware. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pp. 787–791. IEEE.

Wainer, J., Robins, B., Amirabdollahian, F., e Dautenhahn, K. (2014). Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism. *IEEE Transactions on Autonomous Mental Development*, 6(3):183–199.

Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R., e Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, 24(2):159–178.

Wang, F. e Shen, X. (2023). Research on speech emotion recognition based on teager energy operator coefficients and inverted mfcc feature fusion. *Electronics*, 12(17):1–16.

Wang, W. e Huang, X. (2012). Toward speech and nonverbal behaviors integration for humanoid robot. *International Journal of Advanced Robotic Systems*, 9(3):1–10.

Wauthia, E., Lefebvre, L., Huet, K., Blekic, W., El Bouragui, K., e Rossignol, M. (2019). Examining the hierarchical influences of the big-five dimensions and anxiety sensitivity on anxiety symptoms in children. *Frontiers in psychology*, 10:1185.

Wu, Y. (2023). Facial expression recognition in classroom environment based on attention mechanism. In *Advances in Artificial Intelligence, Big Data and Algorithms*, pp. 689–695. IOS Press.

Xia, Y. e LeTendre, G. (2020). Robots for future classrooms: A cross-cultural validation study of "negative attitudes toward robots scale" in the us context. *International Journal of Social Robotics*, pp. 1–12.

Xu, G., Gao, X., Pan, L., Chen, S., Wang, Q., Zhu, B., e Li, J. (2018). Anxiety detection and training task adaptation in robot-assisted active stroke rehabilitation. *International Journal of Advanced Robotic Systems*, 15(6):1–18.

Yang, E. e Dorneich, M. C. (2017). The emotional, cognitive, physiological, and performance effects of variable time delay in robotic teleoperation. *International Journal of Social Robotics*, 9(4):491–508.

Yang, Y., Que, Y., Huang, S., e Lin, P. (2016). Multimodal sensor medical image fusion based on type-2 fuzzy logic in nsct domain. *IEEE Sensors Journal*, 16(10):3735–3745.

Yohanan, S. e MacLean, K. E. (2012). The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *International Journal of Social Robotics*, 4(2):163–180.

Zadeh, A., Chen, M., Poria, S., Cambria, E., e Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zhang, S.-F., Zhai, J.-H., Xie, B.-J., Zhan, Y., e Wang, X. (2019). Multimodal representation learning: Advances, trends and challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 1–6. IEEE.

Zielonka, M., Piastowski, A., Czyżewski, A., Nadachowski, P., Operlejn, M., e Kaczor, K. (2022). Recognition of emotions in speech using convolutional neural networks on different datasets. *Electronics*, 11(22):3831.

Zillig, L. M. P., Hemenover, S. H., e Dienstbier, R. A. (2002). What do we assess when we assess a big 5 trait? a content analysis of the affective, behavioral, and cognitive processes represented in big 5 personality inventories. *Personality and Social Psychology Bulletin*, 28(6):847–858.