Dissertação de Mestrado

Detecção de microações em vídeos faciais para análise de carga cognitiva em ambientes de aprendizado multimídia

Cristóvão da Silva Rodrigues Costa

Orientador:

Thales Miranda de Almeida Vieira

Maceió, Julho de 2023

Cristóvão da Silva Rodrigues Costa

Detecção de microações em vídeos faciais para análise de carga cognitiva em ambientes de aprendizado multimídia

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Thales Miranda de Almeida Vieira

Maceió, Julho de 2023

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Thales Miranda de Almeida Vieira - Orientador
Universidade Federal de Alagoas

Bruno Almeida Pimentel - Examinador
Universidade Federal de Alagoas

Diego Carvalho Nascimento - Examinador
Universidad de Atacama, Chile

Maceió, Julho de 2023

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecária: Helena Cristina Pimentel do Vale CRB4 - 661

C837d Costa, Cristóvão da Silva Rodrigues .
Detecção de microações em vídeos faciais para análise de carga cognitiva em ambientes de aprendizado multimídia / Cristóvão da Silva Rodrigues Costa. – 2023.
75 f. : il.

Orientador: Thales Miranda de Almeida Vieira.
Dissertação(mestrado em Informática) – Universidade Federal de Alagoas, Instituto de Computação. Maceió, 2023.

Bibliografia: f. 72-75.

1. Aprendizado de máquina. 2. Teoria da carga cognitiva. 3. Visão computacional. 4. Aprendizagem significativa. 5. Rede neural artificial. I. Título.

CDU: 004.78

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PPGI
Instituto de Computação/UFAL
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Folha de Aprovação

CRISTÓVÃO DA SILVA RODRIGUES COSTA

DETECÇÃO DE MICROAÇÕES EM VÍDEOS FACIAIS PARA ANÁLISE DE CARGA COGNITIVA EM AMBIENTES DE APRENDIZADO MULTIMÍDIA

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 31 de julho de 2023.

Banca Examinadora:

Prof. Dr. THALES MIRANDA DE ALMEIDA VIEIRA
UFAL – Instituto de Computação
Orientador

Prof. Dr. BRUNO ALMEIDA PIMENTEL
UFAL – Instituto de Computação
Examinador Interno

Prof. Dr. DIEGO CARVALHO DO NASCIMENTO
Universidad de Atacama-UDA/Chile
Examinador Externo

Dedico este trabalho a minha filha Catherine Santana Rodrigues
e a minha mãe Maria Cícera da Silva e pai José Rodrigues Costa.

Agradecimentos

Agradeço de coração a todos que contribuíram para a realização deste trabalho de mestrado. Em primeiro lugar, gostaria de expressar minha profunda gratidão aos meus pais, Maria e José, por me ensinarem os princípios da vida e valores que levo comigo sempre.

Aos meus queridos irmãos, que sempre estiveram ao meu lado, compartilhando momentos de alegria e desafios, agradeço por serem meus eternos amigos e por todo o suporte que me deram durante essa caminhada.

Gostaria também de expressar minha gratidão aos meus co-orientadores, Bruno Pimentel e Diego Nascimento, pela oportunidade de aprender com suas vastas experiências e conhecimentos. Sua orientação e orientação ajudaram a moldar este trabalho e me guiaram rumo ao sucesso.

Agradeço aos professores e colegas que estiveram presentes ao longo dessa jornada, pelos debates enriquecedores, pela troca de conhecimentos e pelo incentivo constante. Suas contribuições foram essenciais para o aprimoramento deste trabalho.

Por fim, quero estender meus agradecimentos especiais ao meu orientador principal, Thales Vieira, por seu constante apoio, dedicação, paciência e sabedoria. Sua orientação foi inestimável, e sem ela, essa dissertação não seria possível. Sua confiança em meu potencial e seu incentivo incansável me motivaram a superar desafios e buscar a excelência.

A todos os mencionados e a tantos outros que de alguma forma contribuíram para esta conquista, meu mais sincero agradecimento. Vocês foram peças fundamentais nessa trajetória e compartilham comigo a alegria desta realização.

"Quanto mais aumenta nosso conhecimento,
mais evidente fica nossa ignorância.
(John F. Kennedy)

Resumo

A evolução tecnológica dos últimos anos, acelerada pela pandemia do Covid19, foi responsável por uma rápida e contínua quebra de paradigmas no ensino, com a adoção de ambientes de aprendizagem multimídia. Durante o processo de aprendizagem nestes ambientes, a absorção do conteúdo pelo discente cai com o aumento do volume de informação transmitida aos alunos, ou seja, quando há uma sobrecarga cognitiva em um ou em ambos os canais visual e verbal. Atualmente são escassos os estudos que utilizam ferramentas de Visão Computacional e Ciência de Dados para a análise da Carga Cognitiva. Ferramentas desta natureza possibilitariam uma análise automatizada de grandes volumes de vídeos, e consequentemente a avaliação e geração de conteúdo multimídia que otimizem o aprendizado dos alunos. Neste trabalho adotou-se um estudo piloto com uma amostra de 13 alunos da faculdade de medicina da Universidad de Atacama, Chile. Assim, foi desenvolvida uma metodologia para extrair e investigar correlações entre características visuais da face dos alunos e a carga cognitiva. Foi usada uma base de vídeos faciais dos alunos assistindo, pela tela do computador, aulas com recursos multimídia. Esta base de vídeos foi inicialmente organizada e pré-processada, aplicando-se em seguida modelos de Aprendizado Profundo para extrair pontos de interesse visuais da face em cada quadro. As micro ações foram previamente anotadas pelo pesquisado e os dados resultantes foram avaliados para identificação de padrões relevantes relacionados à carga cognitiva. Além de responder à investigação principal desta pesquisa, os resultados deste estudo incluem uma prova de conceito para a análise da correlação das expressões faciais com a nota da prova do indivíduo para posterior análise da carga cognitiva em ambientes de aprendizagem multimídia. O código do experimento e das ferramentas foi disponibilizado publicamente através da URL: https://github.com/cristovaor/CAST

Palavras-chaves: Teoria da Carga Cognitiva; Aprendizado de Máquina; Visão Computacional; Aprendizado Profundo; Rede Neural Artificial.

Abstract

The technological evolution of recent years, accelerated by the Covid-19 pandemic, has been responsible for a rapid and continuous paradigm shift in education, with the adoption of multimedia learning environments. During the learning process in these environments, the absorption of content by students decreases as the volume of information transmitted to them increases. In other words, there is a cognitive overload in one or both visual and verbal channels. Currently, there is a scarcity of studies that use Computer Vision and Data Science resources for the analysis of Cognitive Load. Tools of this nature would enable automated analysis of large volumes of videos and, consequently, the evaluation and generation of multimedia content that optimize student learning. In this work, a pilot study was conducted with a sample of 13 students from the School of Medicine at the Universidad de Atacama, Chile. Thus, a methodology was developed to extract and investigate correlations between visual characteristics of the students' faces and cognitive load. A database of facial videos of students watching multimedia-enhanced lectures on their computer screens was used. This video database was initially organized and preprocessed, followed by the application of Deep Learning models to extract visual points of interest from the face in each frame. Micro-actions were previously annotated by the researcher, and the resulting data were evaluated to identify relevant patterns related to cognitive load. In addition to addressing the main investigation of this research, the results of this study include a proof of concept for analyzing the correlation of facial expressions with individual exam scores, for further analysis of cognitive load in multimedia learning environments. The code for the experiment and the tools was made publicly available at the URL: https://github.com/cristovaor/CAST

Keywords: Cognitive Load Theory; Machine Learning; Computer vision; Deep Learning; Artificial Neural Network.

Lista de Figuras

2.1 Tipos de Processamento Cognitivo Adaptado de Mayer (2009). 7
2.2 Fluxograma do Processo Cognitivo Adaptado de Mayer (2009). 7
2.3 Equilíbrio da Carga Cognitiva. Adaptado de Mayer (2009). 8
2.4 Aprendizado Supervisionado e Não Supervisionado Faceli et al. (2021). 11
2.5 Exemplo de Rede Neural Artificial. Fonte: Faceli et al. (2021). 13
2.6 Exemplo de Arquitetura de uma Rede CNN. Fonte: Oliveira, Alves e Malqui (2017). 15
2.7 Arquitetura de uma CNN. Fonte: Faceli et al. (2021). 15
2.8 Arquitetura da MobileNet com Depthwise Separable Convolution. Adaptado de Sandler et al. (2018). 16
2.9 Processo de Depthwise Separable Convolution. Adaptado de Sandler et al. (2018). 17
2.10 Exemplo de LSTM - Smagulova e James (2020) 18
2.11 Exemplo de BILSTM - Mao et al. (2022). 19

4.1 Fluxograma da Metodologia Aplicada no Estudo - Autor 2023 23
4.2 Vídeo exemplo dos estudantes. Autor 2023 24
4.3 Fluxo de coleta de dados por participante: primeiro deve ser medido o seu conhecimento prévio à aula (Fase 1); em seguida, devem ser gravados os vídeos faciais durante a transmissão de uma aula gravada (Fase 2); por fim, será medido o conhecimento com um pós-teste ao final da aula virtual (Fase 3) - Autor (2023). 25
4.4 Arquitetura do Face Mesh - MediaPipe (2020) 26
4.5 Exemplo de detecção de íris com o MediaPipe - MediaPipe (2020) 26
4.6 Numeração dos pontos coletados pela MediaPipe - MediaPipe (2020). 27
4.7 Ferramenta para visualização dos frames - Autor (2023). 30
4.8 Diagrama para geração da Amostra - Autor (2023). 32
4.9 Arquitetura dos classificadores de micro ação: a partir de uma sequência de pontos de interesse de 7 quadros consecutivos, o classificador foi projetado para reconhecer se uma ação específica está sendo executada no último quadro da sequência - Autor (2023). 33

5.1 Ganho cognitivo por Grupo de Teste - Autor. (2023). 37
5.2 Ganho na aprendizagem por Grupo - Autor. (2023). 38
5.3 Imagem do Aplicativo em Flask - Autor (2023). 40
5.4 Visualização t-SNE da Interação 1 e 2. 41
5.5 Visualização t-SNE da Interação 3 e 4. 42
5.6 Visualização t-SNE da Interação 5 e 6. 43
5.7 Visualização t-SNE da Interação 7 e 8. 44
5.8 Visualização t-SNE da Interação 9. 45
5.9 Boxplot por tipo de teste e ganho de aprendizado - Autor (2023) 59
5.10 Boxplot por tipo de teste x ação - Autor (2023) 61
5.11 Correlaçoes entre as variáveis - Autor (2023) 62
5.12 Boxplot por Tipo de Teste por Segundo - Autor (2023) 63
5.13 Correlaçoes entre as variaveis no tempo - Autor (2023) 64
5.14 Correlações entre as variáveis normalizada no tempo - Autor (2023) 65
5.15 Correlações entre as variáveis do tipo R - Autor (2023) 66
5.16 Correlações entre as variáveis do tipo NR - Autor (2023) 67

Lista de Tabelas

4.1 Conjunto de Pontos FaceMesh por Região Facial. MediaPipe (2020) 28
4.2 Exemplo de Anotação por Vídeo 30

5.1 Distribuição de genero dos Dados Originais segundo grupo de aulas. 36
5.2 Ganho de aprendizagem e diferença percentual relativa por aluno. 36
5.3 Conjunto de vídeos usado para experimentos com classificadores de micro ações. 38
5.4 Resultados dos testes aplicados nos vídeos com qualidade adequada. 39
5.5 Ações anotadas manualmente nos dados - Autor (2023) 39
5.6 Conjunto de características e ações correspondentes -Autor 2023 39
5.7 Melhores combinações de hiper parâmetros - Autor 2023 40
5.8 Divisão de Treino e Teste dos Dados 45
5.9 Tabela de erro absoluto dos classificadores de micro ações - Autor (2023) . . . 46
5.10 Tabela de erro relativo dos classificadores de micro ações - Autor (2023) 46
5.11 Resultados para cada classificador de micro ação: os descritores de vídeo de groundtruth e previstos são comparados, revelando erro relativo baixo para todas as ações. 47
5.12 Resumo dos resultados para o classificador olho fechado 49
5.13 Metricas coletadas em cada um dos treinamentos do modelo, seguindo a estratégia *leave-one-video-out* para o classificador olho fechado. 50
5.14 Resumo dos resultados - Olhando para o Canto 52
5.15 Metricas coletadas em cada um dos treinamentos do modelo, seguindo a estratégia *leave-one-video-out* para o classificador olhando para o canto. 52
5.16 Resumo dos resultados dos Interações de treinamento do Classificador - Ação Mexeu Lábios 54
5.17 Metricas coletadas em cada um dos treinamentos do modelo, seguindo a estratégia *leave-one-video-out* para o classificador mexeu lábios 54
5.18 Cortes executados nos frames. 56
5.19 Métricas de classificadores - Média Variação 57
5.20 Análise Estatística do Ganho de Aprendizado por Tipo de Aula 60

5.21 Dados Normalizados no Tempo 63

Lista de Abreviaturas e Siglas

AUC Area Under The Curve

AUs Action Units

CLM Constrained Local Model

CNN Convolutional Neural Network

ECG Eletrocardiograma

FACS Facial Action Coding System

GPU Graphics Processing Unit

HSV Sistema de cor Hue, Saturation, Value

KNN K-Nearest Neighbors

MAE Mean Absolute Error

MSE Mean Squared Error

MTCNN Multi-Task Cascaded Convolutional Neural Network

OLFE Olho Fechado

OLPC Olhando para Canto

MSO Mexeu Sobrancelha

MLA Mexeu Lábios

OPENCV Open Source Computer Vision Library

PCA Análise dos Componentes Principais

PET Tomografia por emissão de pósitrons

RAM Memória de Acesso Aleatório

RFE Recursive Feature Elimination

RGB Sistema de cor Red, Green, Blue

RNA Rede Neural Artificial

SVM Support Vector Machine

t-SNE t-Distributed Stochastic Neighbor Embedding

VR Virou Rosto

Lista de Símbolos

@	Arrouba
$\div$	divisão (e.g., $a \div b$)
$\equiv$	equivalente a (e.g., $a \equiv b$)
$\sum$	somatório (e.g., $\sum_{i=1}^{n} x_i$)
y	Rótulos dos dados de entrada
$\hat{y}$	Predição do modelo para uma entrada x
θ	Parâmetros do modelo
w	Vetor de pesos
ReLU	Função de ativação ReLU (Retified Linear Unit)
conv	Camada de convolução
pool	Camada de pooling

Sumário

Lista de Abreviaturas e Siglas **xii**

Lista de Símbolos **xiii**

1 Introdução **1**

1.1 Contextualização 1

1.2 Justificativa 2

1.3 Objetivo Geral e Objetivos Específicos 3

1.4 Estrutura do trabalho 4

2 Referencial Teórico **5**

2.1 Teoria da carga cognitiva 5

2.1.1 Teoria da carga cognitiva e o aprendizado multimídia 6

2.2 Expressões faciais e o estado psicológico 9

2.3 Aprendizado de Máquina 9

2.3.1 Conjunto de Dados 10

2.3.2 Formas de aprendizado de máquina 11

2.4 Visão Computacional 12

2.5 Redes Neurais Artificiais 13

2.5.1 REDES NEURAIS CONVOLUCIONAIS 14

2.5.2 MobileNet 16

2.5.3 Long Short-term Memory - LSTM 17

3 Trabalhos Relacionados **20**

3.1 Detecção Facial 20

3.2 Reconhecimento de pontos de interesse da face 21

3.3 Reconhecimento de Expressões Faciais 21

3.4 Carga Cognitiva e Aprendizado Multimídia 22

4 Metodologia **23**

4.1 Organização e Pré-Processamento das Imagens 24

4.2 Estimativa do ganho de aprendizagem 25

4.3 Extração das Características 25

4.3.1 Classificadores de micro ações 29

4.3.2 Arquitetura dos Classificadores de Micro Ação 31

4.3.3 Sumarização dos Resultados 33

5 Experimentos **35**

5.1 Coleta de dados 35

5.2 Análise dos testes de aprendizagem 36

5.3 Classificadores de Micro Ações 37

5.3.1 Visualização de dados de micro ações 39

5.3.2 Validação e avaliação dos resultados 45

5.3.3 Resultados por micro ação 48

5.4 Análise de Dados Estatística 59

5.4.1 Normalização dos Resultados no Tempo 62

5.4.2 Produtos de variáveis 64

5.4.3 Grupo R 65

5.4.4 Grupo NR 66

5.4.5 Discussão 67

5.5 Limitações e desafios da detecção de micro ações faciais 68

6 Conclusão e Trabalhos Futuros **69**

Referências bibliográficas **72**

Capítulo 1

Introdução

1.1 Contextualização

Diante das rápidas transformações tecnológicas e das evoluções no campo educacional, tem-se observado uma crescente adoção de aulas ministradas de forma remota e/ou pré-gravadas, impulsionada principalmente pela disseminação da pandemia de Covid-19. Nesse contexto, a incorporação de recursos tecnológicos e multimídia para otimizar o processo de aprendizado tornou-se mais presente. Esse fenômeno acelerou o processo de digitalização do ensino e consequentemente aumentou a aplicação de conteúdo multimídia pelos docentes na elaboração de suas aulas, para a adaptação a esse novo formato (Martins, 2020).

Durante o processo de aprendizado que envolve recursos de multimídia, a assimilação do conteúdo pelo estudante tende a decair à medida que a quantidade de informações transmitidas aumenta, indicando uma sobrecarga cognitiva em um ou em ambos os canais visual e verbal. Essa afirmação encontra respaldo na Teoria da Carga Cognitiva (CLT), proposta por Sweller (1988), que sugere que nossa memória é capaz de absorver apenas uma pequena quantidade de informações ao mesmo tempo. Dessa forma, o excesso de informações pode dificultar a absorção do conteúdo transmitido pelo interlocutor.

A compreensão aprofundada das complexidades subjacentes ao processo de aprendizagem tem sido alvo de estudo em diversas áreas da psicologia educacional. Nesse contexto, dois trabalhos de relevância notável emergem como pilares teóricos fundamentais. O primeiro, delineado no livro "Cognitive Load Theory" por Paas, Renkl e Sweller (2016), apresenta uma estrutura conceitual ampla conhecida como Teoria da Carga Cognitiva, que examina como a capacidade limitada de processamento da mente impacta a aquisição e assimilação de conhecimento. Em complemento, a pesquisa de Tianlong et al. (2020) examina o elo entre os movimentos oculares e os distintos tipos de carga cognitiva, oferecendo insights cruciais para o entendimento da interação entre processamento de informações e esforço mental.

Em paralelo, a pesquisa liderada por Hussain, Calvo e Chen (2014b) explorou de maneira abrangente o domínio das expressões faciais como um indicador sensível de carga cognitiva e estados emocionais durante a execução de tarefas complexas. Ao investigar a relação entre carga cognitiva, expressões faciais e emoções, o estudo contribuiu para a compreensão da interação dinâmica entre o processamento cognitivo e a expressão emocional. Esses estudos oferecem uma base sólida e complementar para a presente pesquisa, enriquecendo o panorama da interconexão entre processos cognitivos, expressões faciais e respostas emocionais, e fornecendo um arcabouço teórico e metodológico valioso e um enriquecedor ponto de partida para a presente pesquisa.

Com base nessa suposição, no âmbito da Visão Computacional, objetivamos neste trabalho avaliar a possibilidade de reconhecer a carga cognitiva por meio dos movimentos e características faciais, além do rastreamento da íris. Isso possibilitaria a análise dos variados recursos tecnológicos e multimídia de aprendizagem, visando melhorar a absorção de conteúdos pelos alunos.

Portanto, propomos neste estudo investigar as possíveis relações entre características visuais da face dos alunos e a carga cognitiva. Com este objetivo, apresentamos neste trabalho uma metodologia baseada em Visão Computacional e Ciência de Dados para realizar a análise automática da carga cognitiva de estudantes a partir de vídeos faciais de estudantes assistindo aulas com recursos multimidia. Para validar e experimentar a metodologia proposta, e responder as principais questões de pesquisa deste trabalho, utilizamos uma base de vídeos coletada na Universidad de Atacama, Chile.

1.2 Justificativa

Este trabalho se justifica pela relevância do campo de estudo da Teoria da Carga Cognitiva, onde se busca avaliar se as expressões faciais podem estar relacionadas com a eficiência do aprendizado através da análise da carga cognitiva, utilizando vídeos de estudantes em ambientes de aprendizagem multimídia. Destaca-se a importância de tal estudo com o crescimento da aplicação de conteúdo multimídia em formatos síncrono e assíncrono em graduações, pós-graduações e cursos na internet.

Além disso, a proposta deste trabalho é contribuir para o desenvolvimento de algoritmos avançados capazes de extrair de maneira automatizada informações sobre a carga cognitiva dos alunos. Isso permitirá uma avaliação objetiva da eficácia de diversos tipos de recursos pedagógicos multimídia, bem como possibilitará o monitoramento contínuo do processo de aprendizagem dos alunos ao longo do tempo.

A abordagem central desta pesquisa busca estabelecer uma prova de conceito que demonstra a viabilidade de identificar tanto a carga cognitiva quanto a atenção do aluno por meio da

aplicação de algoritmos de Visão Computacional focados na extração de movimentos e características faciais relevantes. Com essa finalidade, propõe-se uma metodologia robusta baseada em Aprendizado Profundo e Ciência de Dados, a fim de aprofundar a compreensão da dinâmica que subjaz à manifestação das expressões faciais em relação à carga cognitiva.

O cerne desse tema reside na intersecção com a linha de pesquisa em Computação Visual e Inteligência Artificial, com um enfoque particular na utilização de técnicas de Aprendizado Profundo e sua aplicação em Visão Computacional. A solução proposta, ancorada nesses princípios, busca abordar de forma inovadora e substancial a problemática de pesquisa delineada, contribuindo assim para o avanço da compreensão sobre as complexas relações entre cognição, expressões faciais e eficácia da aprendizagem em ambientes multimídia.

1.3 Objetivo Geral e Objetivos Específicos

O objetivo geral desta proposta é responder à seguinte questão de pesquisa: "É possível empregar redes neurais profundas e análise estatística de dados para identificar automaticamente micro ações em vídeos faciais de sessões de aprendizado multimídia?". Caso essa hipótese se confirme, pretendemos também investigar: i) se existem padrões relevantes relacionados a ocorrências de micro ações e o ganho no aprendizado; ii) e se existem diferentes padrões de ocorrências de micro ações para diferentes tipos de aulas de aprendizado multimídia.

Para atingir este objetivo geral, os seguintes objetivos específicos foram definidos:

a) Organizar e processar uma base de vídeos pré-existentes de estudantes em ambientes de aprendizagem multimídia.

b) Desenvolver um pipeline automatizado para a extração de características visuais da face e movimentos da íris em vídeos.

c) Utilizar modelos baseados em Aprendizado Profundo para extrair com precisão a posição de diversas características visuais da face, a partir de uma imagem ou vídeo RGB.

d) Aplicar técnicas de aprendizado não supervisionado para o agrupamento de expressões faciais e outros padrões, e para a análise de semelhanças entre os grupos.

e) Aplicar técnicas de aprendizado supervisionado para selecionar características relevantes e realizar a predição da carga cognitiva dos estudantes a partir das características visuais.

f) Avaliar e comparar os resultados, visando responder à pergunta de pesquisa principal.

1.4 Estrutura do trabalho

O trabalho será dividido de tal forma que, em cada um dos seguintes capítulos, abordaremos cada aspecto da pesquisa:

No capítulo 2 serão apresentados os estudos existentes na literatura acerca dos temas propostos no trabalho. Serão verificadas as abordagens utilizadas por outros autores para os problemas tratados nesta dissertação, apontando os pontos fortes, que evidenciam a importância do estudo, e os pontos fracos, apontando onde há espaço a ser explorado.

No capítulo 3 mostraremos os principais conceitos indispensáveis ao desenvolvimento desse trabalho, bem como suas eventuais aplicações, referendando as tecnologias escolhidas para este problema em questão.

No capítulo 4 daremos uma estruturação ao método adotado na pesquisa, detalhando como se dará cada etapa deste estudo.

No capítulo 5 veremos a realização dos experimentos e coleta de seus resultados alcançados, comparando abordagens e tecnologias, e iremos nos aprofundar nos resultados obtidos durante o experimento, buscando pontos relevantes para validar a prova de conceito proposta no trabalho.

No capítulo 6, por fim, serão apresentadas as observações e conclusões finais acerca do estudo e seus resultados atingidos, com possibilidades de aplicações futuras das tecnologias analisadas.

Capítulo 2

Referencial Teórico

2.1 Teoria da carga cognitiva

Os psicólogos Sweller, Ayres e Kalyuga (2011) observaram através de suas pesquisas que o aprendizado torna-se mais efetivo quando os recursos informativos estão alinhados com o processo cognitivo humano, isto é, quando a carga de informações é compatível com a capacidade cognitiva do indivíduo.

Esses estudos foram replicados por diversos pesquisadores, que buscaram criar um ambiente de aprendizagem adequado. De acordo com Santos e Tarouco (2015), um ambiente de aprendizagem adequado, segundo os princípios da Teoria da Carga Cognitiva, minimiza o uso de recursos mentais desnecessários, e, em vez disso, emprega-os de maneira a maximizar a aprendizagem. Por isso, a Teoria da Carga Cognitiva tem recebido atenção crescente nos campos da pesquisa científica, que buscam interconectar diversas áreas para criar o ambiente de aprendizagem ideal.

Na estrutura cognitiva humana para aprendizagem, destacam-se os processos de armazenamento e consolidação da informação. Esses processos podem ser divididos em três etapas: aquisição, consolidação e evocação. O processamento e consolidação acontecem na memória, passando pela memória sensorial, de curta duração (ou de trabalho) e de longa duração.

Segundo a Teoria da Carga Cognitiva, a elaboração de materiais didáticos, especialmente aqueles que utilizam recursos multimídia, deve seguir padrões e princípios específicos para reduzir a sobrecarga cognitiva do estudante e potencializar o seu aprendizado (Mayer, 2009).

Atualmente, existem três técnicas diferentes para medir a carga cognitiva: autoavaliação através de questionários; coleta de medidas fisiológicas; e o método das tarefas secundárias. Segundo Weller, Merriënboer e Paas (2019), a técnica mais utilizada é a autoavaliação.

Hoje em dia, a medição da Carga Cognitiva vem sendo realizada através de imagens

cerebrais obtidas por meio da tomografia por emissão de pósitrons (PET) e também por meio da ressonância magnética (Murata, 2005).

Bruken defende que as medidas fisiológicas, como a medição do ritmo cardíaco e da dilatação da pupila, não são medidas diretas da Carga Cognitiva. Elas possuem apenas uma correlação indireta com a Carga Cognitiva. Por exemplo, uma alta carga cognitiva pode levar a um alto nível de estresse em um indivíduo, o que pode causar mudanças no ritmo cardíaco - uma resposta emocional do indivíduo ao material de ensino (Brunken, Plass e Leutner, 2003).

Contrariamente a essas medidas fisiológicas indiretas, as medidas obtidas por meio de imagens neurológicas são diretas, pois refletem diretamente a atividade cerebral. No entanto, estas são mais caras e com menor acessibilidade e possibilidade de escala. Nesse sentido, a análise das expressões faciais se torna bastante interessante para avaliar a carga cognitiva, pois permite uma maior escalabilidade da solução e tem um custo de implementação menor.

Vários autores já investigaram a relação entre as expressões faciais e a carga cognitiva, obtendo resultados positivos. O primeiro trabalho nesse sentido foi desenvolvido por Hussain, Calvo e Chen (2014a), com outros se focando na análise da relação entre a carga cognitiva e as emoções básicas, como raiva, alegria, etc., como no estudo de Zhou, Ghose e Lukowicz (2020).

2.1.1 Teoria da carga cognitiva e o aprendizado multimídia

Já em 1956, George Miller referia-se a um "número mágico" 7-2 ou 7+2: o sistema cognitivo humano consegue processar somente um número limitado de informações, que variam entre 5 a 9 elementos simultaneamente. Quando esses limites são excedidos, o raciocínio e a aprendizagem diminuem em desempenho, sobrecarregando a estrutura cognitiva.

Mayer (2009) corroborou esse fato em seu estudo utilizando aplicações multimídia, onde verificou que quanto mais canais de percepção simultâneos eram utilizados, maior era a desorientação e a falta de estímulo do aluno. A partir disso, formulou três pressupostos: h

1. O ser humano possui dois canais distintos para o processamento da informação: o canal visual e o canal verbal;
2. Existe uma capacidade limitada de processamento de informação em cada canal;
3. A aprendizagem requer um processamento cognitivo essencial em ambos os canais.

O processamento ativo, de acordo com o terceiro pressuposto de Mayer (2009), é composto por cinco processos cognitivos, representados no infográfico da figura 2.1:

Assim formando segundo Mayer (2009) o processo cognitivo, que podemos analisar com a figura 2.2 abaixo:

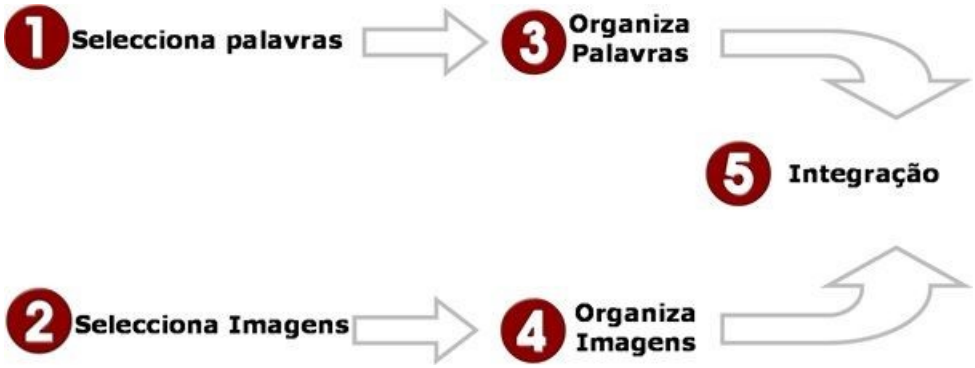


Figura 2.1: Tipos de Processamento Cognitivo Adaptado de Mayer (2009).

Figura 2.2: Fluxograma do Processo Cognitivo Adaptado de Mayer (2009).

Em seus estudos, Mayer (2009), também define três tipos de processamento que são definidos como:

- Essencial: Direcionado para o "fazer sentido" do material apresentado. Ou seja, o selecionar, organizar e integrar palavras e imagens.
- Incidental: Direcionado para os aspectos não essenciais do material.
- Exploração Representacional: Tendente a explorar as representações verbais e visuais na memória de trabalho/curto prazo.

Conforme Mayer (2009), podemos segregar as medidas da carga cognitiva em:

- Carga Intrínseca: inerente a complexidade do conteúdo do material de ensino;
- Carga Extrínseca ou irrelevante: não interfere na construção de esquemas e desperdiça recursos mentais limitados que poderiam ser usados para a auxiliar a carga natural;
- Carga Natural ou relevante: é a carga necessária para promover a aprendizagem.

Dentro desses processamentos, temos alguns cenários que reduzem a eficiência do aprendizado devido ao aumento da carga cognitiva. Podemos sintetizar em 3 tipos de sobrecarga:

- Sobrecarga de informação essencial num canal: a informação excede a capacidade cognitiva do canal visual.
- Sobrecarga de informação essencial em ambos os canais: material conceptualmente complexo excede a capacidade cognitiva em ambos os canais.
- Sobrecarga por combinação de informação essencial com incidental: quando a capacidade cognitiva é sobrecarregada pela adição de material incidental (supérfluo ou redundante).

Dessa forma, o modelo exibido na Figura 2.3 foi apresentado para o desenvolvimento de um ambiente de aprendizagem de boa qualidade, visando permitir um equilíbrio entre as cargas cognitivas.

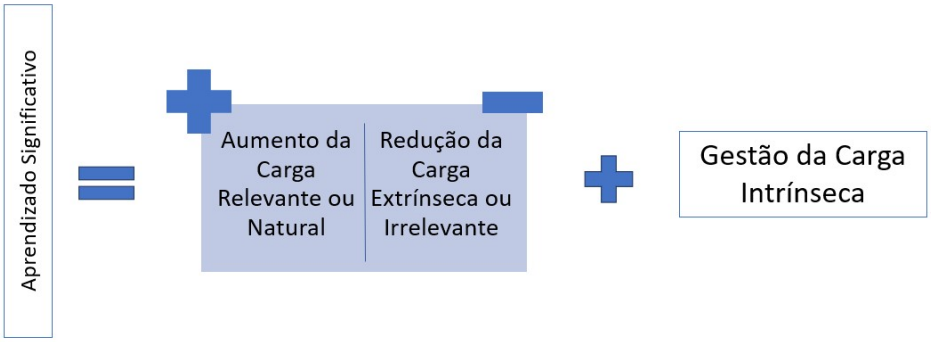


Figura 2.3: Equilíbrio da Carga Cognitiva. Adaptado de Mayer (2009).

Mayer (2014), em continuidade ao estudos de Sweller, elaborou os seguintes princípios a serem seguidos para minimizar a carga cognitiva:

- Princípio de Representação Múltipla: os alunos aprendem melhor quando se combinam palavras e imagens, do que quando se usam somente palavras.
- Princípio de Proximidade Espacial: esse princípio diz respeito à proximidade de palavras e imagens, ou seja, é quando palavras e imagens correspondentes estão próximas em vez de afastadas.
- Princípio da Não Divisão ou da Proximidade Temporal: nesse princípio tem se a apresentação de palavras e imagens simultaneamente em vez de sucessivamente, uma vez que a apresentação de um texto e de uma animação na mesma tela divide a atenção do aluno.
- Princípio das Diferenças individuais: sabe-se que estudantes com maior nível de conhecimento, sobre um determinado assunto e com grau maior de orientação espacial possuem maiores condições de organizar e processar seu próprio conhecimento ao interagir com o assunto.

- Princípio da Coerência: refere-se à exclusão de palavras, imagens ou sons não relevantes para o assunto. Quanto mais simples e objetiva for a apresentação do conteúdo, mais livre ficará a memória de trabalho para processar um número maior de conhecimentos.
- Princípio da Redundância: nesse princípio, ressalta-se que o uso da animação e narração, quando usadas simultaneamente no processo de ensino, potencializa o conhecimento, diferente de quando usadas separadamente.

2.2 Expressões faciais e o estado psicológico

Segundo Ekman e Friesen (1977), expressões faciais são consideradas a maneira mais indicativa e eficiente para reconhecer as emoções e o estado psicológico durante o processo de comunicação. Neste mesmo estudo, os autores afirmam que nem todas as expressões faciais retratam alguma alteração no estado emocional. Assim, investigaremos neste trabalho o uso dessas movimentações e expressões faciais para identificar se há a possibilidade de avaliar de forma acurada a carga cognitiva dos alunos durante o processo de aprendizagem.

Essas movimentações e expressões faciais foram estudadas profundamente pela psicologia e pela neurociência. Um dos primeiros trabalhos que analisou as expressões faciais foi o de Darwin (1871), onde foi estudada a semelhanças e diferenças entre a forma que humanos e animais expressam suas emoções onde o pesquisador chegou à conclusão que: “os jovens e os velhos de diferentes raças, independente de homem ou animal, expressam o mesmo estado de espírito com os mesmos movimentos”.

Após esse estudo, a comunidade científica despertou o interesse no tema das emoções por cientistas, onde damos destaque ao estudo de Ekman e Friesen (1977), onde foi defendido a existência de seis emoções base, que eram reconhecidas e produzidas sem depender do contexto.

Esse trabalho foi o ponto de partida para a entrada e surgimento de metodologias como a Facial Action Coding System - FACS publicado por Paul Ekman e Wallace V. Friesen, 1978. Estes pesquisadores criaram um sistema composto por códigos denominados como Action Units (AUs), Action Descriptors (ADs), Moviments, Gross Behaviors Codes e Visibility Codes. Outros dois sistemas que também são muito utilizados são o da análise de expressões emocionais espontâneas e a análise da dinâmica temporal destas expressões (Lucey, Lucey e Cohn, 2010).

2.3 Aprendizado de Máquina

Dentro dos ramos da inteligência artificial se encontra o aprendizado de máquina, onde Samuel (1959), definiu o aprendizado de máquina como “um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados para tal”.

Já para Goodfellow, Bengio e Courville (2016), o Aprendizado de Máquina é uma forma de estatística aplicada com ênfase no uso de computadores para estimar funções complicadas, e diminuir a ênfase em provar intervalos de confiança ao redor dessas funções. Faceli et al. (2021), define o aprendizado de máquina como dispositivos computacionais que são programados para aprender a partir de experiências passadas. Para tal, frequentemente empregam um princípio de inferência denominado indução, que permite extrair conclusões genéricas a partir de um conjunto particular de exemplos.

Em resumo, o aprendizado de máquina, de maneira simplificada, envolve o processo de capacitar um sistema computacional para executar tarefas específicas, utilizando novos dados, com base em conhecimentos e padrões adquiridos durante o aprendizado e treinamento prévios, muitas vezes empregando princípios de indução. Este processo é habilitado por técnicas variadas, como aprendizado supervisionado, não supervisionado e de reforço, conforme elucidado por Faceli et al. (2021).

2.3.1 Conjunto de Dados

Um conjunto de dados consiste em uma amostra que foi coletada a partir de observações. Estes conjuntos podem ser classificados como conjuntos de dados estruturados e não estruturados.

Dados Estruturados

Faceli et al. (2021), define que os dados estruturados são um conjunto de informações em formato tabular, uma matriz atributo-valor, em que cada linha representa um objeto (instância ou exemplo) e cada coluna representa um atributo (característica ou variável). Os atributos podem ser divididos em atributos preditivos, cujos valores descrevem características dos objetos, que formam um vetor de entrada, e atributo alvo, cujo valor rotula o objeto, com uma classe ou valor numérico. Essas denominações têm por origem o frequente uso dos valores dos atributos preditivos de um objeto para predizer o valor de seu atributo alvo.

Dados não Estruturados

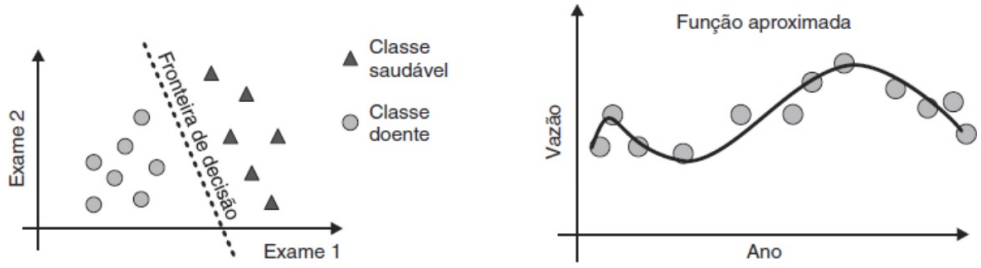
Outra forma de dados, são os não estruturados que, diferente do estruturados, representam informação de forma implícita, sem uma estrutura/padrão definido. Exemplos comuns deste tipo de dados são textos, imagens, vídeos e áudios (Faceli et al., 2021).

2.3.2 Formas de aprendizado de máquina

O aprendizado supervisionado é um paradigma em aprendizado de máquina no qual os algoritmos aprendem uma função que mapeiam dados em rótulos, ou seja, aprendem com um conjunto de dados previamente rotulados (Ghotra, McIntosh e Hassan, 2015). Matematicamente, isso pode ser representado como $y = f(x)$, onde y é a saída desejada, x é a entrada e f é a função que o algoritmo aprende a aproximar com base nos dados de treinamento.

De acordo com as colocações de Faceli et al. (2021), é possível conceituar os problemas de classificação como aqueles que envolvem conjuntos de dados previamente rotulados em uma ou mais categorias. Nesse contexto, os algoritmos de aprendizado de máquina têm como objetivo explorar os dados de treinamento, os quais já possuem rótulos atribuídos, a fim de identificar a fronteira de decisão que permitirá classificar de maneira eficaz novos dados submetidos ao modelo. A partir desse processo, ocorre a demarcação das diferentes categorias, conforme ilustrado na Figura 2.4a.

Já os problemas de regressão são um conjunto vasto de técnicas estatísticas usadas para modelar relações entre variáveis e predizer o valor numérico de uma ou mais variáveis dependentes (ou de resposta) a partir de um conjunto de variáveis independentes (Figura 2.4b).



(a) [Classificação] (b) [Regressão]

Figura 2.4: Aprendizado Supervisionado e Não Supervisionado Faceli et al. (2021).

Porém, quando as amostras não possuem rótulos, ainda é possível aplicar algoritmos de aprendizado não supervisionado. Estes algoritmos aprendem padrões significantes nos dados para detectar semelhança entre os dados e agrupá-los por algum padrão de similaridade que o algoritmo aprendeu através do treinamento.

Temos uma ampla variedade de algoritmos à disposição para o aprendizado não supervisionado, cada um com sua aplicação específica e abordagem única. Alguns dos algoritmos mais amplamente utilizados merecem destaque, como o K-means e o Hierarchical Clustering Murphy (2012), que é frequentemente empregado para tarefas de agrupamento, auxiliando na identificação de padrões intrínsecos nos dados. Além disso, o uso de técnicas como os Autoencoders Goodfellow, Bengio e Courville (2016) oferece uma abordagem valiosa para a redução de dimensionalidade, permitindo a construção de representações compactas dos dados.

Para tratar a maldição da dimensionalidade, também encontramos o uso de métodos como

o PCA (Principal Component Analysis) e o Kernel-PCA (Murphy, 2012). Essas abordagens visam a redução da dimensionalidade, preservando a estrutura relevante dos dados, o que pode ser especialmente útil em tarefas de visualização ou pré-processamento para aprendizado subsequente.

Por fim, vale mencionar também a existência dos algoritmos de aprendizado por reforço, onde um agente (robô) trabalha em prol de atingir um objetivo explícito, fazendo interações com seu ambiente e mudando o estado atual do ambiente através de escolha de ações e recebendo uma recompensa para mudança de estado (Mohri, Rostamizadeh e Talwalkar, 2018).

2.4 Visão Computacional

A Visão Computacional está para a imagem assim como o Processamento de Linguagem Natural está para o texto (Chollet, 2017). Conforme Szeliski (2022), podemos definir Visão Computacional como sendo um conjunto de algoritmos através dos quais sistemas baseados em computadores podem extrair informações dos pixels que compõem a imagem.

Uma definição alternativa de Visão Computacional é dada pelo processo de descobrir, por meio de imagens, o que está presente no mundo e onde se está localizado, ou seja, uma tarefa de processamento de informação através de imagens (Marr, 1982).

Já na visão de Crowley e Christensen (1994), Visão Computacional é a área da ciência que estuda e procura desenvolver teorias e métodos voltados à extração automática de informações úteis contidas em imagens. Tais imagens são capturadas por dispositivos como câmera de vídeo e scanner.

Portanto, os processos de Visão Computacional são formados por um conjunto de métodos e técnicas através dos quais sistemas computacionais tentam ser capazes de interpretar imagens, muitas vezes realizando tarefas de Processamento de Imagens.

Assim, as tarefas de Visão Computacional incluem métodos para adquirir, processar, analisar e entender imagens digitais e extrair dados de alta dimensão do mundo real para produzir informações numéricas ou simbólicas, por exemplo, nas formas de decisões. Ou seja, como define Szeliski (2022), compreender e descrever o mundo visto em uma ou mais imagens e reconstruir suas propriedades como forma, iluminação e distribuição de cores.

Os dados coletados a partir das imagens podem assumir várias formas, como sequências de vídeo, visualização de várias câmeras, dados multidimensionais de um scanner 3D ou dispositivo de digitalização médica. Dados que possibilitam a utilização do Aprendizado de Máquina e do Aprendizado Profundo para a resolução de problemas do cotidiano.

2.5 Redes Neurais Artificiais

Na constante busca por simular o processo de aprendizado, pesquisadores como McCulloch e Pitts (1943) se inspiraram no cérebro humano. Mais precisamente, se basearam no funcionamento dos neurônios e na forma em que os mesmos trocam informações por meio de sinapses, sua forma de conexão e a forma da atuação em paralelo. Assim seguiu-se pela busca de um sistema computacional cuja capacidade de processamento se aproximasse da do cérebro humano, motivando o desenvolvimento dos métodos conexionistas de aprendizado de máquina, as chamadas Redes Neurais Artificiais (RNAs) Faceli et al. (2021).

Em seu livro, Faceli et al. (2021) define as Redes Neurais Artificiais como sistemas computacionais distribuídos compostos de unidades de processamento simples, densamente interconectadas. Essas unidades, conhecidas como neurônios artificiais, computam funções matemáticas.

A estrutura do neurônio artificial é composto por uma ou mais camadas interligadas por um grande número de conexões. Na maioria das arquiteturas, essas conexões, que simulam as sinapses biológicas, possuem pesos associados que ponderam a entrada recebida por cada neurônio da rede. Uma RNA é, portanto, caracterizada por dois aspectos básicos: arquitetura e aprendizado. Enquanto a arquitetura está relacionada com o tipo, o número de unidades de processamento e a forma como os neurônios estão conectados (Figura 2.5), o aprendizado diz respeito às regras utilizadas para ajustar os pesos da rede e à informação que é utilizada por essas regras (Faceli et al., 2021).

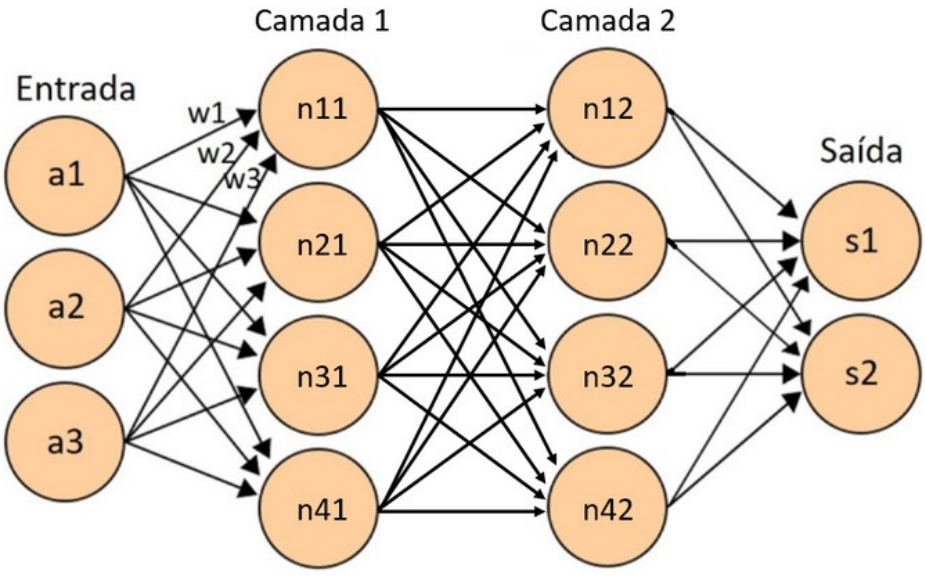


Figura 2.5: Exemplo de Rede Neural Artificial. Fonte: Faceli et al. (2021).

Em uma RNA, os neurônios podem ser estruturados em uma ou mais camadas. Quando

duas ou mais camadas são utilizadas, um neurônio pode receber em seus terminais de entrada valores de saída de neurônios da camada anterior e/ou enviar seu valor de saída para terminais de entrada de neurônios da camada seguinte. A Figura 2.5, ilustra um exemplo de RNA com duas camadas. Essa rede recebe como entrada valores de três atributos de entrada e gera dois valores em sua saída.

Faceli et al. (2021) define que as redes que possuem mais de uma camada de neurônios são nomeadas como multicamadas, sendo compostas pelas camadas de entrada, intermediárias ou escondidas e as de saída. As redes neurais que operam com uma ou duas camadas ocultas conectadas entre si são conhecidas como redes neurais rasas (shallow neural networks). Já as redes neurais de aprendizagem profunda (deep neural networks), podem ter até dezenas de camadas ocultas. Em uma rede multicamadas, as conexões entre os neurônios podem apresentar diferentes padrões de conexão. De acordo com esses padrões, a rede pode ser classificada em:

- Completamente conectada: quando os neurônios da rede estão conectados a todos os neurônios da camada anterior e/ou seguinte.
- Parcialmente conectada: quando os neurônios estão conectados a apenas alguns dos neurônios da camada anterior e/ou seguinte.
- Localmente conectada: são redes parcialmente conectadas, em que os neurônios conectados a um neurônio se encontram em uma região bem definida.

2.5.1 REDES NEURAIS CONVOLUCIONAIS

As redes neurais convolucionais (RNC), ou Convolutional Neural Network (CNN), são arquiteturas de RNA que buscam imitar o funcionamento do córtex visual. Consequentemente são bastante utilizadas para problemas envolvendo imagens e dados não estruturados. Sua arquitetura é diferenciada conforme podemos verificar na Figura 2.6, onde é possível observar que a CNN pode ser formada por uma ou mais camadas de convolução, seguidas por camadas densas.

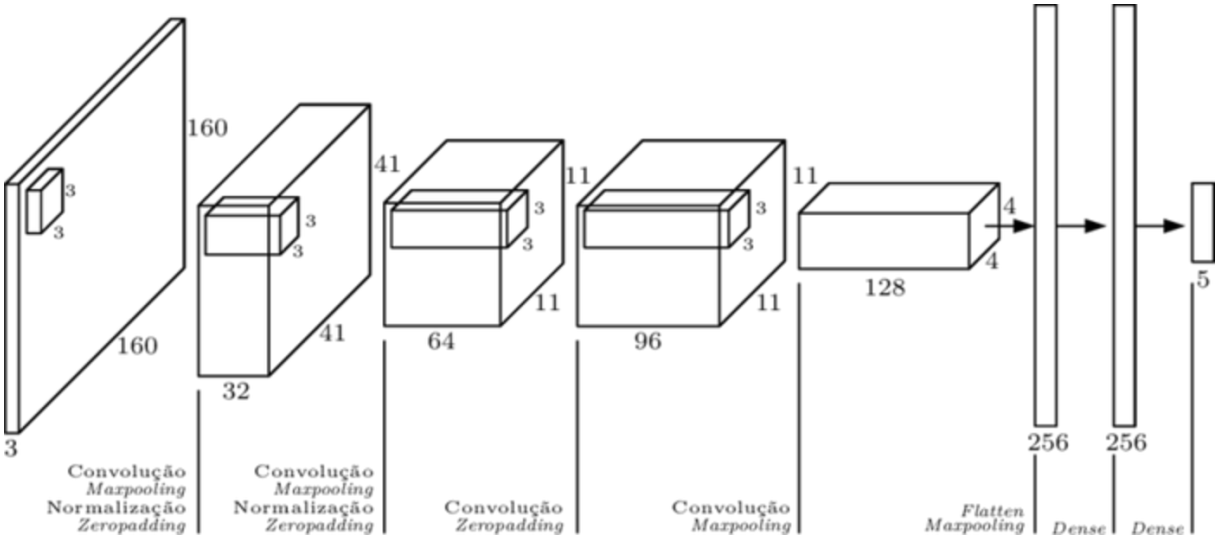


Figura 2.6: Exemplo de Arquitetura de uma Rede CNN. Fonte: Oliveira, Alves e Malqui (2017).

A chamada camada convolucional é composta por um conjunto de filtros ou kernel de convolução. Assim, para cada filtro é aplicado à toda imagem de entrada resultando em um mapa de característica (feature map). Após o processo de convolução, as ativações são passadas para uma segunda camada, a de subamostragem (subsampling), onde são calculados os valores máximo e médio de uma área de tamanho pré-definido, gerando assim uma representação de entrada com a resolução reduzida. O resultado dessa etapa é um outro mapa de características com resolução menor que proporciona invariância a pequenas translações (Lecun et al., 1998). Podemos ver um exemplo de arquitetura CNN para classificação de imagens na Figura 2.7.

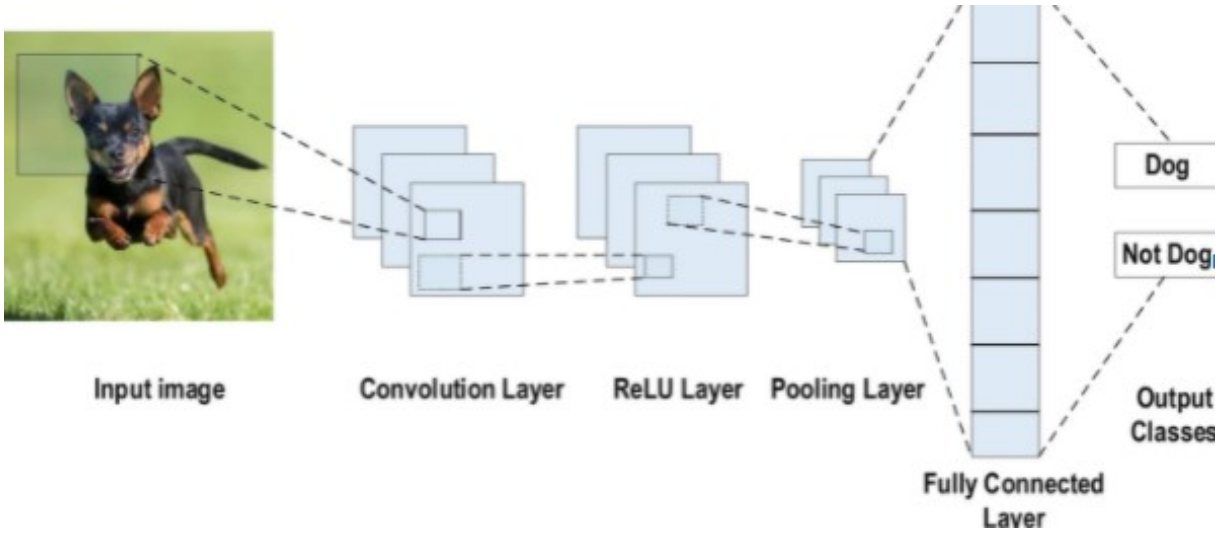


Figura 2.7: Arquitetura de uma CNN. Fonte: Faceli et al. (2021).

2.5.2 MobileNet

Em 2017, o Google lançou o MobileNet, que foi desenvolvido com a proposta de ser uma rede neural profunda e de baixo custo computacional para aplicações móveis e embarcadas de Visão Computacional. Ele foi baseado no conceito de *Depthwise Separable Convolution*, que é uma forma de convolução fatorada. Esse tipo de convolução separa a convolução padrão em duas partes: a convolução em profundidade e a convolução pontual, sendo capaz de reduzir o número de parâmetros, quando comparamos com uma rede neural tradicional, conforme a Figura 2.8.

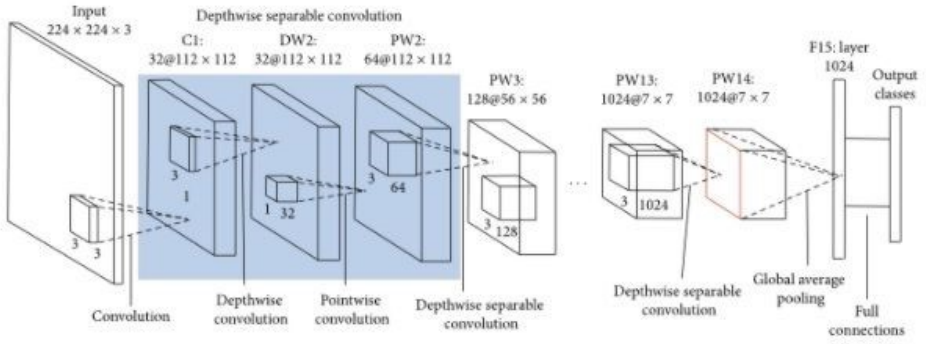


Figura 2.8: Arquitetura da MobileNet com Depthwise Separable Convolution. Adaptado de Sandler et al. (2018).

O processo do Depthwise Separable Convolution, pode ser separado em duas etapas: a de Depthwise convolution e Pointwise convolution. Onde o Depthwise convolution é um convolução espacial DK×DK orientada por canal. Já o Pointwise convolution é uma convolução 1x1 para mudança de dimensão, conforme figura 2.9 abaixo.

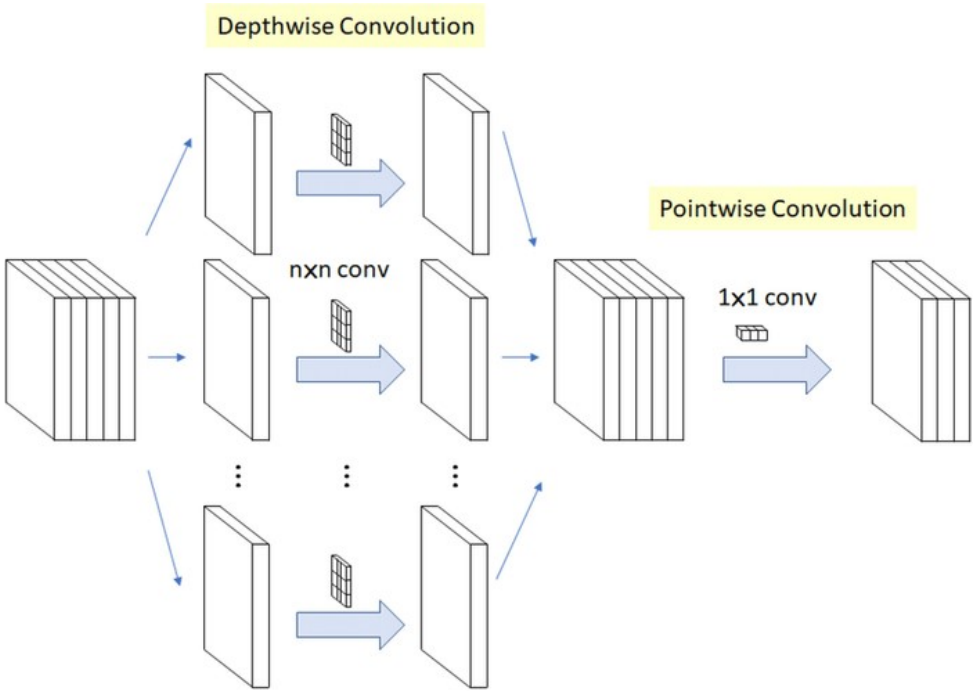


Figura 2.9: Processo de Depthwise Separable Convolution. Adaptado de Sandler et al. (2018).

Em 2018 foi apresentada a MobileNet v2, que apresentou melhorias em relação à arquitetura original. Isto resultou no aumento da velocidade de treinamento e no aumento na acurácia dos modelos, além da redução no número de parâmetros. No ano seguinte, a MobileNet v3 foi apresentada com uma nova arquitetura, trazendo a sua versão "Large" and "Small", que podem ser utilizados em casos de utilização com poucos ou muitos recursos.

2.5.3 Long Short-term Memory - LSTM

O *Long Short Term Memory* (LSTM) é uma rede neural criada para resolver o problema de memória de longo prazo que as RNNs simples possuem, devido ao problema da dissipação do gradiente. A LSTM tem uma memória de longo prazo e possui três tipos de portões: esquecimento, entrada e saída. Cada portão é uma camada de rede neural MLP, com pesos, vieses e uma função de ativação, geralmente a sigmóide. A LSTM também tem a capacidade de reter informações de estados distantes, já que o estado $\boldsymbol{c}_t$ flui livremente entre as células e a influência do *vanishing gradient* é menor. O estado da célula $\boldsymbol{c_t}$ armazena informações das saídas anteriores e entrada atual, funcionando como uma memória. A LSTM tem uma influência menor no estado $\boldsymbol{c}_t$ porque não há múltiplas multiplicações de pesos, comparado à multiplicação entre estado e pesos para uma RNN simples. Os passos de propagação dentro de uma célula LSTM são mostrados na Figura 2.10, e nas Equações 2.1 e 2.2).

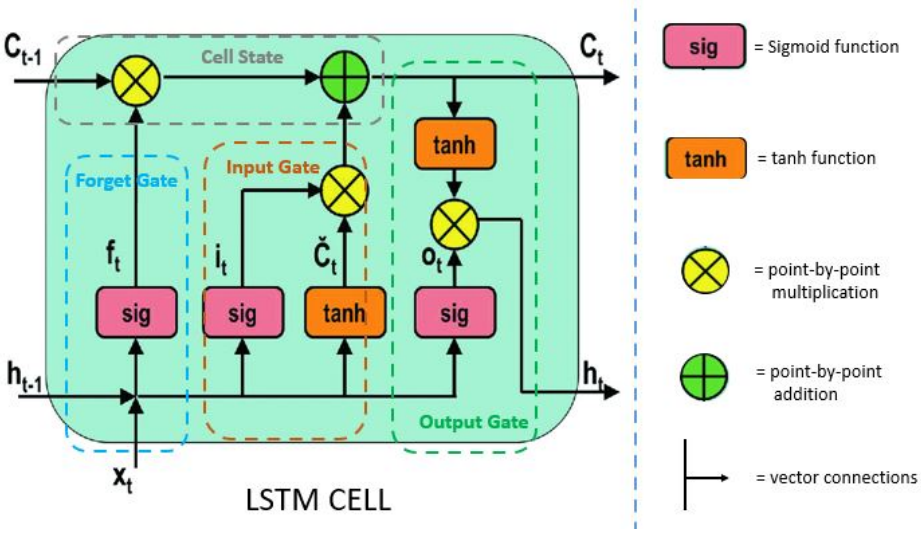


Figura 2.10: Exemplo de LSTM - Smagulova e James (2020)

.

$$f_t = \sigma(W_t^f \cdot [y_{t-1}, x_t] + b_f) \tag{2.1}$$

$$y_t = tanh(c_t) * a_t \tag{2.2}$$

O portão de entrada é calculado pela Equação 2.1, com o valor produzido pelo portão denotado por f_t, os pesos utilizados no portão de esquecimento denotados por W_t^f e os vieses por b_f. O portão de esquecimento calcula a influência da saída anterior $yt-1$ e da entrada atual x_t no estado da célula C_t. A operação $[yt-1, x_t]$ é uma concatenação de vetores. Em seguida, o valor do portão de entrada i_t é obtido pela Equação 2.3, com os pesos desse portão sendo a matriz W_t^i e vieses b_i. A matriz $\hat{C}_t$ é computada pela Equação 2.4, com os pesos W_t^c e vieses b_c que representam os pesos e vieses respectivamente para o estado da célula atual. A matriz $\hat{C}_t$ representa um possível estado novo da célula. O estado da célula LSTM C_t é descrito pela Equação 2.5.

$$i_t = \sigma(W_t^i \cdot [h_{t-1}, x_t] + b_i) \tag{2.3}$$

$$\hat{c}_t = tanh(W_t^c \cdot [h_{t-1}, x_t] + b_c) \tag{2.4}$$

$$c_t = f_t * c_{t-1} \bigoplus i_t \times \hat{c}_t \tag{2.5}$$

$$a_t = \sigma(W_t^o \cdot [h_{t-1}, x_t] + b_a) \tag{2.6}$$

A arquitetura Long Short-term Memory(LSTM) consiste de neurônios que processam dados sequenciais, tornando-a ideal para lidar com reconhecimento de textos. Tais células efetuam operações em cima das informações passadas e são capazes de guardar dados ao longo do tempo, evitando a perda de informações importantes para o processo de aprendizado. A quantidade de células da LSTM é um dos parâmetros que pode ser ajustado ao longo do treinamento, variando o poder de aprendizado e abstração do modelo.

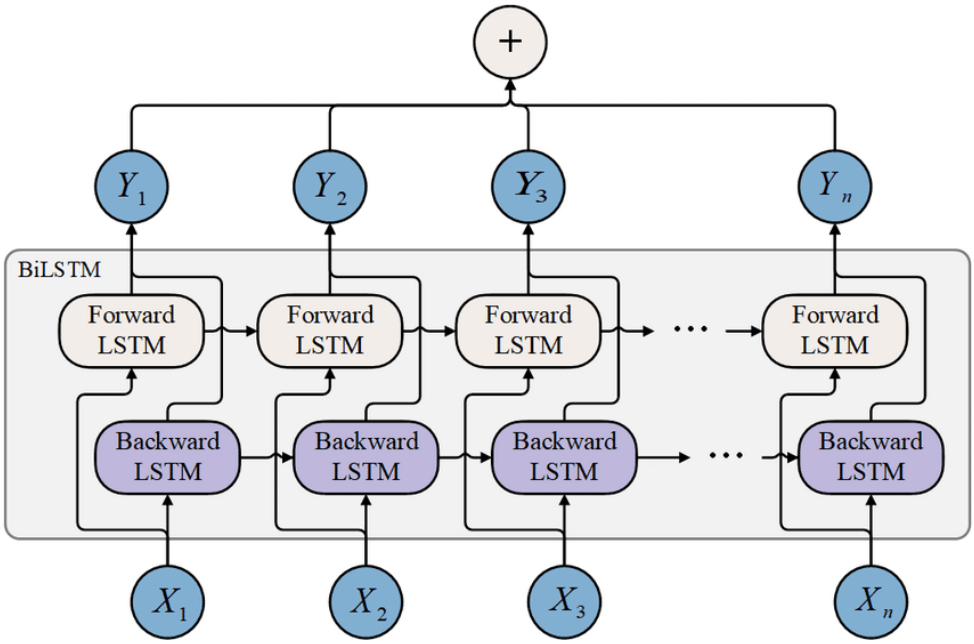


Figura 2.11: Exemplo de BILSTM - Mao et al. (2022).

A BiLSTM (Bidirectional Long Short-Term Memory) é uma arquitetura de rede neural recorrente que se destaca por sua capacidade de processar sequências de dados em duas direções simultaneamente, capturando tanto informações contextuais anteriores quanto posteriores em uma sequência. Ela é uma extensão da LSTM (Long Short-Term Memory), que é projetada para evitar o problema de dissipação do gradiente em sequências longas. A BiLSTM tem sido amplamente utilizada em tarefas de Processamento de Linguagem Natural, devido à sua habilidade de capturar relações complexas em dados sequenciais. O funcionamento da BiLSTM envolve a passagem da sequência de entrada através de duas camadas LSTM independentes, uma processando a sequência no sentido direto e outra no sentido reverso. A saída final combina as informações das duas direções, resultando em uma representação contextual mais rica e abrangente. Referências como Schuster e Paliwal (1997), Graves e Schmidhuber (2005), detalham a arquitetura e os benefícios da BiLSTM no contexto de tarefas sequenciais como podemos ver na Figura 2.11.

Capítulo 3

Trabalhos Relacionados

Nesta sessão apresentamos trabalhos relacionados a esta pesquisa, organizados da seguinte forma. Serão abordados trabalhos relacionados à Visão Computacional, como detecção facial, extração de características da face e reconhecimento de expressões faciais; e trabalhos na linha da Teoria de Carga Cognitiva e suas perspectivas.

3.1 Detecção Facial

O processo de detecção facial é a tarefa da localização de faces humanas em imagens. Dada uma imagem qualquer, o objetivo do algoritmo é determinar se há faces na imagem e caso sejam encontradas retornar a sua localização (Yang, Kriegman e Ahuja, 2002). Esse procedimento que é facilmente executado por nós humanos, no âmbito da Visão Computacional é um desafio devido às inúmeras complexidades da tarefa. Isto inclui fatores ocorridos na aquisição da imagem, como iluminação, e fatores locais como cores similares entre objetos, exibição parcial da face por oclusão, variações de expressão facial e outros fatores.

Yang, Kriegman e Ahuja (2002), dividiu as técnicas de detecção em técnicas baseadas em características e baseadas em modelo. As primeiras tem como foco localizar um conjunto de características faciais na imagem, como olhos, boca e nariz. A partir da localização desses pontos de interesse, verifica-se se estes estão em um arranjo geométrico plausível para concluir se existe ou não uma face (Szeliski, 2022).

Na abordagem baseada em modelos, tenta-se aprender modelos a partir de conjuntos de dados de faces, buscando características relevantes que possam distinguir rostos em imagens (Szeliski, 2022).

Podemos ressaltar alguns algoritmos que se destacaram nessa tarefa, como o Viola e Jones (2001), queestá disponível em bibliotecas como a OpenCV. Outras abordagens utilizam redes neurais profundas como a Multi-Task Cascaded Convolutional Neural Network (MTCNN).

Essa arquitetura usa uma estrutura em cascata com três redes, onde o primeiro passo é o redimensionamento da imagem para diferentes tamanhos. O primeiro modelo propõe áreas que podem ser características da face. A segunda rede, por sua vez, filtra as marcações que foram feitas na primeira rede e por fim a terceira e última rede propõe as landmarks faciais (Zhang e al., 2016).

3.2 Reconhecimento de pontos de interesse da face

Diferentemente da detecção facial, o reconhecimento de pontos de interesse da face tem como objetivo a detecção e o rastreamento de pontos de interesse (ou *landmarks*) em faces humanas. Atualmente não há uma definição formal para os pontos de interesse das faces, porém comumente são determinados pelas características do rosto como boca, nariz, olhos, sobrancelhas e outros pontos de interesse (Hamid, Mohammed e Aksasse, 2016).

Segundo Wu e Ji (2016), os métodos utilizados para o reconhecimento das pontos de interesse podem ser divididos em três grupos principais: métodos holísticos, métodos baseados em regressão, e métodos locais com restrições (Constrained Local Models). Cada grupo utiliza um conjunto de técnicas para a extração das características faciais, o método holístico utiliza modelos para representar a aparência e formato da face de forma global. Os métodos baseados em regressão buscam capturar implicitamente o formato e aparência da face. E por fim o CLM, busca características de forma global porém cria o modelo de formato da face de forma local (Wu e Ji, 2016).

Trabalhos como o de Pantic e Rothkrantz (2000) buscaram aprofundar o reconhecimento das expressões faciais, como também o trabalho de Murphy-Chutorian e Trivedi (2009), o qual buscava criar um algoritmo para estimar a posição da pose da face, foram fortemente baseados na extração de características da face.

O processo de extração de características também se estende aos pontos ao redor dos olhos, que foram utilizados no trabalhos de Hansen e Ji (2010), para a detecção e rastreamento dos olhos.

3.3 Reconhecimento de Expressões Faciais

Revina e Emmanuel (2021) fazem um estudo comparativo entre os modelos de reconhecimento de expressões, tais modelos atualmente são baseados em extração de características e vem apresentando melhorias na precisão da classificação das expressões e sentimentos, com os avanços nos métodos de seleção de características e atributos que segundo os autores é a parte fundamental para a criação dos modelos. Os autores afirmam que o rosto oferece três tipos

de sinal que definem como estático, rápido e lento. Os primeiros englobam aspectos da face, como pigmentação, formato da face, construção dos ossos, cartilagem, localização e tamanho de características faciais como sobrancelhas, olhos, nariz e boca. Os sinais rápidos são, por exemplo, a elevação das sobrancelhas e os movimentos dos músculos faciais, onde esses movimentos duram apenas alguns milissegundos. Os lentos englobam as mudanças na aparência facial como o tônus muscular e a mudanças na textura da pele que ocorrem de forma lenta ao longo do tempo. Em seus experimentos, através da análise comparativa dos modelos, chegaram a conclusão, que métodos que utilizam o Gabor Filter, possuem menos complexidades e provêm uma acurácia entre 82,5% e 99%. E que a maior acurácia foi obtida através do classificador baseado em Support Vector Machine.

3.4 Carga Cognitiva e Aprendizado Multimídia

Alguns trabalhos como o de Paul et al. (2021), investigaram a possibilidade da utilização de métricas fisiológicas para a análise da carga cognitiva. Avaliando e investigando, a hipótese da utilização de técnicas que analisam efeitos fisiológicos para detectar diferentes entre a carga cognitiva intrínseca causada por tarefas de vários níveis de complexidade. Em seus estudos os autores, dividiram as medidas fisiológicas em quatro grandes categorias que são: Medidas cardiorrespiratórias, medidas de atividade ocular, medidas cerebrais e por fim medidas eletrodérmica. Concluiu-se que há elementos suficientes para sustentar a afirmação de que é possível fazer a análise da carga cognitiva através de métodos de análise de dados fisiológicos, coletados do coração, pulmão, olhos, pele e do cérebro. Porém, para uma maior acurácia, os dados devem ser analisados e preditos de forma individual por medida.

Segundo o mesmo estudo, as medidas que são mais sensíveis à variação da carga cognitiva são a taxa de piscada dos olhos, frequência cardíaca, dilatação da pupila e as ondas *alpha*.

Outro ponto bastante discutido é a movimentação da íris, que foi estudada nos artigos de Zagermann, Pfeil e Reiterer (2016) através da análise do movimento dos olhos cruzada com as informações da aplicação de aulas multimídias e aplicações de testes após a aula, comprovando que a movimentação da íris tem impacto no aprendizado cognitivo dos alunos.

Seguindo nessa linha esse estudo é corroborado pelo o de Tianlong et al. (2020), que utiliza o movimento dos olhos para medir as diferentes cargas cognitivas, e que replicou com sucesso o estudo de DeLeeuw e Mayer (2008).

Capítulo 4

Metodologia

No presente capítulo será descrita a metodologia proposta para analisar carga cognitiva em vídeos. Considera-se como entrada vídeos faciais capturados de alunos assistindo diferentes tipos de aula com recursos multimídia. Os vídeos devem ser anotados manualmente. Em seguida, são extraídos automaticamente os pontos de interesse usando a biblioteca MediaPipe (2020) e cruzados posteriormente com as anotações coletadas para responder a pergunta de pesquisa. A metodologia adotada para alcançar os objetivos propostos neste estudo envolve um fluxo estruturado de etapas interconectadas conforme a Figura 4.1.

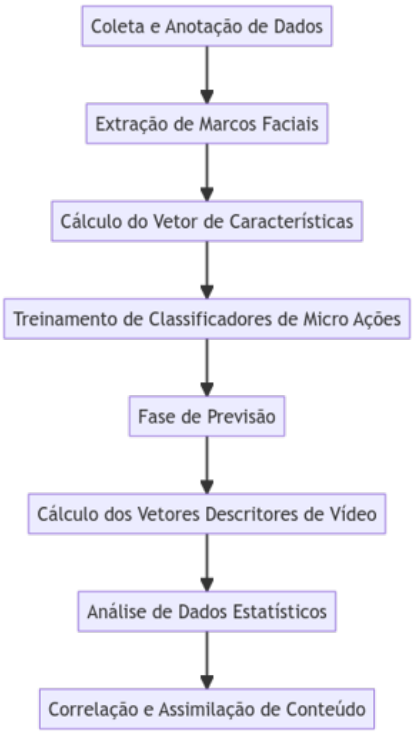


Figura 4.1: Fluxograma da Metodologia Aplicada no Estudo - Autor 2023

Inicialmente, a coleta e anotação dos dados é realizada para criar um conjunto de

referência que servirá como base para análise. A partir dessa base, os pontos de interesse faciais são extraídos dos vídeos coletados. Com estes pontos, o cálculo do vetor de características é conduzido para construir a representação das micro ações faciais ao longo do tempo.

A próxima fase compreende o treinamento de classificadores de micro ações, que se baseiam nos vetores de características previamente calculados. Esses classificadores são desenvolvidos para discernir padrões sutis nas expressões faciais, contribuindo para a detecção e análise de micro ações específicas. Com os classificadores treinados, a fase de predição entra em cena e a seguir, os vetores descritores de vídeo são calculados, o que permite a agregação de informações das micro ações detectadas em sequências temporais mais amplas. Esses vetores fornecem uma representação global das expressões faciais, permitindo a análise de padrões de comportamento e tendências em níveis mais abrangentes. A análise de dados estatísticos entra em cena para interpretar e contextualizar as informações extraídas, identificando relações e insights relevantes.

Finalmente, a correlação e assimilação de conteúdo são realizadas para integrar os resultados obtidos com o conhecimento existente na área de estudo. Isso permite a interpretação mais profunda das variações nas expressões faciais, associando-as a contextos específicos e potencialmente revelando insights sobre comportamentos subjacentes.

4.1 Organização e Pré-Processamento das Imagens

Consideramos como entrada vídeos faciais curtos (poucos minutos) de alunos assistindo sessões multimídia de aprendizado. Estes alunos podem pertencer a grupos distintos. No caso de nossos experimentos, os alunos foram alocados em dois grupos diferentes conforme a Tabela 5.1. O primeiro grupo está relacionado aos alunos que foram submetidos à aula informativa não redundante (NR), versus o segundo grupo aos alunos com aula informativa redundante (R). Um exemplo de imagem de um dos vídeos é exibido na Figura 4.2.

Figura 4.2: Vídeo exemplo dos estudantes. Autor 2023

Não se fez necessária a aplicação de nenhuma técnica para o melhoramento de imagem, remoção de ruídos e tratamento da imagem. Todos os vídeos foram usados em seu formato e qualidade original para o segundo passo, que é a extração das características.

4.2 Estimativa do ganho de aprendizagem

Independentemente do tipo de aula, cada aluno deve ser submetido a um pré-teste e pós-teste, sendo que a mesma aula foi apresentada com 3 minutos de duração diferindo apenas com a adição de elementos interativos na aula redundante conforme Figura 4.3.

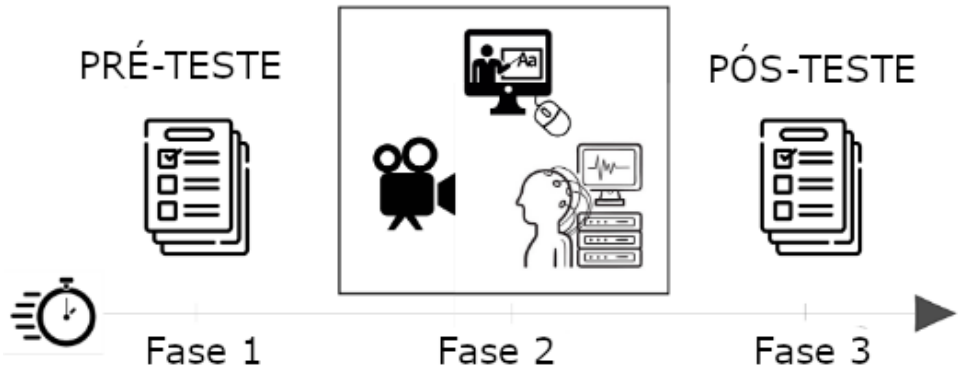


Figura 4.3: Fluxo de coleta de dados por participante: primeiro deve ser medido o seu conhecimento prévio à aula (Fase 1); em seguida, devem ser gravados os vídeos faciais durante a transmissão de uma aula gravada (Fase 2); por fim, será medido o conhecimento com um pós-teste ao final da aula virtual (Fase 3) - Autor (2023).

Para comparar as notas obtidas (pré-teste/pós-teste), adotamos o teste de Wilcoxon (U-Test) para comparar o valor absoluto da diferença entre a pontuação do pré-teste e pós-teste, e obter a medida de ganho de aprendizado cognitivo, conforme a Equação 4.1. Espera-se que uma resposta maior seja apresentada no Pós-teste dentro de cada sujeito, mas em magnitudes diferentes dependendo do grupo.

$$\text{LEARNING GAIN} = |\text{GRADEPOST} - \text{GRADEPRE}| \tag{4.1}$$

4.3 Extração das Características

Finalizada a primeira etapa de pré-processamento, propõe-se o framework do Google o Media Pipe para a extração das características da face. Este framework utiliza Aprendizado de Máquina para inferir a superfície facial em 3D a partir de uma única imagem colorida de uma

câmera de baixo custo, sem a necessidade de um sensor de profundidade dedicado.

O pipeline do MediaPipe, no que tange Aprendizado de Máquina, é composto por dois modelos de redes neurais profundas que funcionam em tempo real e de forma conjunta: uma que opera na imagem completa e calcula os locais e bordas da face e a segunda utiliza um modelo de referência de face em 3D, prevendo uma superfície 3D do rosto por meio de regressão que em seguida é usada para a identificação e extração das características da face e da íris dos vídeos, conforme podemos verificar nas Figuras 4.4 e 4.5.

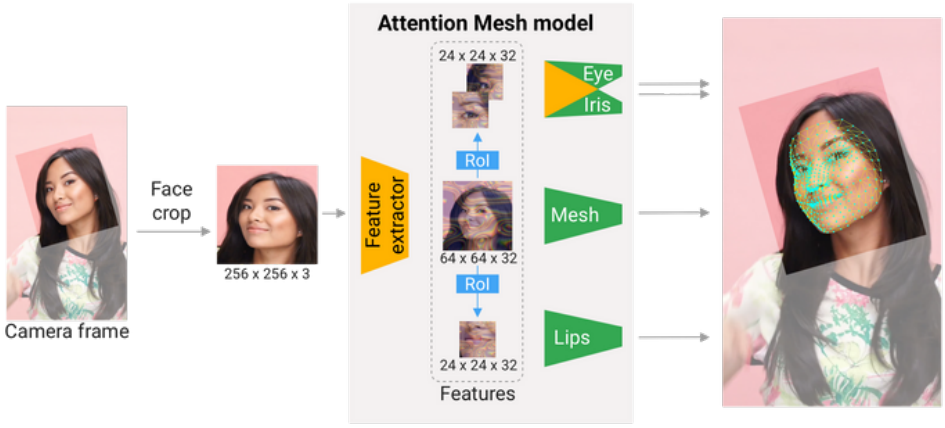


Figura 4.4: Arquitetura do Face Mesh - MediaPipe (2020)

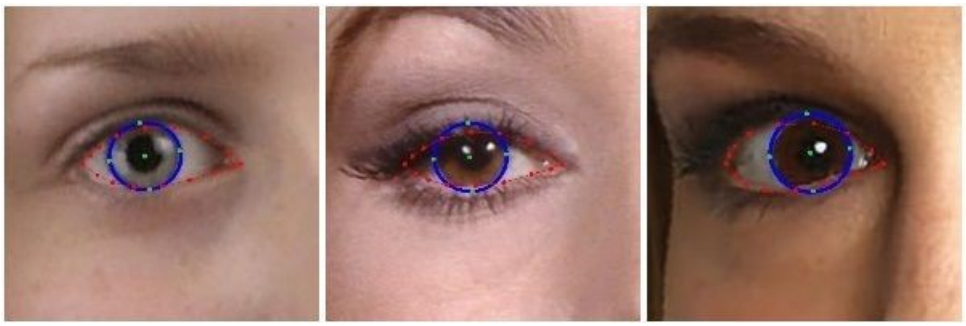


Figura 4.5: Exemplo de detecção de íris com o MediaPipe - MediaPipe (2020)

Cada frame do vídeo é convertido pelo modelo em uma lista com 468 pontos conforme Figura 4.6, onde cada ponto é composto pelas coordenadas x, y e z. As coordenadas x e y representam a largura e a altura da imagem respectivamente. Já a coordenada z determina a profundidade, sendo que quanto menor o valor, mais próximo o ponto estará da câmera.

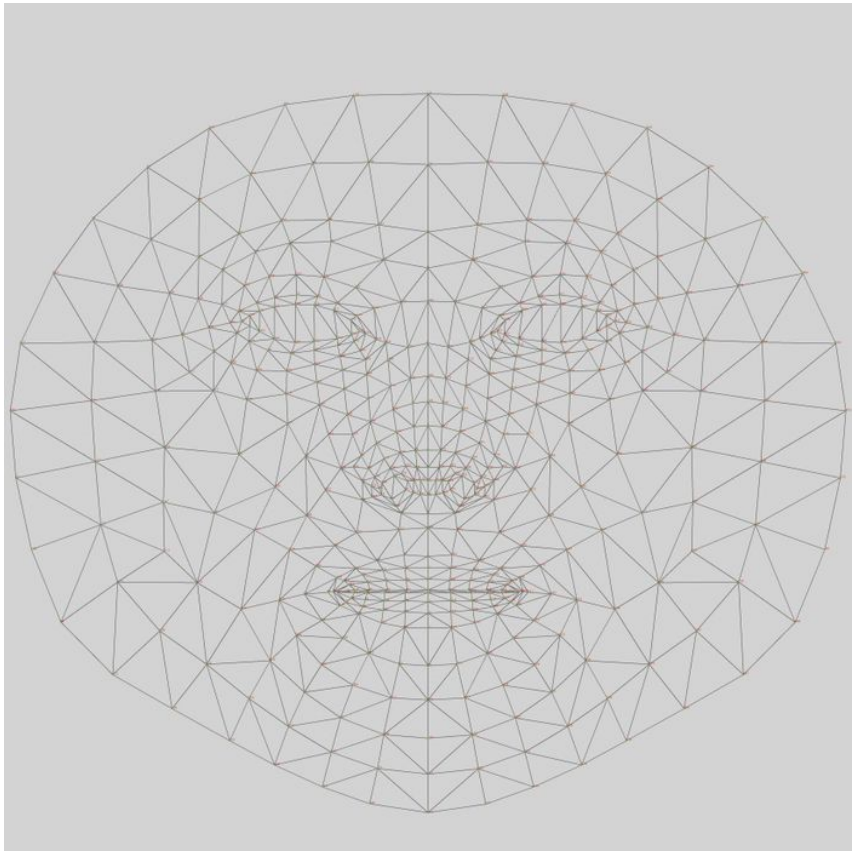


Figura 4.6: Numeração dos pontos coletados pela MediaPipe - MediaPipe (2020).

Região Facial	**Índices dos pontos**
Sobrancelha Direita	46, 53, 52, 65, 55, 70, 63, 105, 66, 107
Sobrancelha Esquerda	276, 283, 282, 295, 285, 300, 293, 334, 296, 336
Olho Direito	33, 7, 163, 144, 145, 153, 154, 155, 133, 246, 161, 160, 159, 158, 157, 173
Olho Esquerdo	263, 249, 390, 373, 374, 380, 381, 382, 362, 466, 388, 387, 386, 385, 384, 398
Iris Direita	469, 470, 471, 472
Iris Esquerda	474, 475, 476, 477
Labios	61, 146, 91, 181, 84, 17, 314, 405, 321, 375, 291, 185, 40, 39, 37, 0, 267, 269, 270, 409, 78, 95, 88, 178, 87, 14, 317, 402, 318, 324, 308, 191, 80, 81, 82, 13, 312, 311, 310, 415
Nariz	168, 6, 197, 195, 5, 4, 1, 19, 94, 2, 98, 97, 326, 327, 294, 278, 344, 440, 275, 45, 220, 115, 48, 64, 98
Contorno do Rosto	10, 338, 297, 332, 284, 251, 389, 356, 454, 323, 361, 288, 397, 365, 379, 378, 400, 377, 152, 148, 176, 149, 150, 136, 172, 58, 132, 93, 234, 127, 162, 21, 54, 103, 67, 109, 10

Tabela 4.1: Conjunto de Pontos FaceMesh por Região Facial. MediaPipe (2020)

Os pontos faciais coletado através do Facemesh são enumerados de 0 a 468 e por padrão alguns desses pontos estão agrupados por região facial conforme 4.1. Esse agrupamento pode ser encontrado na documentação do MediaPipe (2020) - FaceMesh disponibilizada pelo Google ou através de funções dentro do próprio código da biblioteca.

No entanto, nem todos os 468 pontos coletados dos vídeos foram utilizados nas etapas seguintes. Uma vez que as micro ações são caracterizadas pelo movimento de pontos de interesse em regiões específicas do rosto, portanto apenas subconjuntos específicos desses pontos são usados para treinar cada classificador de micro ação.

De maneira sucinta, pontos relacionados às regiões dos olhos, íris, sobrancelhas, boca e contornos faciais foram utilizados para pelo menos uma micro ação, totalizando assim 100 pontos representados por suas coordenadas x e y. Além disso, as coordenadas z mostraram-se irrelevantes em experimentos preliminares, e, portanto, foram descartadas.

Para alcançar uma representação facial invariante às variações de formas de rosto dos

indivíduos, foi incluída uma etapa de normalização. Como cada micro ação depende de regiões específicas, a normalização foi realizada por região. Especificamente, a caixa delimitadora bidimensional dos pontos de cada região é primeiro calculada, e então seus pontos correspondentes são normalizados de forma que a caixa delimitadora resultante se torne unitária e centrada na origem. Dessa maneira, é alcançada a invariância à translação e à escala de cada região. Para tal foi executada a normalização dos pontos em cada frame por área de interesse de cada vídeo através do cálculo das caixas delimitadoras (*bounding boxes*) nas áreas de interesse utilizando a Equação 4.2.

$$x^R_{\min} = \min_{i=1}^{n_R} x_i \tag{4.2}$$

$$x^R_{\max} = \max_{i=1}^{n_R} x_i \tag{4.3}$$

$$y^R_{\min} = \min_{i=1}^{n_R} y_i \tag{4.4}$$

$$y^R_{\max} = \max_{i=1}^{n_R} y_i \tag{4.5}$$

$$\text{bbox}^R = [x^R_{\min}, y^R_{\min}, x^R_{\max}, y^R_{\max}], \tag{4.6}$$

onde bbox^R é calculada apenas considerando os pontos $\{x^R_1, \ldots, x^R_{n_R}\}$ da região R. Após encontrar a bounding box de cada região de interesse, é aplicada uma normalização espacial individualmente para cada região, conforme a Equação 4.7.

$$\tilde{x}^R_i = \frac{x^R_i - x^R_{\min}}{x^R_{\max} - x^R_{\min}} \qquad \tilde{y}^R_i = \frac{y^R_i - y^R_{\min}}{y^R_{\max} - y^R_{\min}} \tag{4.7}$$

4.3.1 Classificadores de micro ações

Para treinar classificadores de micro ações através de aprendizado supervisionado, foi necessário gerar uma base de dados com as anotações das micro ações a nível de frame, para um conjunto de vídeos. As anotações devem ser realizadas manualmente conforme a Tabela 4.2, onde 4 micro ações são contempladas, além da micro ação neutra: NEUTRA, OLHO_FECHADO (OF), OLHANDO_PARA_CANTO (OC), MEXEU_LABIOS (ML) e VIROU_ROSTO (VR).

Frame	Neutro	OF	OC	ML	VR
1	0	1	0	0	0
2	0	1	0	0	0
3	0	1	0	0	0
4	1	0	0	0	0
5	1	0	0	0	0
6	0	0	1	0	0
N	1	0	0	0	0

Tabela 4.2: Exemplo de Anotação por Vídeo

Para facilitar esta tarefa entediante de analisar e anotar cada quadro, foi desenvolvida uma ferramenta em Flask e Dash que plotava em tela os frames com maiores diferenças por região, ilustrada na Figura 4.7, possibilitando assim a análise dos frames de forma visual dentro da ferramenta.

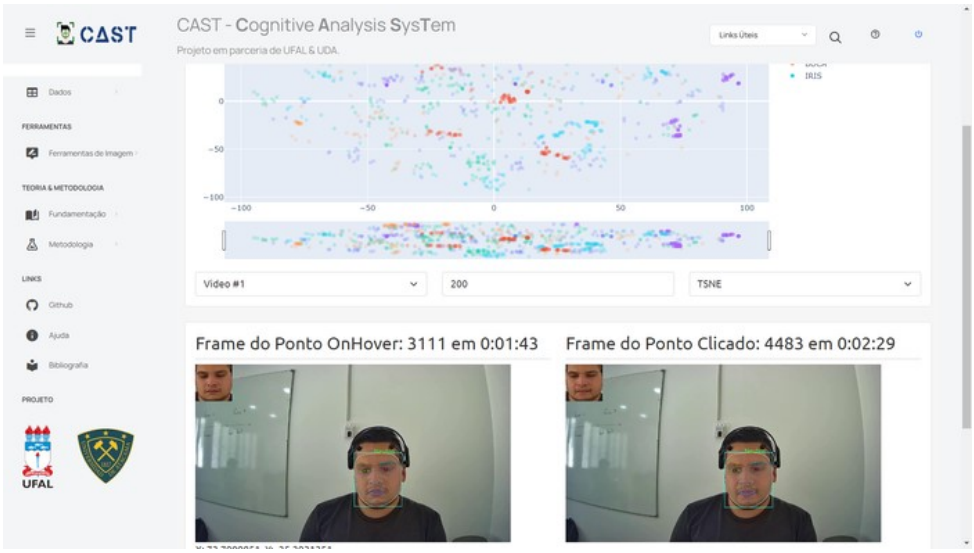


Figura 4.7: Ferramenta para visualização dos frames - Autor (2023).

Para o treinamento dos modelos, os pontos foram agrupados de acordo com cada micro ação a ser reconhecida e suas regiões de interesse. A Tabela 4.1 descreve as regiões utilizadas em nossos experimentos e os índices dos pontos de cada região.

Para o treinamento não foi feita nenhuma normalização e/ou mudança de escala além das executadas na extração das posições conforme a seção 4.2. Os classificadores foram treinados para detectar micro ações, a nível de quadro, considerando uma sequência de características dos quadros anteriores e do quadro a ser predito. Mais especificamente, o classificador foi desenhado para receber como entrada uma sequência de 7 quadros consecutivos, e classificar a micro ação do último destes quadros.

4.3.2 Arquitetura dos Classificadores de Micro Ação

Seja $\mathcal{D} = \{(\mathcal{V}_1, l_1), (\mathcal{V}_2, l_2), \ldots, (\mathcal{V}_n, l_n)\}$ o conjunto de pontos representando n vídeos, onde $\mathcal{V}_i = \{v_1^i, v_2^i, \ldots, v_{n_i}^i\}$ é a sequência de descritores para o vídeo i, v_j^i representa as posições dos pontos relevantes do quadro j no vídeo i, e l_i são as anotações do vídeo. Cada v_j^i é composto por posições 2D de 100 pontos, conforme descrito na 4.1, mas apenas alguns deles são necessários para treinar cada classificador de micro ação.

Os classificadores foram projetados para detectar micro ações individuais, em nível de quadro, tomando como entrada uma sequência de 7 quadros consecutivos, e classificando a micro ação ocorrendo no último destes quadros.

Assim, para treinar cada classificador, foi adotada a seguinte abordagem supervisionada. Cada vídeo foi processado para gerar muitas pequenas amostras compostas de 7 quadros consecutivos, por meio de uma abordagem de janela deslizante. Especificamente, cada vídeo $\mathcal{V}_i$, que é composto por n_i quadros, gera $n_i - 6$ instâncias de 7 quadros consecutivos. O alvo para cada instância é inicialmente uma micro ação anotada manualmente. Note que a anotação manual de ações nos vídeos é necessária apenas em uma fase de treinamento, e então a análise de carga cognitiva pode ser realizada em outros vídeos sem repetir essa etapa.

Durante a fase de treinamento, cada instância é associada a uma micro ação anotada manualmente, que serve como rótulo alvo. Importante ressaltar que a anotação manual das ações nos vídeos é necessária apenas nessa etapa inicial de treinamento. Posteriormente, essa base de conhecimento é utilizada para alimentar os classificadores, permitindo a análise de micro ações em outros vídeos sem a necessidade de repetir a anotação.

O processo ilustrado no diagrama 4.8 abaixo, que demonstra a sequência de operações envolvidas na preparação das amostras para treinamento dos classificadores, proporcionando um fluxo contínuo e eficiente para a análise de expressões faciais.

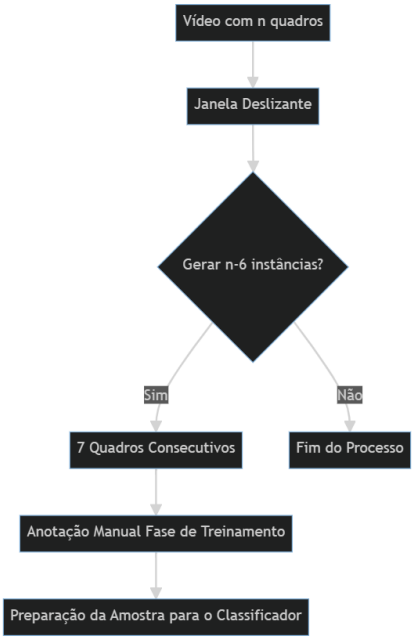


Figura 4.8: Diagrama para geração da Amostra - Autor (2023).

A seguir, para cada micro ação a, um classificador f_a é treinado, restringindo primeiro os pontos relevantes em $\mathcal{D}$ para sua região de interesse $\mathcal{R}_a$, de acordo com a Tabela 4.1. Além disso, o alvo para cada quadro de cada vídeo é convertido em um número binário que terá valor 1 apenas se a micro ação anotada do quadro correspondente for a.

Cada classificador de micro ação f_a é uma rede neural profunda com uma arquitetura baseada em LSTM, como representado na Figura 4.9. A camada de entrada considera uma sequência de 7 quadros, com 200 números (coordenadas x e y de 100 pontos, embora esse número possa ser menor dependendo do número de pontos de interesse em $\mathcal{R}_a$). Então, uma camada densa é aplicada em paralelo para cada quadro, para incorporar os dados de cada quadro em um espaço de menor dimensão, de 64. Em seguida, três camadas LSTM consecutivas tentarão capturar e reter dependências de longo prazo nestes dados sequenciais. Essas camadas têm, respectivamente, 32, 16 e 16 unidades. Cada uma dessas camadas tem a conhecida ativação ReLU. Finalmente, uma camada de saída com 2 unidades e ativação softmax é empregada para realizar a classificação binária final do último quadro da sequência de entrada.

Também vale mencionar que se espera um alto desbalanceamento das classes ao longo do tempo, onde na maioria dos quadros os alunos não estão realizando uma micro ação específica. Para mitigar esse problema, foi empregado um esquema de ponderação inversa para dar maior peso aos quadros que contêm micro ações.

Neste trabalho, treinamos e avaliamos 4 micro ações relevantes: OLHO_FECHADO (OF), OLHAR_PARA_CANTO (OC), MEXEU_LABIOS (ML) e VIROU_ROSTO (VR).

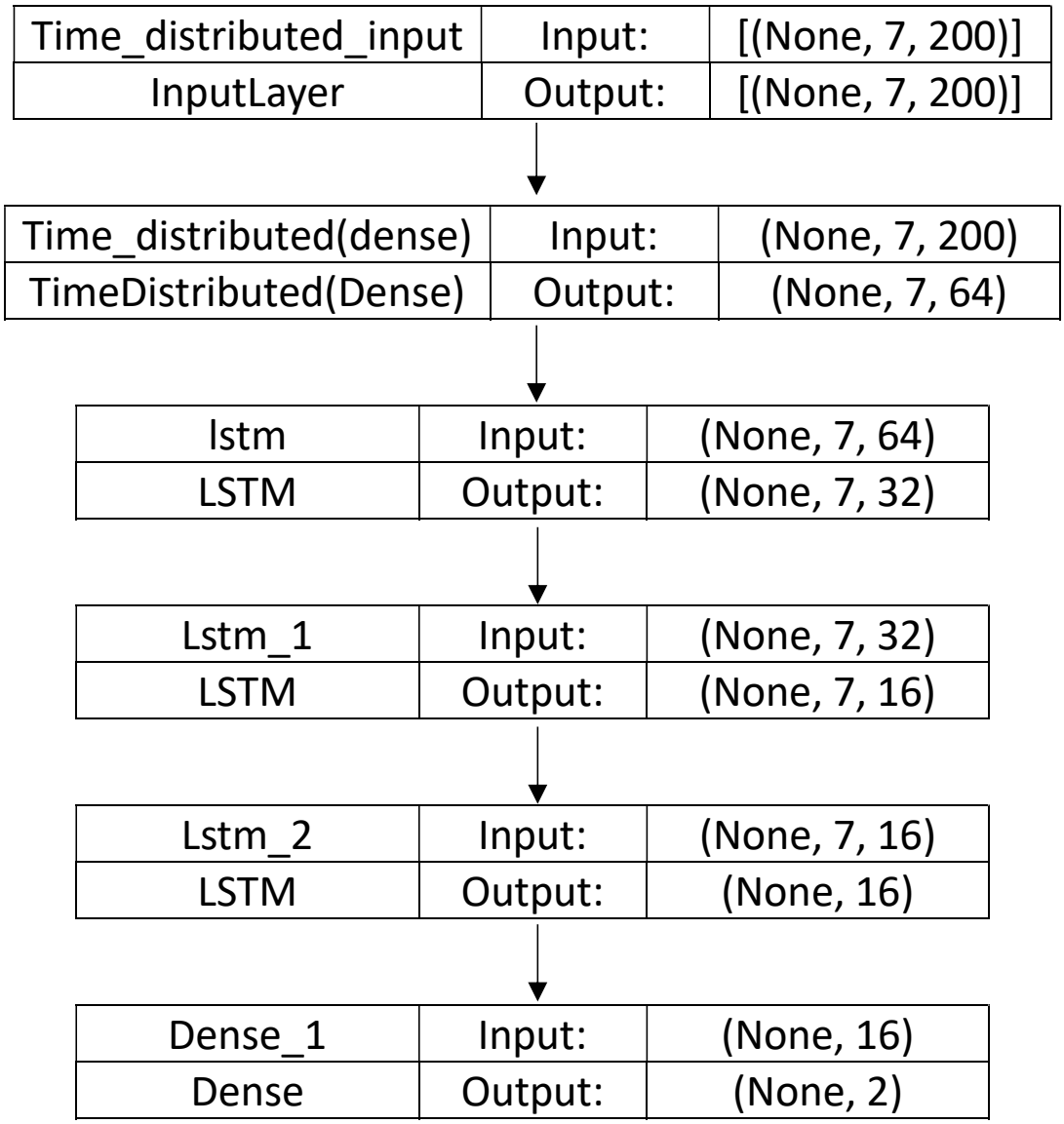


Figura 4.9: Arquitetura dos classificadores de micro ação: a partir de uma sequência de pontos de interesse de 7 quadros consecutivos, o classificador foi projetado para reconhecer se uma ação específica está sendo executada no último quadro da sequência - Autor (2023).

4.3.3 Sumarização dos Resultados

Para cada vídeo i, cada classificador de micro ação deve ser capaz de identificar os quadros em que sua micro ação correspondente ocorre. Seja $f_a(\mathcal{V}_i)$ a sequência de previsões fornecida pelo classificador f_a. Para calcular quantas vezes uma micro ação específica treinada a ocorre no vídeo i, previsões consecutivas em $f_a(\mathcal{V}_i)$ são removidas. Por exemplo, 00000111111100111 se torna 0101. Então, observe que a soma dos elementos resultantes é o número de ocorrências da micro ação a. Ao contar o número de ocorrências para cada micro ação, um descritor de vídeo

pode ser gerado para o vídeo i como $\delta_i = (c_1, c_2, \ldots, c_m)$, onde m é o número de ocorrências das micro ações treinadas (no nosso caso, $m = 4$). Este descritor resume o vídeo de maneira concisa, fornecendo informações relevantes para a análise de carga cognitiva.

Capítulo 5

Experimentos

Nesta seção serão descritos os experimentos realizados e os resultados obtidos para responder as perguntas de pesquisa levantadas na Seção 1.3. Especificamente, estes experimentos buscam validar e avaliar os classificadores de micro ações, e realizar uma prova de conceito para demonstrar como a metodologia proposta é capaz de identificar padrões relevantes de ocorrências de micro ações e assimilação de conteúdo.

O experimento e as aplicações de apoio foram desenvolvidas em Python Rossum (1995), devido a sua flexibilidade para a construção de aplicações analíticas e da disponibilidade de bibliotecas como o Keras Chollet et al. (2015), Scikit-learn [1], OpenCV: [2] and Flask[3].

Os experimentos foram processados através em um conjunto de dados específicos de vídeos previamente coletados, utilizando uma máquina com processador Intel Core i7-7500U de 2.7 GHz com 4 núcleos, 32GB de memória RAM, placa de Video NVidia 1650Gtx e 1,5T de disco (SSD).

5.1 Coleta de dados

Foi utilizado para o experimento vídeos de 13 alunos da Faculdade de Medicina da Universidade do Atacama (UDA), que foram alocados em dois grupos diferentes. O primeiro grupo está relacionado aos alunos que foram submetidos à aula informativa não redundante (NR), versus o segundo grupo aos alunos com aula informativa redundante (R), conforme a Tabela 5.1. Independentemente do tipo de aula, cada aluno foi submetido a um pré-teste e pós-teste, sendo que a mesma aula foi apresentada com 3 minutos de duração diferindo apenas com a adição de elementos interativos na aula redundante.

[1] <https://scikit-learn.org/>
[2] <https://opencv.org/>
[3] <https://flask.palletsprojects.com/en/2.3.x/>

Aula informativa - Grupo	Homens	Mulheres	Total
#1 (Não redundante)	4	3	7
#2 (Redundante)	2	4	6

Tabela 5.1: Distribuição de genero dos Dados Originais segundo grupo de aulas.

Os vídeos dos alunos foram coletados previamente à realização desta pesquisa, possuindo um tamanho médio de 400MB, com a resolução de 1920×1080 e em formato MP4. O conjunto de dados é composto por 13 vídeos de 3 minutos em média conforme Figura 4.2.

5.2 Análise dos testes de aprendizagem

Cada aluno foi submetido a um pré-teste e pós-teste, sendo que a mesma aula foi apresentada com 3 minutos de duração diferindo apenas com a adição de elementos interativos na aula redundante. O Teste de Wilcoxon (U-Test) foi aplicado, resultando nos resultados revelados na Tabela 5.2.

Estudante/vídeo	Tipo	Ganho de Aprendizagem	Variação
1	NR	3.6	25.71%
2	R	8	57.14%
3	NR	2.5	17.86%
4	R	8.5	60.71%
5	NR	8.9	63.57%
6	NR	3	21.43%
7	NR	7.2	51.43%
8	R	8	57.14%
9	NR	2.5	17.86%
10	R	4.2	30.00%
11	NR	7.3	52.14%
12	R	4	28.57%
13	R	9.5	67.68%

Tabela 5.2: Ganho de aprendizagem e diferença percentual relativa por aluno.

Escolhendo um nível de significância igual a 5%, rejeitamos uma hipótese nula (H0: $MEDIANA_{REDU} = MEDIANA_{NON-REDU}$) de que o ganho mediano obtido do grupo Redundante (R) é igual à mediana do grupo Não Redundante (NR), ou seja, as medianas entre os grupos são estatisticamente iguais. Através do teste U, a probabilidade que esses resultados sejam iguais é de p-value = 0,04974, e considerando um nível de significância de 5%, se rejeita a hipótese de igualdade. Visualmente também é possível se observar a distribuição empírica entre os grupos e

a sua disparidade (Figura 5.1). Os elementos indicam que a redundância de aula, em uma aula de 3 minutos, pode resultar em um aprendizado cognitivo diferente que uma aula não redundante.

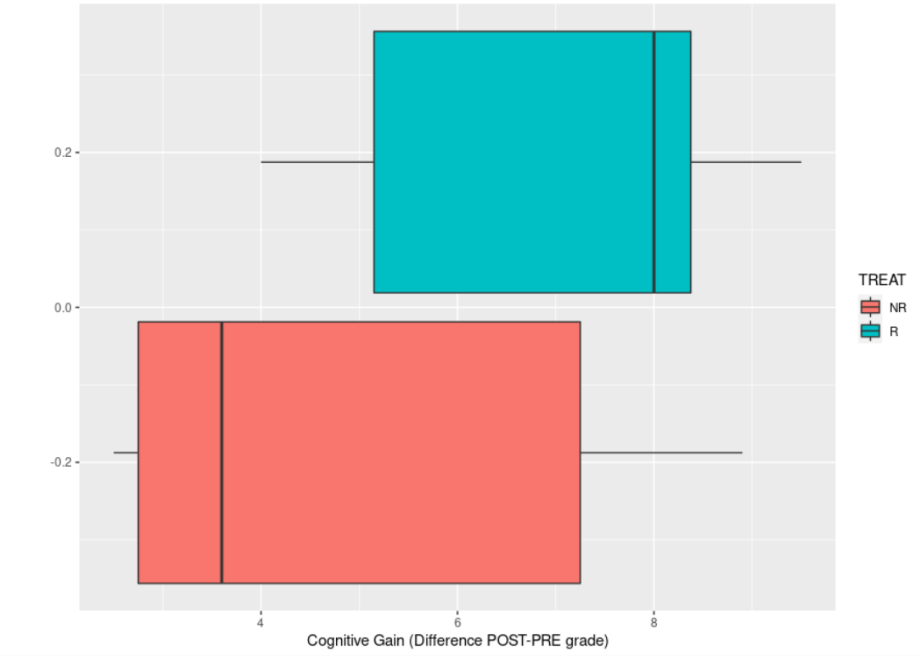


Figura 5.1: Ganho cognitivo por Grupo de Teste - Autor. (2023).

Além disso, a Figura 5.2 descreve a diferença de pontuação entre os grupos nas questões (os pontos são a pontuação média por grupo, enquanto a cor vermelha representa o grupo NR e a cor azul representa R, as barras são a representação da variabilidade das respostas por grupo). Uma vez que o ganho de aprendizagem cognitiva é perceptível, a próxima investigação é relacionar aos pontos de interesse faciais que podem estar relacionadas às reações dos participantes usando Visão Computacional.

5.3 Classificadores de Micro Ações

Inicialmente, previa-se a utilização dos dados recolhidos da face dos 13 participantes iniciais. Porém, dados de dois vídeos foram retirados do experimento devido a sua baixa qualidade de gravação e de outros dois vídeos foram retirados devido a utilização de máscara de proteção facial. Consequentemente, usou-se um total de 9 vídeos com a composição de cada grupo para o experimento conforme Tabela 5.3.

Com a exclusão dos vídeos, obtemos uma nova distribuição de IDs e suas referências conforme a Tabela 5.4 abaixo, esses novos códigos serão referenciados nos experimentos.

As anotações de micro ações foram feitas manualmente no dataset a nível de frame, conforme a Tabela 5.5. Algumas micro ações tiveram duração maior do que um frame como,

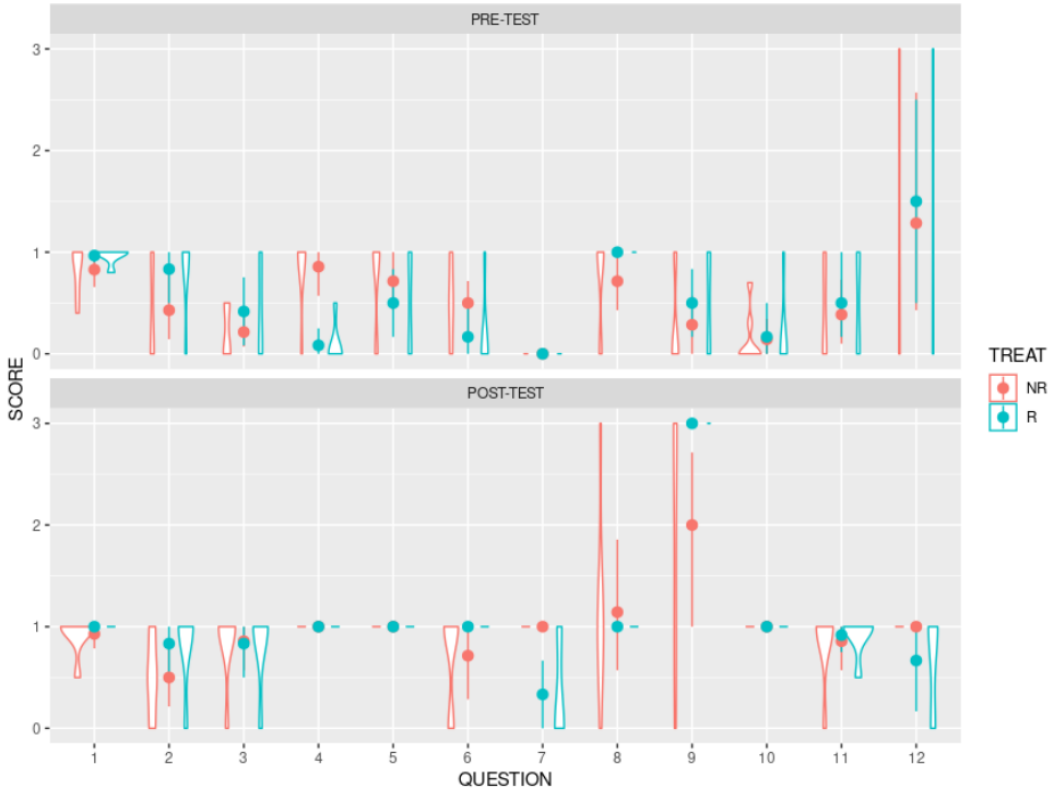


Figura 5.2: Ganho na aprendizagem por Grupo - Autor. (2023).

Grupo	Homens	Mulheres	Total
#1	2	3	4
#2	2	3	5

Tabela 5.3: Conjunto de vídeos usado para experimentos com classificadores de micro ações.

por exemplo, a ação OLHO_FECHADO (OF) que demora em média em nossas anotações de 3 à 5 frames. Vale ressaltar que algumas ações podem ocorrer em conjunto como por exemplo MEXEU_LABIOS (ML) e OLHO_FECHADO (OF). A ação VIROU_ROSTO (VR) não apresentou muitos registros, mas poderá ser primordial em algumas ações e a ação OLHANDO_PARA_CANTO (OC) foi a que apresentou maior frequência.

Como descrito anteriormente, o treinamento dos classificadores de micro ações considera apenas pontos de regiões de interesse específicas de cada micro ação. Em nossos experimentos, os pontos das regiões utilizados de acordo com cada micro ação é descrito na Tabela 5.6.

Os classificadores foram treinados usando o conhecido otimizador Adam, durante até 40 épocas, com um tamanho de lote de 32 e uma taxa de aprendizado de 10^{-4}. A parada precoce (*early stopping*) também foi usada para evitar o overfitting. Esses hiper parâmetros específicos foram ajustados através de otimização bayesiana conforme a Tabela 5.7.

ID	Video/Aluno	Tipo	Ganho de Aprendizado	Variação
#1	4	R	8.5	60.71%
#2	5	NR	8.9	63.57%
#3	6	NR	3	21.43%
#4	7	NR	7.2	51.43%
#5	8	R	8	57.14%
#6	10	R	4.2	30.00%
#7	11	NR	7.3	52.14%
#8	12	R	4	28.57%
#9	13	R	9.5	67.68%

Tabela 5.4: Resultados dos testes aplicados nos vídeos com qualidade adequada.

Vídeo	OF	OC	ML	VR
#1	70	165	2	0
#2	77	99	7	0
#3	38	237	7	1
#4	150	242	2	0
#5	84	158	0	0
#6	126	201	1	0
#7	77	205	3	0
#8	96	238	1	0
#9	73	174	0	0
Total	791	1719	23	1

Tabela 5.5: Ações anotadas manualmente nos dados - Autor (2023)

Regiões de interesse	Micro ação
Olho Direito, Olho Esquerdo, Iris	OLHO_FECHADO
Olho Direito, Olho Esquerdo, Iris	OLHANDO_PARA_CANTO
Boca	MEXEU_LABIOS
Olho Direito, Olho Esquerdo, Iris, Boca	VIROU_ROSTO
Olho Direito, Olho Esquerdo, Iris, Boca	NEUTRO

Tabela 5.6: Conjunto de características e ações correspondentes -Autor 2023

5.3.1 Visualização de dados de micro ações

Os vetores normalizados que descrevem as expressões faciais de cada quadro dos vídeos possuem uma alta dimensionalidade. Para possibilitar uma exploração visual destes dados de forma interativa, utilizamos o t-distributed Stochastic Neighbor Embedding (t-SNE) Cook et al. (2007), a fim de reduzir a dimensionalidade dos dados conforme Figura 5.3.

Hiperparametro	Valor
Num epoch	40
Batch size	34
Leraning rate	0.00010548643264689491
Decay Rate	$\frac{LearningRate}{Nepoch}$

Tabela 5.7: Melhores combinações de hiper parâmetros - Autor 2023

Figura 5.3: Imagem do Aplicativo em Flask - Autor (2023).

O t-SNE Cook et al. (2007) é um algoritmo de aprendizado de máquina não supervisionado que mapeia dados de alta dimensionalidade em um espaço de baixa dimensionalidade, geralmente 2D ou 3D, preservando as relações de proximidade entre os pontos. Isso permite uma representação visual dos dados que revela agrupamentos e padrões intrínsecos. Neste estudo, aplicamos o t-SNE para visualizar a distribuição dos dados em nove interaçoes diferentes. Os resultados são apresentados na Figura **??**, onde cada imagem representa um interação específico. As imagens são intituladas "TSNE - Interação X | Com Ação" e "TSNE - Interação X | Sem Ação", onde X varia de 1 a 9, indicando o número da interação. Essa visualização nos permite observar como os dados se agrupam e distinguem entre os diferentes interaçoes, fornecendo insights importantes para a análise das micro ações faciais.

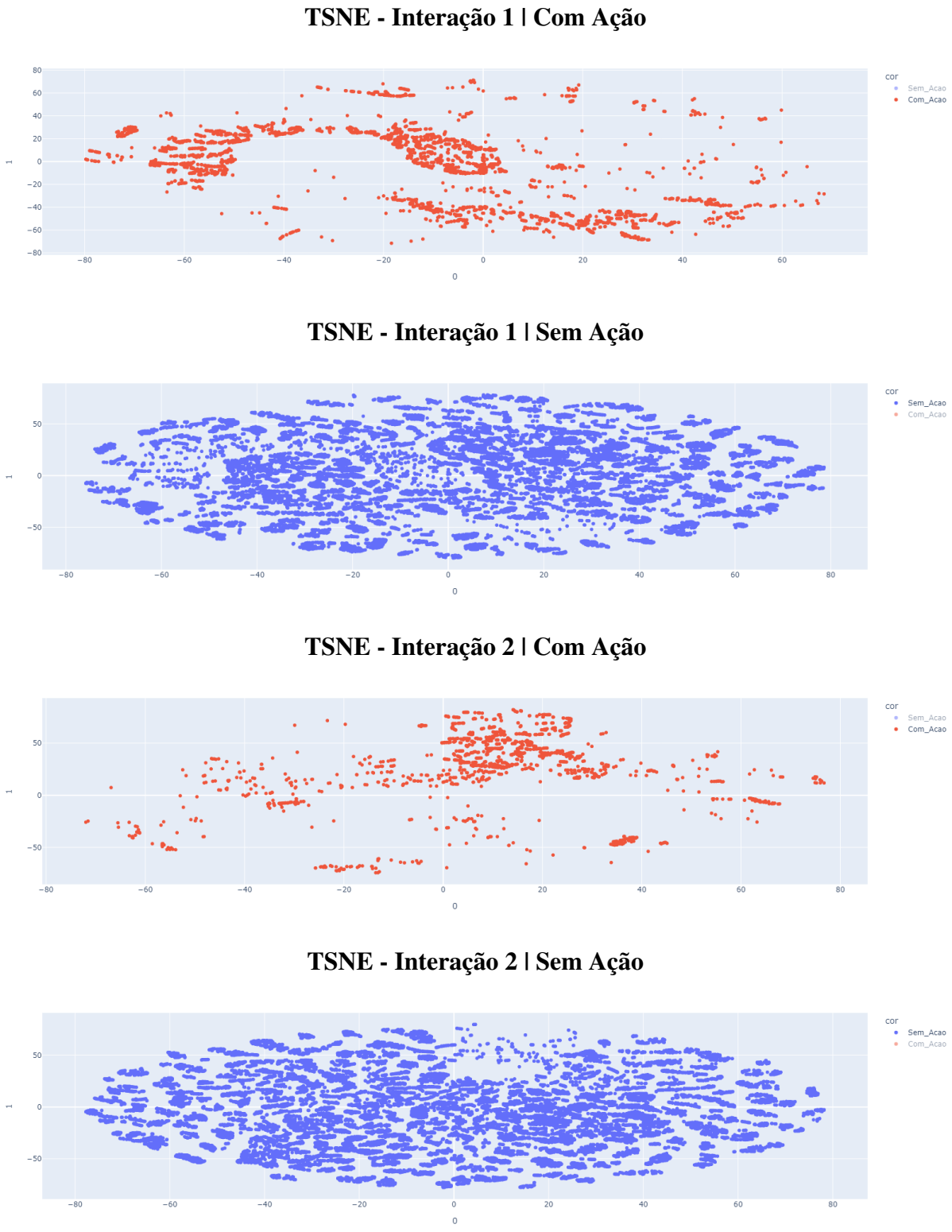


Figura 5.4: Visualização t-SNE da Interação 1 e 2.

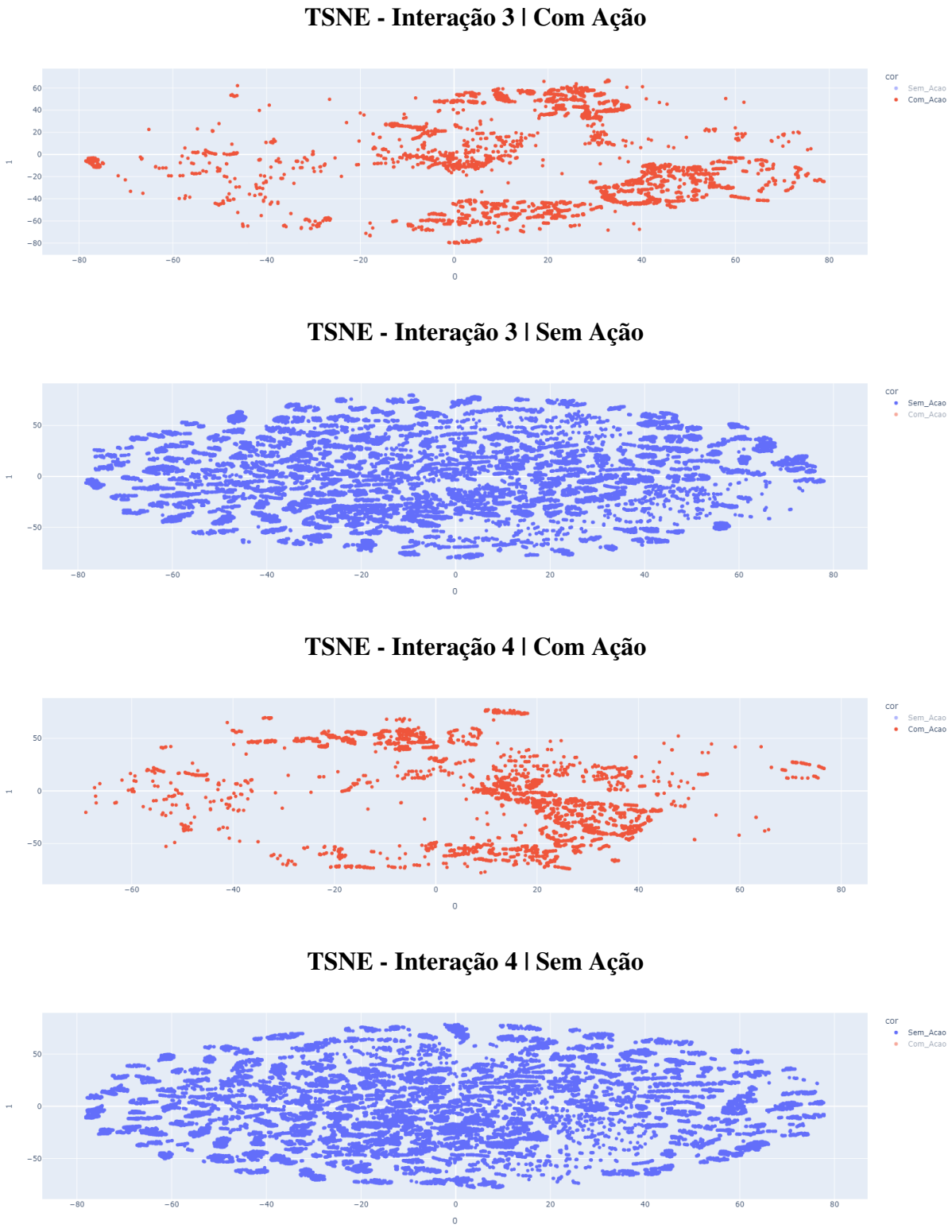


Figura 5.5: Visualização t-SNE da Interação 3 e 4.

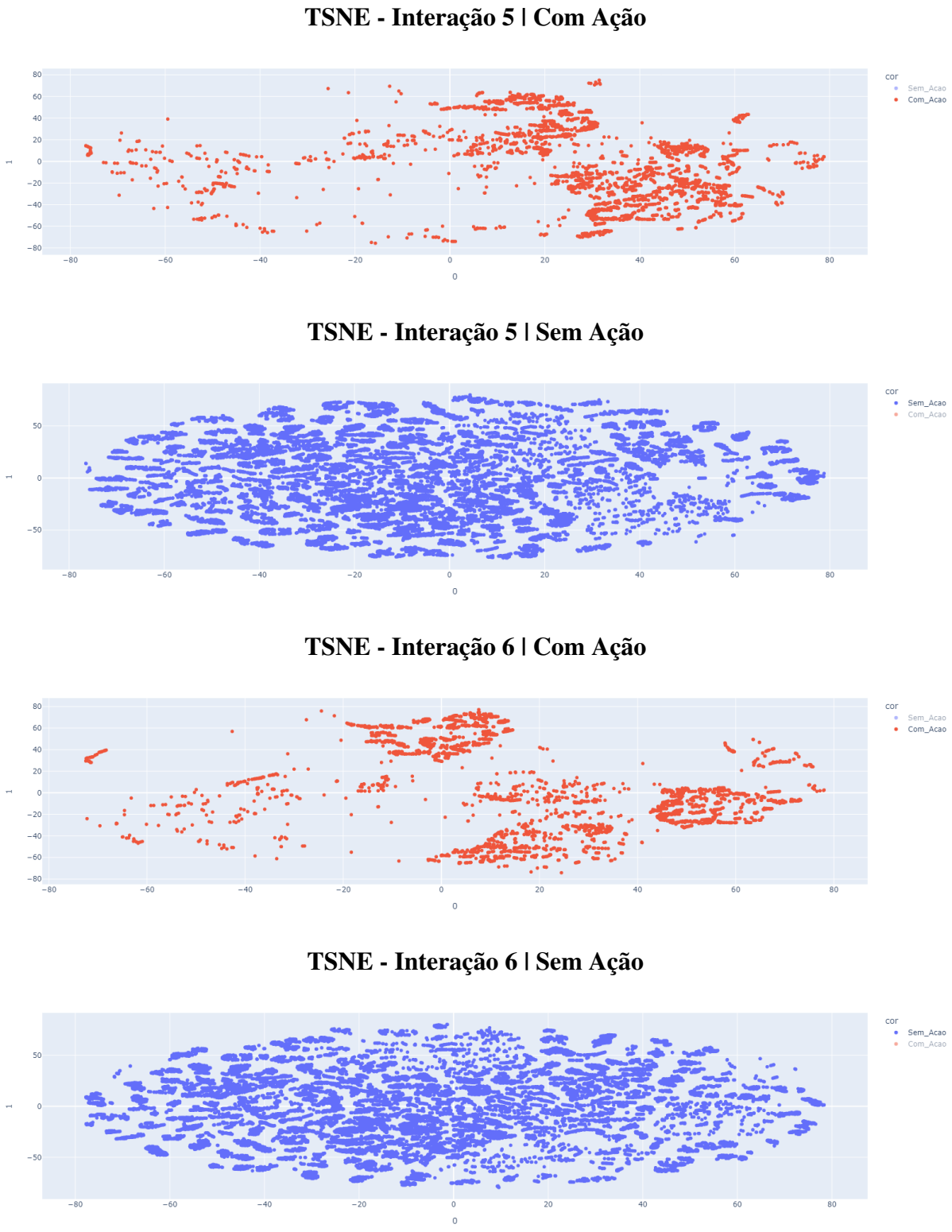


Figura 5.6: Visualização t-SNE da Interação 5 e 6.

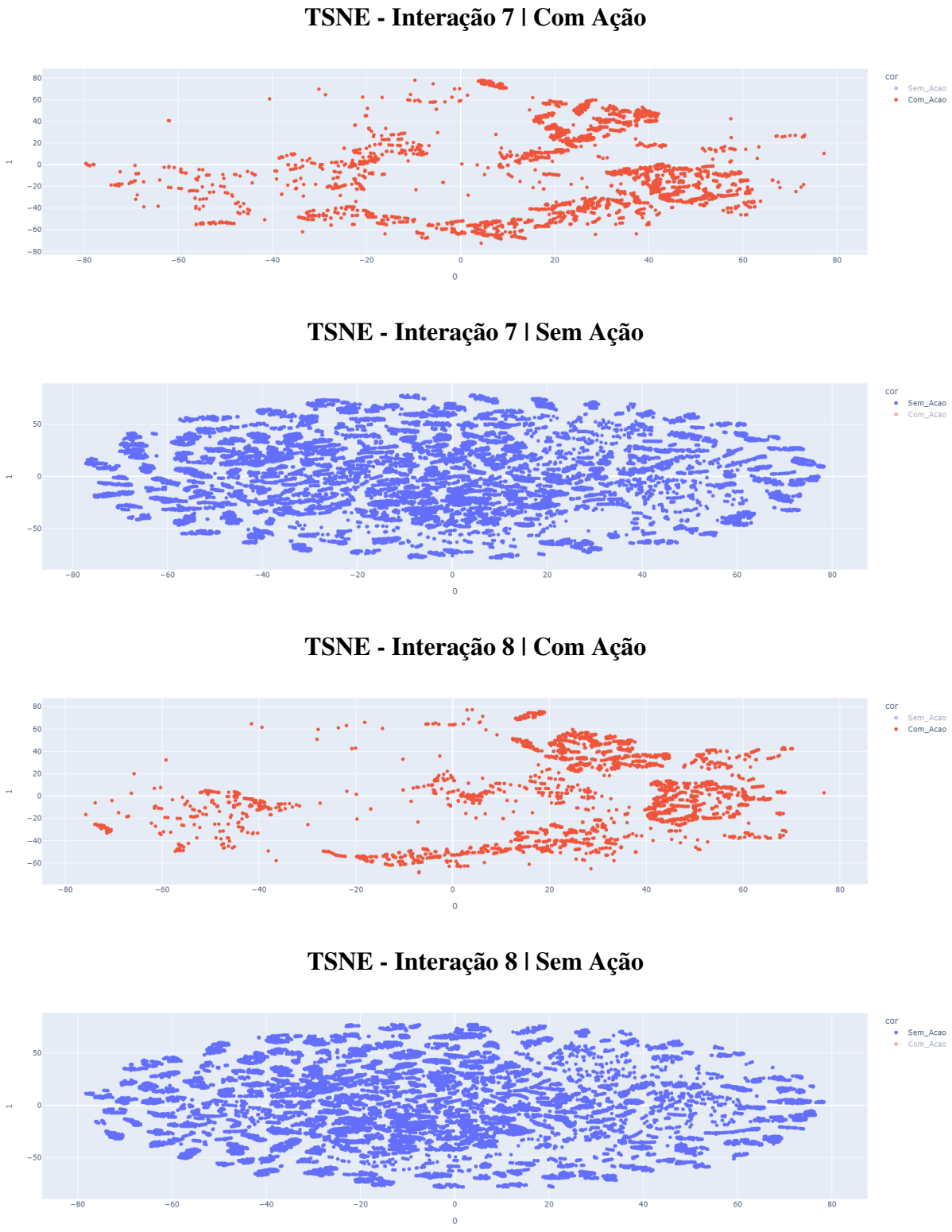


Figura 5.7: Visualização t-SNE da Interação 7 e 8.

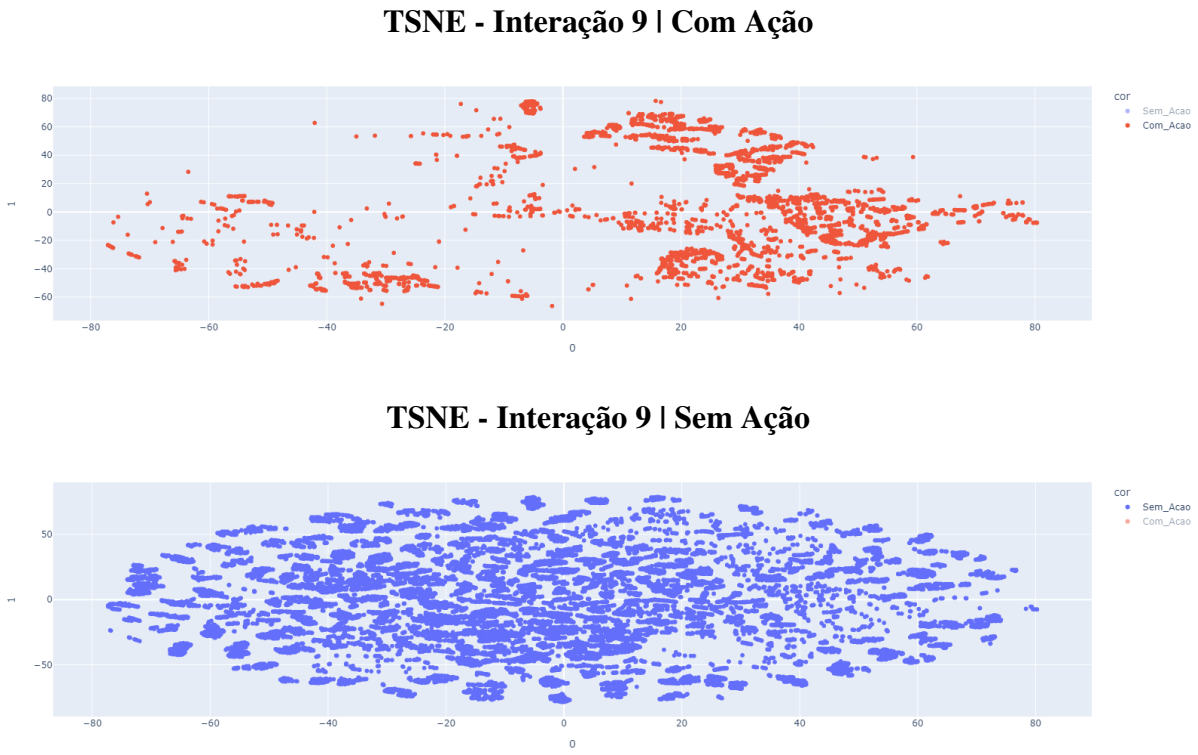


Figura 5.8: Visualização t-SNE da Interação 9.

5.3.2 Validação e avaliação dos resultados

Para a divisão entre treino e teste, foi adotada a estratégia *leave-one-video-out* onde, em cada iteração, um vídeo é usado como conjunto de teste, e os outros compõem o conjunto de treinamento, como descrito na Tabela 5.8.

Vídeo	Treino	Teste
#1	(2-9)	1
#2	(1;3-9)	2
#3	(1,2;4-9)	3
#4	(1-3;5-9)	4
#5	(1-4;6-9)	5
#6	(1-5;7-9)	6
#7	(1-6;8-9)	7
#8	(1-7;9)	8
#9	(1-8)	9

Tabela 5.8: Divisão de Treino e Teste dos Dados

Os resultados obtidos durante o treinamento e teste dos classificadores de micro ações

faciais estão apresentados na Tabela 5.9. Cada vídeo do conjunto possuía em média 5500 frames. Porém, para o cálculo do erro absoluto, foram considerados apenas os frames correspondentes às ações e aos frames neutros, excluindo as repetições do mesmo conforme descrito na Seção 4.3.3.

Vídeo	**OF**	**OC**	**ML**	**VR**
#1	71	113	2	0
#2	47	73	7	0
#3	47	247	6	1
#4	149	265	2	0
#5	58	124	0	0
#6	119	155	1	0
#7	75	160	4	0
#8	94	203	1	0
#9	65	142	0	0
Total	725	1482	23	1

Tabela 5.9: Tabela de erro absoluto dos classificadores de micro ações - Autor (2023)

Na Tabela 5.10 o erro percentual relativo é exibido. O erro relativo permite comparar o desempenho do modelo em diferentes interaçoes, já que ele representa a precisão do modelo em relação ao valor real em cada um deles.

Vídeo	**OF**	**OC**	**ML**	**VR**
#1	1.4%	31.5%	0.0%	0%
#2	39.0%	26.3%	0.0%	0%
#3	23.7%	4.2%	14.3%	0%
#4	0.7%	9.5%	0.0%	0%
#5	31.0%	21.5%	0.0%	0%
#6	5.6%	22.9%	0.0%	0%
#7	2.6%	22.0%	33.3%	0%
#8	2.1%	14.7%	0.0%	0%
#9	11.0%	18.4%	0.0%	0%

Tabela 5.10: Tabela de erro relativo dos classificadores de micro ações - Autor (2023)

A avaliação de cada rede treinada não é direta, uma vez que a anotação de groundtruth e as previsões são realizadas numa base por quadro. No entanto, em muitos casos, não é claro o quadro exato onde uma micro ação começa ou termina. Consequentemente, erros de previsão

são esperados nas extremidades de cada execução de micro ação. Afirmamos que este não é um problema crítico, uma vez que apenas o número de ocorrências de micro ação são relevantes para calcular os descritores de vídeo. Assim, nos concentramos em avaliar como os descritores de vídeo gerados a partir de previsões de rede neural (δ_i) diferem dos descritores de vídeo de groundtruth calculados a partir de dados anotados por humanos, aqui denotados por γ_i, conforme descrito na Seção 4.3.3.

Especificamente, somamos o número de ocorrências de cada micro ação em todas as iterações do experimento de validação cruzada para os rótulos de groundtruth como $S_\gamma = \sum_{it=1}^{9} \gamma_{it}$, onde γ_{it} são rótulos de groundtruth do vídeo it, que é o vídeo usado como conjunto de teste na iteração it. Analogamente, $S_\delta = \sum_{it=1}^{9} \delta_{it}$. Note que S_γ e S_δ são ambos vetores cujos elementos são o número de ocorrências de uma micro ação específica, e assim os erros relativos para cada micro ação podem ser facilmente calculados a partir deles. Esses valores, juntamente com seus erros relativos, são revelados na Tabela 5.11.

Microação	**# Anotado**	**# Previsto**	**Erro relativo**
OLHO_FECHADO	791	725	8.34%
OLHANDO_CANTO	1719	1482	13.79%
MEXEU_LABIOS	23	23	0%
VIROU_ROSTO	1	1	0%

Tabela 5.11: Resultados para cada classificador de micro ação: os descritores de vídeo de groundtruth e previstos são comparados, revelando erro relativo baixo para todas as ações.

O número de ocorrências previstas para cada micro ação (S_γ) alcançou valores muito semelhantes em comparação com os valores de groundtruth (S_δ). O número de ocorrências de ambas as micro ações MEXEU_LABIOS e VIROU_ROSTO, que são ações menos frequentes, foi perfeitamente previsto, como representado por um erro relativo de 0%. Embora isso não signifique necessariamente que todos os quadros foram corretamente previstos, na prática isso indica que as redes neurais forneceriam dados precisos para a próxima etapa da metodologia proposta (análise estatística de dados).

No entanto, o desempenho dos modelos apresenta espaço para melhoria ao lidar com as ações OLHO_FECHADO e OLHANDO_CANTO . Essas ações registraram um erro relativo de 8.34% e 13.79% respectivamente. Vale ressaltar que os erros relativos ocorreram devido a um número subestimado de ocorrências, o que indica que os classificadores falharam em detectar algumas dessas duas micro ações (*i.e.* falsos negativos). Acreditamos que melhores resultados podem ser alcançados com um conjunto de treinamento maior, uma vez que apenas 8 vídeos de 3 minutos foram usados como conjunto de treinamento em nossos experimentos.

As discrepâncias observadas podem ser atribuídas a vários fatores, incluindo a complexi-

dade das próprias ações, a qualidade dos dados de treinamento, e potenciais vieses inerentes ao processo de treinamento. Por exemplo, ações relacionadas aos olhos podem ser mais desafiadoras para detectar com precisão devido à variabilidade nas condições de iluminação, posição do aluno em relação à câmera, e diferenças anatômicas individuais.

Consideramos que esses erros não são substancialmente altos, e assim a questão de pesquisa principal deste trabalho é respondida positivamente, uma vez que as redes neurais treinadas foram capazes de alcançar alta precisão para identificar ocorrências de micro ações em vídeos faciais de sessões de aprendizado multimídia. No entanto, há uma necessidade de otimização adicional, especialmente em relação às micro ações envolvendo os olhos.

Para a análise dos resultados também utilizamos algumas métricas como a acurácia, precisão, revocação, MSE(Mean Squared Error) , Mean Absolute Error(MAE) e AUC.

5.3.3 Resultados por micro ação

Para a análise dos resultados de cada classificador de micro ação, também foram coletadas as seguintes métricas por cada micro ação e interação, obtidas a partir das taxas de verdadeiros negativos (TN); verdadeiros positivos (TP); falsos positivos (FP); e falsos negativos (FN).

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{5.1}$$

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \tag{5.2}$$

$$\text{Precisão} = \frac{TP}{TP+FP} \tag{5.3}$$

$$\text{Revocação} = \frac{TP}{TP+FN} \tag{5.4}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{5.5}$$

$$\text{Sensividade} = Revocao = \frac{TP}{TP+FN} \tag{5.6}$$

$$\text{Especificidade} = \frac{TN}{FP+TN} \tag{5.7}$$

$$AUC = \int_0^1 !TPR(FPR^{-1}(t)), \mathrm{d}t \tag{5.8}$$

Nessa formula TPR é *true posivite rate* e *false positive rate*.

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}} \tag{5.9}$$

Nessa fórmula, x_i e y_i são as observações individuais das variáveis x e y, respectivamente, $\bar{x}$ e $\bar{y}$ são as médias de x e y, respectivamente, e n é o número de observações.

Resultados do Classificador Olho Fechado

A Tabela 5.12 apresenta um resumo dos resultados obtidos para a análise de dados de olho fechado em diferentes interaçoes. Os resultados são baseados nas métricas de precisão, revocação e f1-score, para as categorias 0 e 1 (Neutro e Ação).

Interação	TN	FP	FN	TP
1	66	6	0	71
2	43	5	3	44
3	39	9	21	26
4	79	71	77	72
5	56	3	50	8
6	106	14	14	105
7	1	75	75	0
8	1	94	94	0
9	1	65	65	0

Tabela 5.12: Resumo dos resultados para o classificador olho fechado

Int	Precisão(0)	Precisão(1)	Revocação(0)	Revocação(1)	F1-Score(0)	F1-Score(1)
1	0.52	0.92	1.00	1.00	0.68	0.96
2	0.52	0.90	0.93	0.94	0.67	0.92
3	0.55	0.74	0.65	0.55	0.59	0.63
4	0.65	0.50	0.51	0.48	0.57	0.49
5	0.51	0.73	0.53	0.14	0.52	0.23
6	0.53	0.88	0.88	0.88	0.66	0.88
7	0.99	0.00	0.01	0.00	0.03	0.00
8	0.99	0.00	0.01	0.00	0.02	0.00
9	0.98	0.00	0.02	0.00	0.03	0.00

Tabela 5.13: Metricas coletadas em cada um dos treinamentos do modelo, seguindo a estratégia *leave-one-video-out* para o classificador olho fechado.

Ao analisar os resultados dos classificadores em diferentes interaçoes na Tabela 5.13, podemos observar o desempenho variado em relação às métricas de precisão, revocação, f1-score e acurácia.

Interação 1: Nesta interação, o classificador apresenta um desempenho excelente. A precisão e o revocação são altos para ambas as categorias, indicando que o modelo tem uma alta taxa de acertos e uma baixa taxa de falsos negativos. Além disso, o F1-score, que é uma métrica que combina precisão e revocação, é alto para ambas as categorias, sugerindo que o classificador tem um equilíbrio entre precisão e revocação.

Interação 2: Na segunda interação, o classificador apresenta um desempenho moderado. Embora a precisão para a categoria 0 seja razoavelmente alta, a precisão para a categoria 1 é um pouco mais baixa. Isso sugere que o classificador tem mais dificuldade em identificar corretamente a categoria 1. O revocação e o F1-score para ambas as categorias são bastante altos, o que é uma boa indicação de que o classificador tem um bom equilíbrio entre precisão e revocação.

Interação 3: Na terceira interação, o classificador apresenta um desempenho misto. A precisão para a categoria 0 é razoavelmente alta, mas a precisão para a categoria 1 é significativamente mais baixa. Isso sugere que o classificador tem uma maior taxa de falsos positivos para a categoria 1. O revocação e o F1-score para a categoria 0 são razoavelmente bons, mas são significativamente mais baixos para a categoria 1. Isso indica que o classificador tem mais dificuldade em identificar corretamente a categoria 1.

Interação 4: Nesta interação, o classificador apresenta um desempenho abaixo do esperado. A precisão, revocação e F1-score para ambas as categorias são relativamente baixos, indicando que o classificador tem dificuldade em identificar corretamente ambas as categorias.

Interação 5: Na quinta interação, o classificador apresenta um desempenho misto. A precisão para a categoria 0 é relativamente alta, mas a precisão para a categoria 1 é bastante baixa. Isso indica que o classificador tem uma maior taxa de falsos positivos para a categoria 1. O revocação e o F1-score para a categoria 0 são razoavelmente bons, mas são significativamente mais baixos para a categoria 1, indicando que o classificador tem dificuldade em identificar corretamente a categoria 1.

Interação 6: Nesta interação, o classificador apresenta um bom desempenho, com alta precisão, revocação e F1-score para ambas as categorias. Isso sugere que o classificador tem um bom equilíbrio entre identificar corretamente ambas as categorias e minimizar o número de falsos positivos.

Interação 7: Nesta interação, o classificador apresenta um desempenho extremamente baixo. A precisão para a categoria 1 é de 0, indicando que todas as previsões dessa categoria foram falsas. O revocação e o F1-score para ambas as categorias são igualmente baixos, indicando que o classificador tem dificuldade em identificar corretamente ambas as categorias.

Interação 8: Semelhante a interação 7, o classificador na interação também apresenta um desempenho extremamente baixo. A precisão para a categoria 1 é de 0, sugerindo que todas as previsões dessa categoria foram falsas. O revocação e o F1-score para ambas as categorias são igualmente baixos, indicando que o classificador tem dificuldade em identificar corretamente ambas as categorias.

Interação 9: Como nas interaçoes 7 e 8, o classificador na interação apresenta um desempenho extremamente baixo. A precisão para a categoria 1 é de 0, indicando que todas as previsões dessa categoria foram falsas. O revocação e o F1-score para ambas as categorias são igualmente baixos, sugerindo que o classificador tem dificuldade em identificar corretamente ambas as categorias.

Em geral, os resultados dos classificadores variam dependendo da interação e das categorias em questão. Alguns interaçoes apresentam um desempenho excelente em todas as métricas, enquanto outros têm desempenho desigual ou inferior em uma ou ambas as categorias. Essa análise dos resultados fornece insights importantes sobre a capacidade do modelo em classificar corretamente os dados em diferentes interaçoes.

Resultados do Classificador Olhando para o Canto

O mesmo processo foi utilizado para o classificador de olhando para o lado/canto, o qual gerou os resultados da Tabela 5.14, que demonstra um resumo dos resultados obtidos para a análise de dados em diferentes interaçoes.

Interação	TN	FP	FN	TP
1	109	5	2	111
2	3	71	5	68
3	247	1	1	246
4	218	48	35	230
5	81	44	69	55
6	71	85	49	106
7	99	62	69	91
8	190	14	48	155
9	143	0	142	0

Tabela 5.14: Resumo dos resultados - Olhando para o Canto

Int	Precisão(0)	Precisão(1)	Revocação(0)	Revocação(1)	F1-Score(0)	F1-Score(1)
1	0.51	0.96	0.98	0.98	0.67	0.97
2	0.96	0.49	0.38	0.93	0.54	0.64
3	0.50	1.00	1.00	1.00	0.67	1.00
4	0.55	0.83	0.86	0.87	0.67	0.85
5	0.60	0.56	0.54	0.44	0.57	0.49
6	0.69	0.55	0.59	0.68	0.64	0.61
7	0.62	0.59	0.59	0.57	0.60	0.58
8	0.52	0.92	0.80	0.76	0.63	0.83
9	0.50	0.00	0.50	0.00	0.50	0.00

Tabela 5.15: Metricas coletadas em cada um dos treinamentos do modelo, seguindo a estratégia *leave-one-video-out* para o classificador olhando para o canto.

Interação 1: Nesta interação, a precisão da categoria 0 é moderada, enquanto a precisão para a categoria 1 é muito alta. O revocação e o F1-score para ambas as categorias é próximo de 1, indicando uma alta sensibilidade do modelo e um bom equilíbrio entre a precisão e o revocação. Além disso, as taxas de verdadeiros negativos (TN) e verdadeiros positivos (TP) são muito maiores do que as de falsos positivos (FP) e falsos negativos (FN), sugerindo uma boa performance do classificador.

Interação 2: Nesta interação, o classificador apresenta alta precisão para a categoria 0 e baixa para a categoria 1. O revocação é relativamente baixo para a categoria 0 e alto para a categoria 1. O f1-score é razoável para ambas as categorias. No entanto, a quantidade de falsos positivos é consideravelmente alta, o que prejudica a performance do modelo.

Interação 3: Nesta interação, o classificador apresenta um desempenho quase perfeito, com precisão e revocação de 1,00 para a categoria 1. Para a categoria 0, a precisão é moderada, mas o revocação e o F1-score são elevados. A quantidade de verdadeiros positivos e verdadeiros negativos é muito alta, indicando que o modelo teve uma performance excelente.

Interação 4: Nesta interação, a precisão para a categoria 0 é moderada, enquanto a precisão para a categoria 1 é alta. O revocação e o F1-score para ambas as categorias são elevados, indicando um bom equilíbrio entre precisão e revocação. A quantidade de verdadeiros positivos e verdadeiros negativos é maior do que a de falsos positivos e falsos negativos, sugerindo uma boa performance do classificador.

Interação 5: Nesta interação, tanto a precisão quanto o revocação são moderados para ambas as categorias, com a categoria 1 ligeiramente inferior à categoria 0 em todos os indicadores. O F1-score, que combina precisão e revocação, também é moderado para ambas as categorias. Isso indica que o classificador teve um desempenho equilibrado, mas com espaço para melhorias.

Interação 6: Nesta interação, o classificador apresenta precisão, revocação e F1-score moderados para ambas as categorias. A quantidade de verdadeiros positivos e verdadeiros negativos é ligeiramente menor do que a de falsos positivos e falsos negativos, indicando que o modelo poderia ser melhorado.

Interação 7: Nesta interação, a precisão, revocação e F1-score são moderados para ambas as categorias. A quantidade de verdadeiros positivos e verdadeiros negativos é maior do que a de falsos positivos e falsos negativos, sugerindo uma performance satisfatória do classificador.

Interação 8: Nesta interação, a precisão para a categoria 0 é moderada, enquanto a precisão para a categoria 1 é alta. O revocação e o F1-score para ambas as categorias são relativamente altos, indicando um bom equilíbrio entre precisão e revocação. A quantidade de verdadeiros positivos é alta, indicando que o modelo teve uma performance razoável.

Interação 9: Nesta interação, a precisão para a categoria 0 é moderada, mas para a categoria 1 é nula, indicando que o modelo não conseguiu classificar corretamente qualquer exemplo da categoria 1. O revocação é mediano para a categoria 0 e nulo para a categoria 1, enquanto o F1-score é mediano para a categoria 0 e nulo para a categoria 1. A quantidade de verdadeiros negativos é alta, mas não há verdadeiros positivos, sugerindo que o classificador tem um desempenho insatisfatório na interação.

Resultados do Classificador Mexeu Lábios

Nesta seção, examinaremos em detalhes os resultados obtidos pelo classificador Mexeu Lábios, explorando sua eficácia na identificação e categorização das micro ações faciais relacionadas ao movimento dos lábios durante a realização de atividades educacionais como podemos ver na tabela **??** e **??**.

Interação	TN	FP	FN	TP
1	1	1	1	1
2	7	1	0	7
3	6	1	1	5
4	3	0	0	2
5	1	0	0	0
6	2	0	1	0
7	5	0	2	2
8	2	0	1	0
9	1	0	0	0

Tabela 5.16: Resumo dos resultados dos Interações de treinamento do Classificador - Ação Mexeu Lábios

#Int	Precisão(0)	Precisão(1)	Revocação(0)	Revocação(1)	F1-Score(0)	F1-Score(1)
1	0.67	0.50	0.50	0.50	0.57	0.50
2	0.50	0.88	1.00	1.00	0.67	0.93
3	0.50	0.83	0.86	0.83	0.63	0.83
4	0.40	1.00	1.00	1.00	0.57	1.00
5	0.00	0.00	1.00	0.00	0.00	0.00
6	0.33	0.00	0.67	0.00	0.44	0.00
7	0.44	1.00	0.71	0.50	0.55	0.67
8	0.33	0.00	0.67	0.00	0.44	0.00
9	0.00	0.00	1.00	0.00	0.00	0.00

Tabela 5.17: Metricas coletadas em cada um dos treinamentos do modelo, seguindo a estratégia *leave-one-video-out* para o classificador mexeu lábios

Interação 1: Nesta interação, a precisão, revocação e F1-score para ambas as classes são 0,67 para a classe 0 e 0,50 para a classe 1, indicando um desempenho mediano do classificador na distinção entre as classes.

Interação 2: Nesta interação, a precisão, revocação e F1-score para a classe 0 são 0,50, 1,00 e 0,67 respectivamente. Já para a classe 1, esses valores são 0,88, 1,00 e 0,93, respectivamente, indicando um desempenho bastante robusto do classificador na distinção da classe 1, porém um pouco mais fraco para a classe 0.

Interação 3: Nesta interação, a precisão, revocação e F1-score para a classe 0 são 0,50, 0,86 e 0,63, respectivamente. Para a classe 1, os valores são 0,83 para precisão, revocação e F1-

score. Isso sugere um bom desempenho do classificador para a classe 1, embora a performance para a classe 0 seja inferior.

Interação 4: Nesta interação, a precisão para a classe 0 é de 0,40, enquanto o revocação e F1-score são 1,00. Para a classe 1, todas as métricas são perfeitas (1,00), indicando um excelente desempenho na classificação dessa classe, apesar da precisão baixa na classe 0.

Interação 5: Nesta interação, todas as métricas para ambas as classes são 0,00, indicando que o classificador não realizou nenhuma previsão correta para nenhuma das classes.

Interação 6: Nesta interação, a precisão, revocação e F1-score para a classe 0 são 0,33, 0,67 e 0,44, respectivamente. Para a classe 1, todas as métricas são 0,00. Isso indica um desempenho muito fraco do classificador para ambas as classes, sendo incapaz de classificar corretamente a classe 1.

Interação 7: Nesta interação, a precisão, revocação e F1-score para a classe 0 são 0,44, 0,71 e 0,55, respectivamente. Para a classe 1, a precisão é perfeita (1,00), o revocação é 0,50 e o F1-score é 0,67. Isso sugere um desempenho razoável do classificador na classificação de ambas as classes.

Interação 8: Nesta interação, a precisão, revocação e F1-score para a classe 0 são 0,33, 0,67 e 0,44, respectivamente. Para a classe 1, todas as métricas são 0,00, o que indica um desempenho muito fraco do classificador, sendo incapaz de classificar corretamente a classe 1.

Interação 9: Nesta interação, todas as métricas para a classe 0 são 0,00, o que indica que o classificador não foi capaz de fazer qualquer previsão correta para a classe 0. A classe 1 não apresenta instâncias na interação.

Resultados do Classificador Virou Rosto

Para esse classificador houve apenas um movimento, no vídeo 3, que foi previsto corretamente pelo classificador.

Interação 3: Nesta interação, observamos um desempenho consistente do classificador, com valores elevados para todas as métricas. Isso indica uma classificação correta tanto para a categoria 0 quanto para a categoria 1.

Análise de erros e limitações do modelo

Ao analisarmos a frequência temporal dos erros, percebemos que os erros estão concentrados no início e final dos vídeos, muito devido ao início e termino das vídeo aulas, com isso fizemos alguns cortes nos frames conforme Tabela 5.18.

ID	FRAMES	FRAME INICIAL	FRAME FINAL	FRAMES TOTAIS PÓS CORTE
1	5979	60	5623	5563
2	6269	88	5372	5284
3	6038	0	5633	5633
4	5720	0	5918	5918
5	6083	0	5683	5683
6	6083	0	5942	5942
7	5811	0	6022	6022
8	6184	0	5942	5942
9	6266	0	5943	5943

Tabela 5.18: Cortes executados nos frames.

Outra limitação que foi encontrada é quanto a utilização de acessórios faciais, como a máscara de proteção, que acaba atrapalhando o classificador por cobrir uma grande parte do rosto do indivíduo, tão quanto o modelo não consegue performar em resoluções muito baixas.

O modelo ao final apresenta um bom desempenho no que diz respeito ao erro relativo, que é a medida de precisão em relação ao número de ações encontradas. Isso significa que o modelo é capaz de identificar corretamente a maioria das ações em relação ao total de ações presentes em cada vídeo analisado. No entanto, é importante ressaltar que o modelo enfrenta alguns desafios ao detectar os frames específicos em que essas ações ocorrem. No caso deste experimento, a taxa de erro relativo é calculada para cada ação específica: "Olho Fechado", "Olhando para Canto", "Mexeu Lábios" e "Virou Rosto".

Analisando os dados de erro relativo para cada ação nos vídeos, podemos observar

algumas variações significativas. Por exemplo, no vídeo #1, o modelo apresentou um erro relativo de 1,4% na detecção da ação "Olho Fechado", o que indica uma boa precisão nesse caso. Porém, no vídeo #2, o erro relativo para a mesma ação foi de 39,0%, indicando que houve um número substancial de falsos positivos.

Além disso, podemos observar que algumas ações apresentam erros relativos altos na ação "Olhando para Canto". Por exemplo, no vídeo #1 possui um erro relativo de 31,5%, sugerindo que o modelo teve dificuldades na detecção dessa ação específica. No entanto, é importante mencionar que em alguns casos, a mesma ação como no vídeo #3, o modelo obteve um erro relativo de 4,2%, indicando uma detecção precisa.

Esses resultados destacam a importância de uma análise mais aprofundada dos frames específicos em que ocorrem as ações, a fim de identificar as possíveis fontes de erro e melhorar o desempenho do modelo nesses casos. É fundamental considerar que a detecção de ações em vídeos é um desafio complexo, pois envolve a interpretação de informações visuais em sequências temporais. Portanto, aprimorar a precisão na detecção de frames específicos é uma área de melhoria importante para garantir um desempenho mais consistente e confiável do modelo em relação às ações analisadas.

Com base nos resultados extraídos da análise, podemos observar na tabela **??** algumas tendências e conclusões significativas, vamos calcular as médias e as variações das métricas de cada classificador e depois de forma geral.

Classificador / Métrica	**Precisão**	**Revocação**	**F1-Score**
Olho Fechado (0)	0.64 (0.52 - 0.99)	0.59 (0.01 - 1.00)	0.48 (0.02 - 0.68)
Olho Fechado (1)	0.65 (0.00 - 0.92)	0.55 (0.00 - 1.00)	0.47 (0.00 - 0.96)
Olhando p/ Lado/Canto (0)	0.67 (0.50 - 0.96)	0.65 (0.38 - 1.00)	0.61 (0.50 - 0.67)
Olhando p/ Lado/Canto (1)	0.69 (0.00 - 1.00)	0.64 (0.00 - 1.00)	0.66 (0.00 - 1.00)
Mexeu Lábios (0)	0.46 (0.33 - 0.67)	0.78 (0.50 - 1.00)	0.56 (0.44 - 0.67)
Mexeu Lábios (1)	0.52 (0.00 - 1.00)	0.43 (0.00 - 1.00)	0.45 (0.00 - 1.00)

Tabela 5.19: Métricas de classificadores - Média Variação

Para o classificador "Olho Fechado", a média geral de precisão é 0.64 (com uma variação de 0.52 - 0.99), revocação é 0.59 (com uma variação de 0.01 - 1.00), e F1-Score é 0.48 (com uma variação de 0.02 - 0.68). Estes resultados mostram que, apesar do modelo ser razoavelmente preciso na identificação dos casos de "Olho Fechado", a sua capacidade de revocação, isto é, de identificar todos os verdadeiros positivos é mais baixa. Além disso, a média geral do F1-Score, que combina precisão e revocação, sugere que há espaço para melhorias significativas neste modelo.

Em contraste, para o classificador "Olhando para o Lado/Canto", a média geral de

precisão é de 0.67 (com uma variação de 0.50 - 0.96), revocação é de 0.65 (com uma variação de 0.38 - 1.00), e F1-Score é de 0.61 (com uma variação de 0.50 - 0.67). Este classificador apresenta resultados ligeiramente superiores em todas as métricas em relação ao "Olho Fechado". No entanto, ainda apresenta espaço para aprimoramento, especialmente na consistência de suas previsões, como evidenciado pelo amplo intervalo de variação em todas as métricas.

O classificador "Mexeu Lábios" apresentou a média geral mais baixa em todas as métricas, com precisão de 0.46 (variando de 0.33 a 0.67), revocação de 0.78 (variando de 0.50 a 1.00) e F1-Score de 0.56 (variando de 0.44 a 0.67). Isso sugere que este classificador tem dificuldades particularmente em termos de precisão, isto é, em evitar falsos positivos.

Em suma, enquanto esses classificadores mostram potencial para a detecção e classificação de expressões faciais, ainda há espaço significativo para melhoria. As áreas de aprimoramento incluem o aumento da precisão, aprimorando a capacidade do modelo de evitar falsos positivos; melhorando o revocação, aumentando a habilidade do modelo em identificar todos os verdadeiros positivos; e aumentando a consistência das previsões, como indicado pela variação nos resultados. Tais melhorias podem ser obtidas através da otimização dos modelos, da inclusão de mais dados de treinamento, ou do uso de técnicas de pré-processamento de imagens mais sofisticadas.

5.4 Análise de Dados Estatística

Nessa seção o foco é investigar se existem padrões relevantes relacionados a ocorrências de micro ações e o ganho no aprendizado; e se existem diferentes padrões de ocorrências de micro ações para diferentes tipos de aulas de aprendizado multimídia.

A principio vamos investigar a distribuição dos resultados da variável Ganho de Aprendizado, entre os dois tipos de teste, na Figura 5.9. Podemos verificar um espaçamento menor dos resultados do teste no grupo que recebeu a avaliação não redundante.

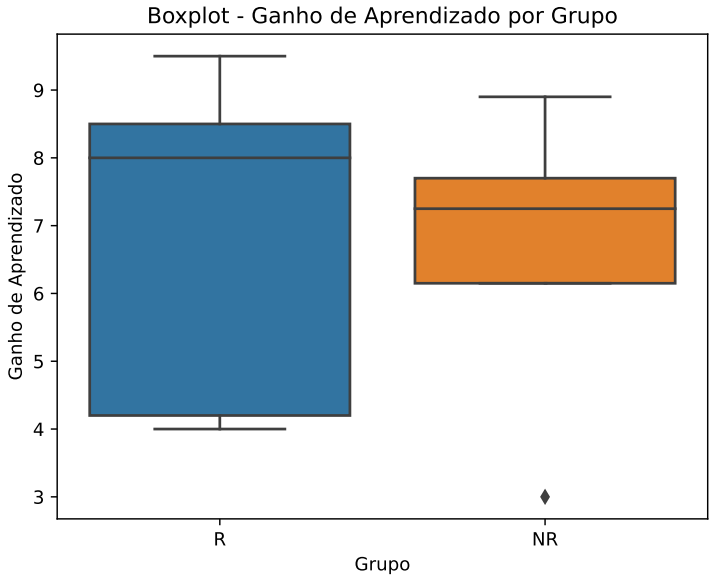


Figura 5.9: Boxplot por tipo de teste e ganho de aprendizado - Autor (2023)

A análise estatística do ganho de aprendizado, conforme apresentada na Tabela 5.20, revela insights importantes sobre o desempenho dos estudantes em diferentes tipos de aulas, especificamente as aulas redundantes (R) e não redundantes (NR). A média geral de ganho de aprendizado é de 6,73, indicando que, em média, os estudantes obtiveram um aumento de conhecimento após as aulas. No entanto, ao desagregar os dados, observa-se que as aulas redundantes têm uma média ligeiramente superior (6,84) em comparação com as aulas não redundantes (6,60). Esse achado sugere que as aulas redundantes podem estar associadas a um desempenho ligeiramente superior dos alunos. No entanto, é importante notar que a variação nos ganhos de aprendizado é considerável, como evidenciado pelo alto desvio padrão geral de 2,25. Isso implica que, apesar da média mais elevada nas aulas redundantes, há uma dispersão significativa nos resultados dos alunos, destacando a complexidade da avaliação de ganhos de aprendizado.

Métrica	Geral	R	NR
Média	6,73	6,84	6,60
Mediana	7,30	8,00	7,25
Desvio Padrão	2,25	2,29	2,19
Variação	5,05	5,24	4,77
Máximo	9,50	9,50	8,90
Mínimo	3,00	4,00	3,00
1º Quartil	4,20	4,20	6,15
3º Quartil	8,50	8,50	7,70
Amplitude Interquartil	4,30	4,30	1,55
Coeficiente de Variação (%)	33,37	33,47	33,11

Tabela 5.20: Análise Estatística do Ganho de Aprendizado por Tipo de Aula

Outro aspecto relevante é a análise do quartil, que fornece uma visão da distribuição dos dados. O primeiro quartil (Q1) das aulas não redundantes é notavelmente mais baixo (6,15) em comparação com as aulas redundantes (4,20), indicando que a maioria dos alunos nas aulas não redundantes teve um ganho de aprendizado mais substancial do que os alunos nas aulas redundantes nesse intervalo. No entanto, o terceiro quartil (Q3) das aulas redundantes e não redundantes é igualmente elevado (8,50), sugerindo que, para os alunos que se destacaram, não houve diferença significativa no ganho de aprendizado entre os tipos de aula. A amplitude interquartil, que mede a dispersão dos dados entre o primeiro e o terceiro quartil, revela uma diferença notável, sendo mais ampla nas aulas não redundantes (1,55) em comparação com as aulas redundantes (4,30). Isso implica que a variabilidade nos ganhos de aprendizado é maior nas aulas não redundantes, onde alguns alunos obtiveram ganhos substanciais, enquanto outros não tiveram tanto sucesso.

Finalmente, o coeficiente de variação (CV), expresso em porcentagem, reflete a variabilidade relativa em relação à média. O CV para as aulas redundantes (33,47%) e não redundantes (33,11%) é muito semelhante, indicando que, em termos relativos, a variabilidade nos ganhos de aprendizado é comparável entre os dois tipos de aula. No entanto, esse valor relativamente alto sugere que os dados têm uma dispersão considerável em relação à média, independentemente do tipo de aula. Essa análise estatística contribui significativamente para a compreensão da eficácia das estratégias de ensino em diferentes tipos de aula, destacando áreas de sucesso e áreas que podem exigir ajustes para otimizar o aprendizado dos alunos.

Após a analise do ganho de aprendizado, vamos aprofundar nossas analises para as ações, conforme o bloxplot na Figura 5.10 abaixo, no qual separamos as ações por tipo de aula.

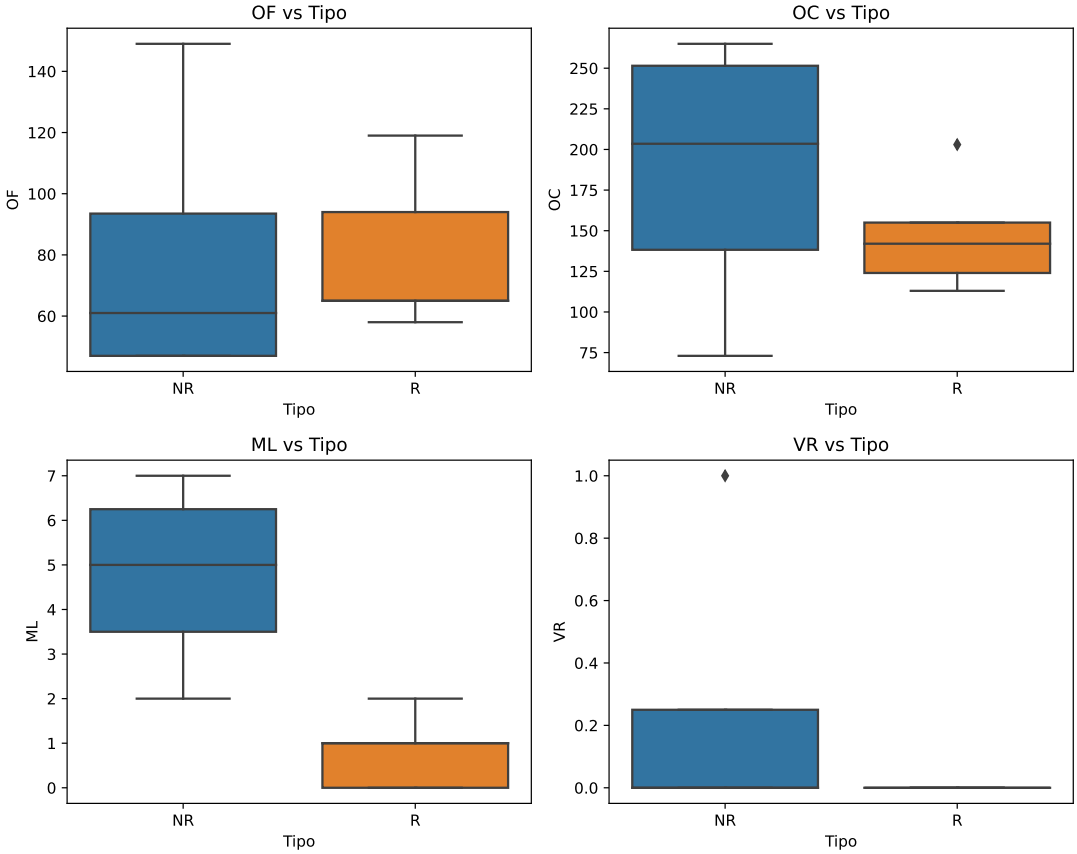


Figura 5.10: Boxplot por tipo de teste x ação - Autor (2023)

Como podemos verificar na Figura 5.10 acima, temos uma distribuição das micro ações bastante diferentes entre os tipos de prova R e NR, sendo o tipo de prova não redundante(NR), contendo uma distribuição mas esparsa durante o intervalo de ações analisadas.

Podemos analisar também a correlação das micro ações com a diferença das notas na avaliação, conforme revela a Figura 5.11 que analisa a correlação das ações com o ganho de aprendizado para todos os alunos do experiento. Com isso temos uma correlação moderada negativa nas ações OC e VR, com o ganho de aprendizado, ou seja, quanto mais ocorrências dessas ações menor o ganho de aprendizado do aluno.

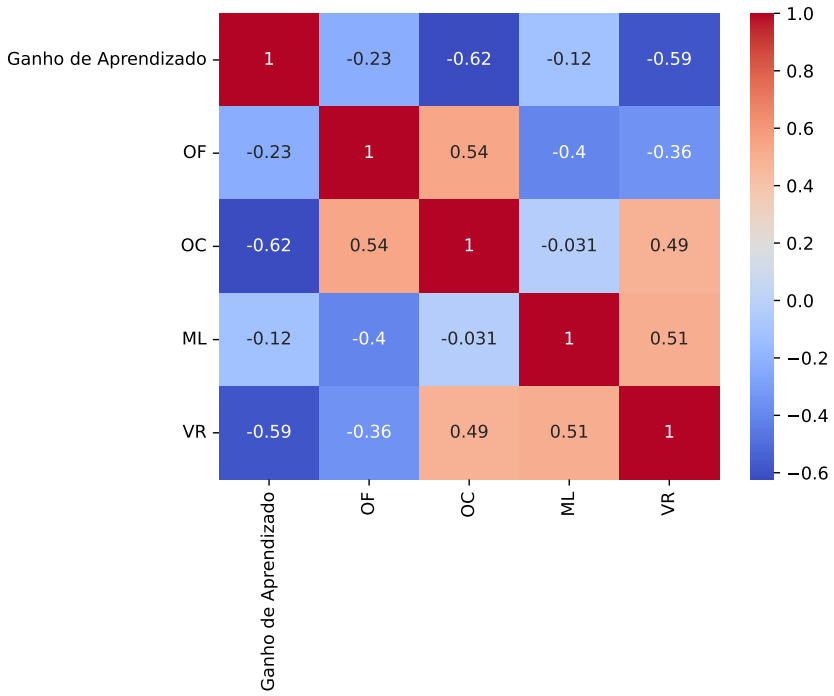


Figura 5.11: Correlaçoes entre as variáveis - Autor (2023)

5.4.1 Normalização dos Resultados no Tempo

A normalização das ações no tempo, ou seja, a conversão do número de ações para a taxa de ocorrência por segundo, desempenha um papel fundamental na análise das micro ações faciais. Ao dividir o número de ações, como olhos fechados (OF), pela duração do vídeo em segundos, obtemos uma medida mais precisa e comparável da frequência com que essas ações ocorrem ao longo do tempo. Essa normalização é essencial para evitar viés devido à variação na duração dos vídeos analisados, permitindo uma comparação justa entre diferentes amostras. Além disso, a taxa de ocorrência por segundo fornece insights valiosos sobre a intensidade e a frequência das micro ações faciais, auxiliando na compreensão de seu impacto e na alteração da nota. Ao normalizar as ações no tempo, somos capazes de extrair informações mais precisas e significativas sobre os padrões e comportamentos das micro ações faciais durante as atividades educacionais, contribuindo para uma análise mais abrangente e confiável dos dados coletados.

Para executar uma análise mais aprofundada, normalizamos as ações pelo tempo, dividindo o número das ações pelo tempo do video, alcançando os seguintes valores expressos na Tabela 5.21 e no bloxplot da Figura 5.12.

Tabela 5.21: Dados Normalizados no Tempo

ID	OF/sec	OC/sec	ML/sec	VR/sec
1	0.3476	0.6043	0.0107	0.0000
2	0.4134	0.4078	0.0391	0.0000
3	0.2513	1.3209	0.0321	0.0053
4	0.7563	1.3452	0.0102	0.0000
5	0.3069	0.6561	0.0000	0.0000
6	0.6010	0.7828	0.0051	0.0000
7	0.3750	0.8000	0.0200	0.0000
8	0.4747	1.0253	0.0051	0.0000
9	0.3283	0.7172	0.0000	0.0000

Ao analisar os dados fornecidos na tabela 5.21, podemos observar algumas informações relevantes abaixo e visualizar de forma gráfica na figura 5.12:

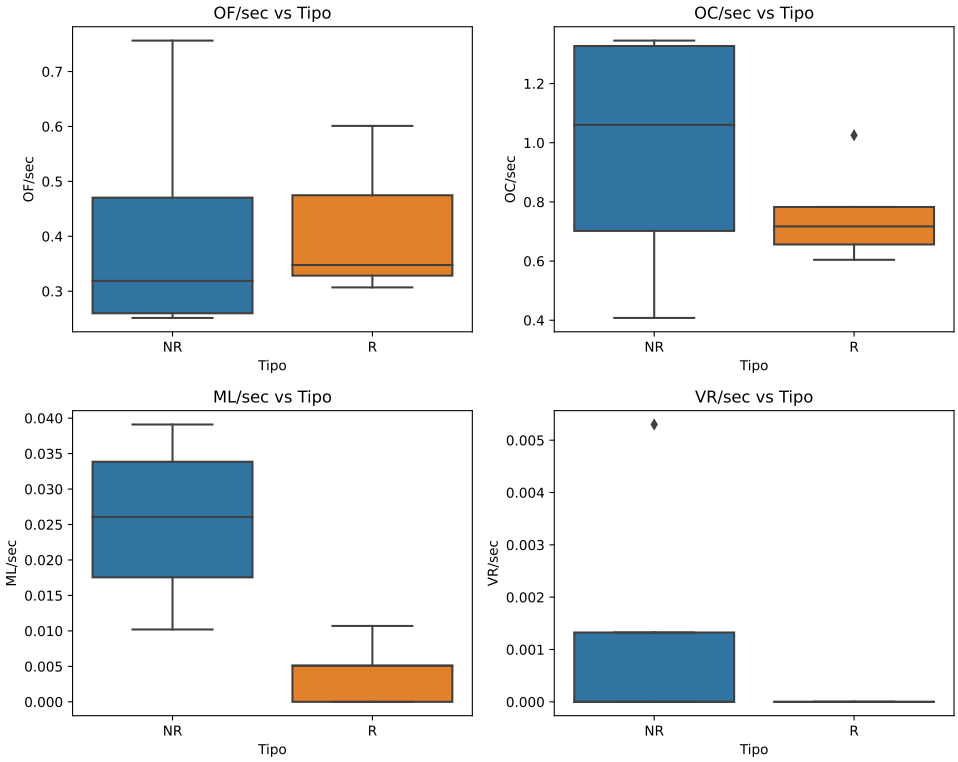


Figura 5.12: Boxplot por Tipo de Teste por Segundo - Autor (2023)

A taxa de ocorrência de Olho Fechado (OF) por segundo varia entre 0,3069 e 0,7563, com uma média de 0,4419. A taxa de ocorrência de Olhando Canto (OC) por segundo varia entre 0,4078 e 1,3452, com uma média de 0,7107. A taxa de ocorrência de Movimento Labial (ML) por segundo varia entre 0,0051 e 0,0391, com uma média de 0,0164. A taxa de ocorrência de Virou de Rosto (VR) por segundo apresenta valores ausentes (representados por "-"), o que indica a falta de dados disponíveis para essa variável em alguns casos.

É possível observar uma maior variabilidade nas taxas de ocorrência de Olhando Canto (OC) em comparação com as demais variáveis. Alguns participantes apresentam uma taxa de ocorrência de Virou de Rosto (VR) igual a zero, indicando que eles não realizaram essa ação durante o período analisado. Os participantes de tipo R parecem apresentar uma tendência de taxas de ocorrência de Olho Fechado (OF) e Olhando Canto (OC) mais baixas em comparação com os participantes de tipo NR. O participante de ID 3, do tipo NR, tem uma taxa de ocorrência de Olhando Canto (OC) por segundo muito alta em comparação com os outros participantes.

Para completar nossa análise, na Figura 5.14 apresentamos a nova correlação com os dados normalizados no tempo, o qual podemos verificar que houve pouca alteração nas correlações entre as variáveis.

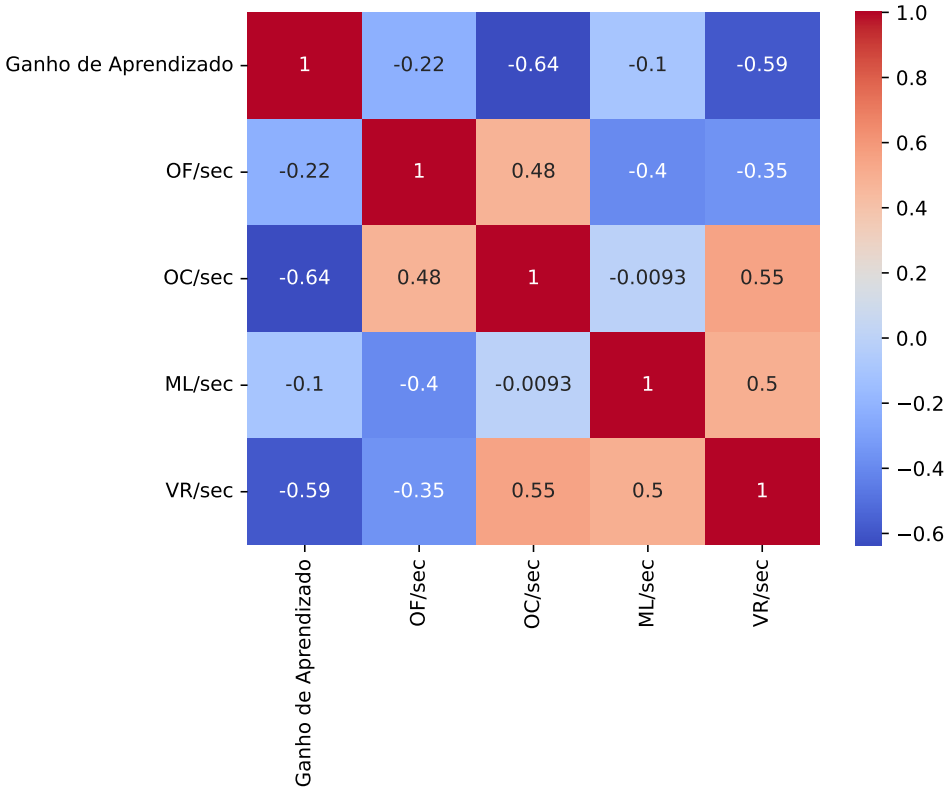


Figura 5.13: Correlaçoes entre as variaveis no tempo - Autor (2023)

5.4.2 Produtos de variáveis

A relação entre os produtos das ocorrências de micro ações normalizadas no tempo, como OF por Seg (olho fechado por segundo) e OC por Seg (olhar para o canto por segundo), e o cruzamento dos dados é um aspecto crucial para aprimorar as análises das micro ações faciais. Ao multiplicar a taxa de ocorrência de uma ação pelo tempo, obtemos uma medida mais abrangente que leva em consideração tanto a frequência quanto a duração das ações. Por exemplo, ao calcular o produto de OF por Seg e OC por Seg, estamos explorando a interação

entre essas duas ações específicas, permitindo-nos identificar padrões mais complexos e entender melhor o comportamento facial durante a atividade multimidia.

Essa abordagem de cruzar os dados, como OF por Seg x OC por Seg, amplia a compreensão das relações entre diferentes micros ações faciais e como elas se correlacionam com a nota da prova. Esses insights combinados nos permitem obter uma visão mais holística dos participantes, contribuindo para uma análise mais refinada e detalhada dos dados obtidos.

Essa abordagem enriquece as análises, fornecendo insights valiosos sobre as interações entre as ações faciais e seu impacto, auxiliando na identificação de padrões significativos. Ao analisarmos a correlação das variáveis dos dois grupos juntos temos os resultados revelados na Figura 5.14.

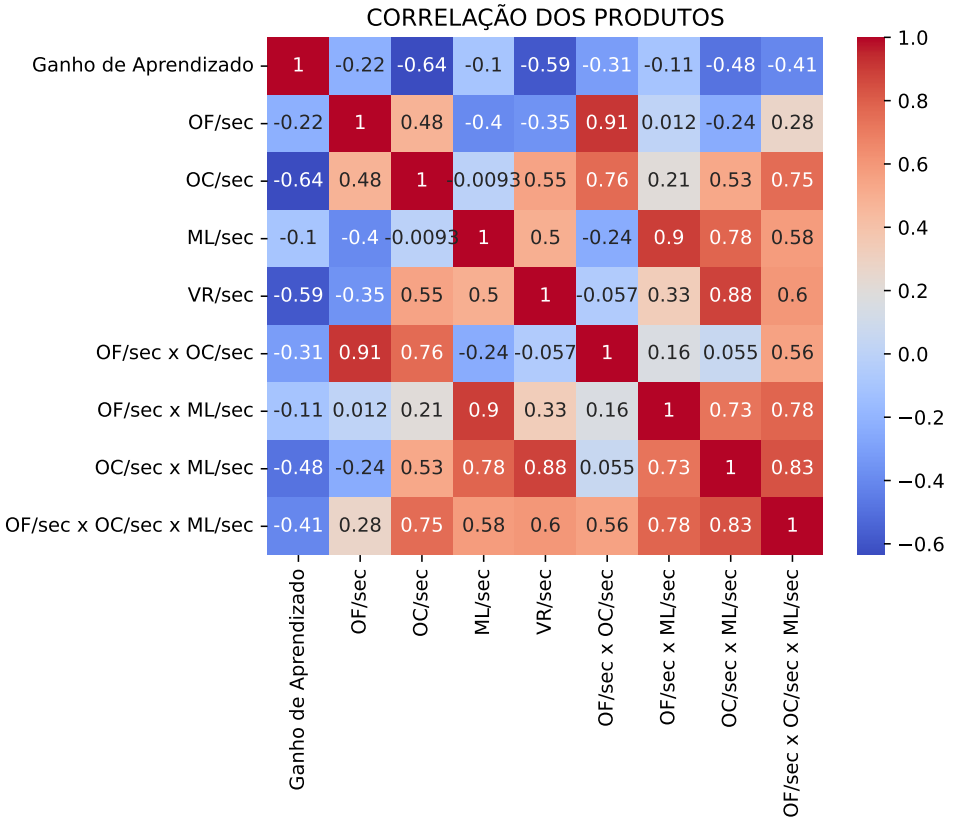


Figura 5.14: Correlações entre as variáveis normalizada no tempo - Autor (2023)

5.4.3 Grupo R

Ao analisar os dados separadamente para os grupos R e NR, podemos observar algumas informações adicionais relevantes. Podemos verificar na Figura 5.15, alguns pontos relevantes como a variável "Ganho de Aprendizado" que apresenta uma forte correlação negativa com as variáveis "OF/sec" e "OC/sec", indicando que um aumento no valor de "Ganho de Aprendizado" está associado a uma diminuição nas taxas de ocorrência dessas ações. A variável "ML/sec"

mostra uma correlação positiva moderada com as variáveis "OF/sec" e "OF/sec x OC/sec x ML/sec", sugerindo que há uma relação entre a taxa de ocorrência de "ML/sec" e o aumento nas taxas de "OF/sec" e no produto dessas três variáveis. As variáveis "OF/sec x OC/sec x ML/sec", "OF/sec x OC/sec" e "OF/sec x ML/sec" têm uma forte correlação negativa com "Ganho de Aprendizado", indicando que à medida que "Ganho de Aprendizado" aumenta, o produto dessas variáveis diminui. As variáveis "OF/sec x OC/sec x ML/sec" e "OF/sec x OC/sec" têm uma forte correlação negativa entre si, sugerindo que essas duas variáveis estão inversamente relacionadas. É importante destacar que a variável "VR/sec" apresenta valores ausentes (NaN) em toda a matriz de correlação, indicando que não há dados disponíveis para calcular a correlação dessa variável com as demais.

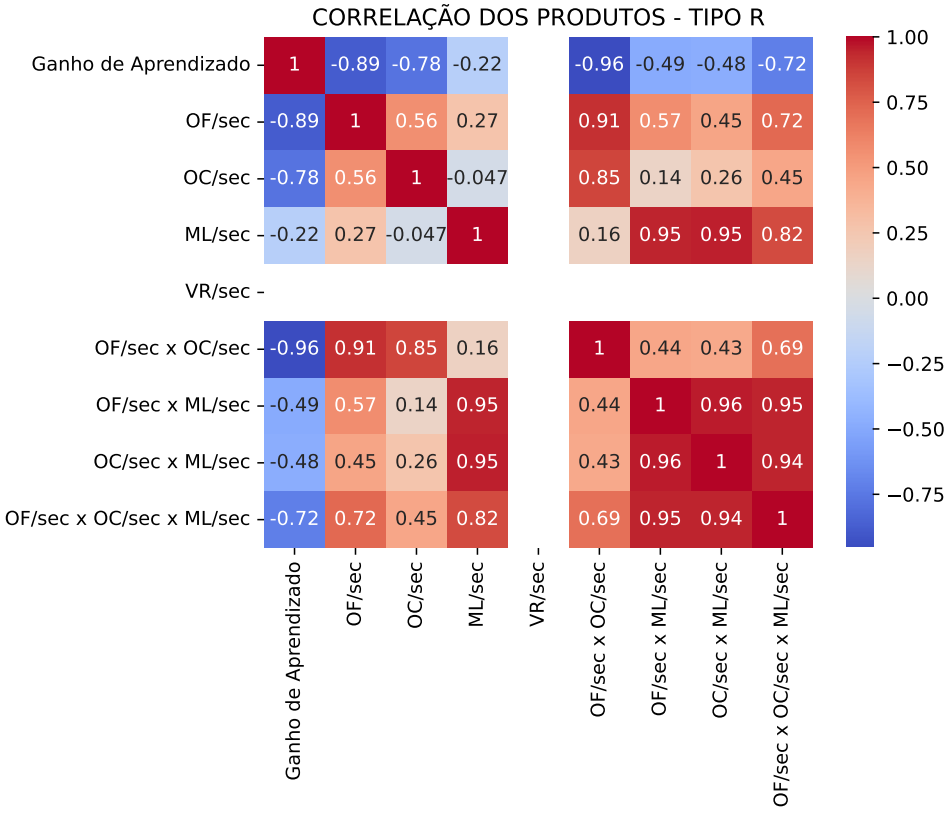


Figura 5.15: Correlações entre as variáveis do tipo R - Autor (2023)

5.4.4 Grupo NR

A variável "Ganho de Aprendizado" apresenta uma correlação positiva moderada com as variáveis "OF/sec" e "OF/sec x ML/sec", indicando que um aumento no valor de "Ganho de Aprendizado" está associado a um aumento nas taxas de ocorrência dessas ações. A variável "OF/sec" mostra uma correlação positiva moderada com as variáveis "OF/sec x OC/sec x ML/sec" e "OF/sec x OC/sec", sugerindo que há uma relação entre a taxa de ocorrência de "OF/sec" e o aumento nas taxas dessas variáveis. A variável "OC/sec" apresenta uma correlação positiva

moderada com as variáveis "OF/sec x OC/sec x ML/sec" e "OC/sec x ML/sec", indicando uma relação entre a taxa de ocorrência de "OC/sec" e o aumento nas taxas dessas variáveis. A variável "ML/sec" mostra uma correlação negativa moderada com as variáveis "OF/sec" e "OF/sec x OC/sec x ML/sec", sugerindo que há uma relação inversa entre a taxa de ocorrência de "ML/sec" e o aumento nas taxas dessas variáveis. A variável "VR/sec" apresenta uma correlação negativa forte com a variável "Ganho de Aprendizado", indicando que um aumento em "VR/sec" está associado a uma diminuição em "Ganho de Aprendizado". É importante destacar que a variável "VR/sec" tem uma forte correlação positiva com a variável "OC/sec x ML/sec", sugerindo que há uma relação entre essas duas variáveis.

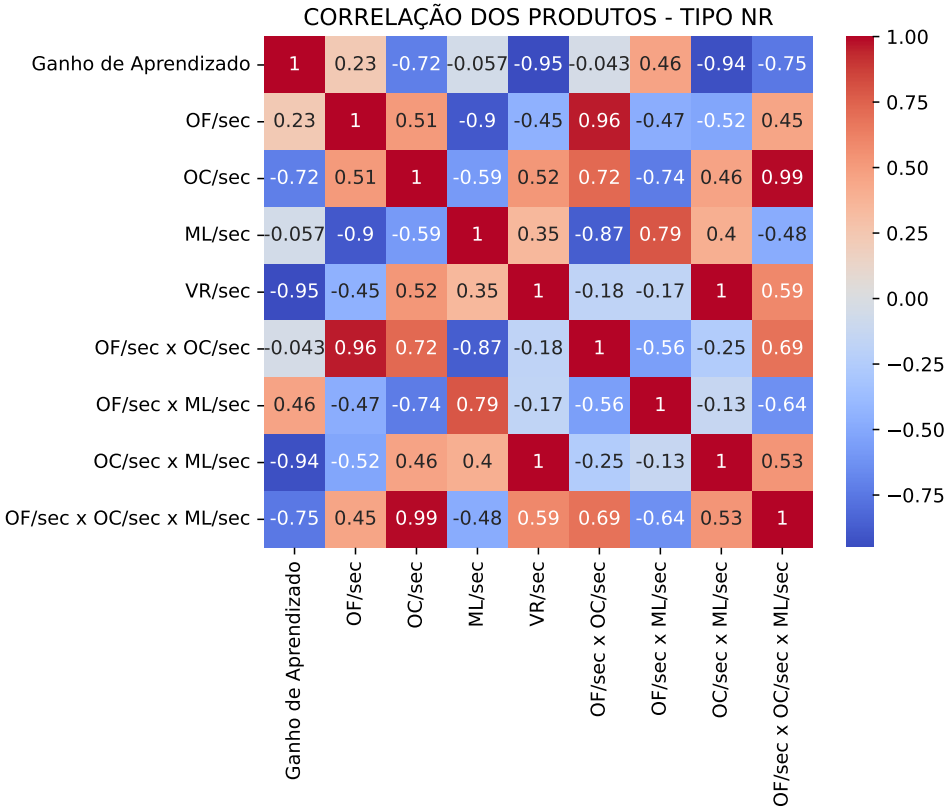


Figura 5.16: Correlações entre as variáveis do tipo NR - Autor (2023)

5.4.5 Discussão

A análise de dados estatística revelou uma correlação expressiva entre os desempenhos dos dois grupos na variável "Ganho de Aprendizado". Essa correlação insinua que as micro ações faciais podem desempenhar um papel fundamental na detecção de alterações nos resultados durante a realização de provas.

Essa conexão nos resultados é de grande importância, pois confirma a validade da detecção de micro ações faciais como prova de conceito. Durante a realização de tarefas educa-

cionais, essas micro ações podem ser ferramentas eficazes para identificar momentos de maior demanda cognitiva e dificuldade na aprendizagem.

Os achados realçam o potencial das micro ações faciais como um indicador não intrusivo e objetivo da carga cognitiva em ambientes de aprendizado multimídia. Para sua validação, é sugerido um estudo mais aprofundado com mais participantes e coleta de dados mais minuciosa e padronizada.

5.5 Limitações e desafios da detecção de micro ações faciais

Embora a detecção de micro ações faciais seja uma técnica promissora, ela enfrenta algumas limitações e desafios. A influência de fatores ambientais, como a iluminação e a qualidade do vídeo, pode afetar a precisão da detecção. Além disso, a variabilidade interpessoal na exibição de microações faciais pode representar um desafio adicional. A calibração e o treinamento do modelo para lidar com essa variabilidade são etapas críticas, mas que exigem um esforço considerável. Além disso, a necessidade de câmeras ou dispositivos para capturar as expressões faciais pode restringir a aplicabilidade em certos contextos ou ambientes. Superar essas limitações exigirá o desenvolvimento de técnicas robustas e a consideração cuidadosa dos requisitos técnicos e práticos para aplicação desse conceito em ambiente produtivo.

Capítulo 6

Conclusão e Trabalhos Futuros

Este estudo estabeleceu e avaliou modelos de redes neurais profundas para a classificação de micro ações faciais. Utilizando uma base de dados compilada a partir de vídeos de alunos capturados durante o processo de ensino, em tempo real, foi possível treinar classificadores LSTM (Long Short-Term Memory) e obter um desempenho promissor.

O classificador "Mexeu Lábios" demonstrou consistência, mas com uma média geral inferior comparada aos dados iniciais, com uma média de 0.46 para precisão em relação à classe 0 e 0.52 para a classe 1. O F1-Score médio foi de 0.56 para a classe 0 e 0.45 para a classe 1. Este resultado, embora mais baixo do que o esperado, ainda indica algum nível de eficiência do modelo em reconhecer movimentos labiais.

O classificador "Olhando para o Lado/Canto" apresentou uma média de precisão de 0.67 para a classe 0 e 0.69 para a classe 1. O F1-Score médio foi de 0.61 para a classe 0 e 0.66 para a classe 1. Esses valores demonstram um desempenho razoável do modelo, mas sugerem a necessidade de aprimoramentos.

Por fim, o classificador "Olho Fechado" apresentou a menor média entre todos os classificadores, com médias de 0.64 e 0.65 para precisão das classes 0 e 1, respectivamente, e F1-Scores médios de 0.48 para a classe 0 e 0.47 para a classe 1. Esse desempenho destaca a dificuldade do modelo em identificar corretamente quando um olho está fechado, o que pode ser influenciado por vários fatores, como a qualidade da imagem ou a sutileza do gesto.

Em média, a precisão, o recall e o F1-Score para todas as métricas foram um pouco inferiores a 0.64, o que sugere um certo grau de promessa nos modelos, mas também ressalta a necessidade de melhorias na aplicação de redes neurais para a predição de micro ações faciais localizadas no tempo.

Além disso a escolha do erro relativo como medida de avaliação do modelo, conforme apresentado na tabela 5.10, foi apropriada devido à sua capacidade de lidar efetivamente com dados proporcionais. Está métrica nos permitiu avaliar a habilidade do modelo em detectar

diversas micro ações faciais em diferentes contextos e cenários.

Observamos que a avaliação direta das redes treinadas apresentou desafios inerentes à natureza do problema, particularmente devido à dificuldade em determinar exatamente quando uma micro ação começa ou termina em cada quadro. No entanto, concluímos que essa imprecisão é menos problemática na prática, pois o objetivo primário é calcular os descritores de vídeo e, portanto, apenas o número total de ocorrências de micro ação é de interesse.

Os resultados resumidos na Tabela 5.11 mostram que a previsão do modelo para o número de ocorrências de cada micro ação foi surpreendentemente precisa, especialmente considerando as micro ações MEXEU_LABIOS e VIROU_ROSTO que foram previstas com perfeição. Embora este resultado não signifique necessariamente que todos os quadros individuais foram previstos corretamente, indica um alto grau de precisão na predição do número total de ocorrências para cada micro ação.

Em resumo, os dados demonstram que as redes neurais empregadas neste trabalho tiveram um desempenho eficaz na detecção de micro ações faciais, indicando a viabilidade de utilizar essas técnicas para análises posteriores na metodologia proposta. O trabalho também ressalta áreas potenciais para futuras melhorias e refinamentos dos modelos, a fim de aumentar ainda mais a precisão e a eficácia desses sistemas.

Além disso, nossa pesquisa também revelou descobertas intrigantes por meio da análise estatística dos dados. Encontramos padrões significativos relacionados à ocorrência de micro ações e ao ganho de aprendizado dos alunos, o que indica a estreita ligação entre a carga cognitiva experimentada pelos estudantes e sua capacidade de assimilar o conteúdo. Essa diferenciação sugere que a dinâmica da carga cognitiva pode variar dependendo do formato ou conteúdo da aula, abrindo novas oportunidades para a personalização do ensino e a otimização da experiência do aluno. Essas descobertas ressaltam a importância da análise de dados avançada em ambientes educacionais e indicam um caminho promissor para melhorar o ensino e a aprendizagem.

Nossos resultados representam um avanço significativo na validação de micro ações faciais como uma métrica útil no contexto educacional. Entendemos a necessidade de melhorar nosso processo de coleta de dados e aumentar a quantidade de amostras para fortalecer a validade de nossas descobertas.

Este trabalho deu origem a um manuscrito que está atualmente sob revisão para publicação na revista *Multimedia Tools and Applications*, demonstrando o potencial desta pesquisa para contribuir para o desenvolvimento do conhecimento nesta área.

Para futuras investigações, pretendemos explorar ainda mais a relação entre as expressões faciais e a carga cognitiva dos alunos, com o objetivo de desenvolver métodos para personalizar materiais de aprendizado. Acreditamos que esses esforços poderão revolucionar a educação personalizada, além de permitir uma aprendizagem otimizada, personalizada e baseada no estado

cognitivo individual do aluno.

Além disso, continuaremos a aprimorar nossa técnica de detecção de micro ações faciais e a superar os desafios e limitações identificados durante este estudo. Nosso objetivo é criar um sistema robusto, eficaz e não invasivo para medir a carga cognitiva em ambientes de aprendizagem, mantendo ao tempo um forte compromisso com as considerações éticas e de privacidade dos participantes.

Acreditamos que nosso trabalho atual irá impulsionar a pesquisa em expressões faciais e aprendizado multimídia, contribuindo para o desenvolvimento de um ambiente de aprendizagem mais engajador e eficaz. Nossas descobertas demonstram o potencial de nossa pesquisa e, ao mesmo tempo, apontam para áreas promissoras de futura investigação e aplicação.

Referências bibliográficas

BRUNKEN, R.; PLASS, J. L.; LEUTNER, D. Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, v. 38, n. 1, p. 53–61, 2003.

CHOLLET, F. *Deep Learning with Python*. 1st. ed. Greenwich, CT, USA: Manning Publications Co., 2017. ISBN 1617294438, 9781617294433.

CHOLLET, F. et al. *Keras*. [S.l.]: GitHub, 2015. <https://github.com/fchollet/keras>.

COOK, J. et al. Visualizing similarity data with a mixture of maps. In: *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*. [S.l.: s.n.], 2007. v. 2, p. 67–74.

CROWLEY, J. L.; CHRISTENSEN, H. I. *Vision as Process: Basic Research on Computer Vision Systems*. [S.l.]: Springer Science Business Media, 1994. v. 1. 435 p.

DARWIN, D. *The Expression of the Emotions in Man and Animals*. John Murray, 1871. v. 1. 378 p. Accessed: March 3, 2023. Disponível em: <http://play.google.com/books/reader?id=IFkIMZvfJy4C&hl=&printsec=frontcover&source=gbs_api>.

DELEEUW, K. E.; MAYER, R. E. A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 2008.

EKMAN, P.; FRIESEN, W. V. *Manual for the Facial Action Coding System*. [S.l.]: Consulting Psychologists Press, 1977.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. 2. ed. [S.l.]: LTC, 2021.

GHOTRA, B.; MCINTOSH, S.; HASSAN, A. E. Revisiting the impact of classification techniques on the performance of defect prediction models. In: IEEE. *2015 IEEE International Conference on Software Engineering*. [S.l.], 2015. p. 789–800.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. v. 1. 800 p.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 2005. In press.

HAMID, O.; MOHAMMED, O.; AKSASSE, B. Facial landmark localization: Past, present and future. In: *International Colloquium on Information Science and Technology (CIST)*. [S.l.: s.n.], 2016. p. 487–493.

HANSEN, D. W.; JI, Q. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence*, v. 32, n. 3, p. 478–500, 2010.

HUSSAIN, M.; CALVO, R.; CHEN, F. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. *Interacting with Computers*, Oxford University Press, v. 26, n. 3, p. 256–268, May 2014.

HUSSAIN, M. S.; CALVO, R. A.; CHEN, F. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. *Interacting with Computers*, v. 26, n. 3, p. 256–268, 2014.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998.

LUCEY, P.; LUCEY, S.; COHN, J. F. Registration invariant representations for expression detection. In: *International Conference on Digital Image Computing: Techniques and Applications*. [S.l.: s.n.], 2010. p. 255–261.

MAO, K. et al. A text classification model via multi-level semantic features. *Symmetry*, v. 14, p. 1938, 09 2022.

MARR, D. V. *A computational investigation into the human representation and processing of visual information*. Cambridge, MA: The MIT Press, 1982. v. 1.

MARTINS, R. X. A covid-19 e o fim da educação a distância: um ensaio. *Rede Revista de Educação a Distância*, v. 7, n. 1, p. 242–256, 2020.

MAYER, R. E. *Multimedia learning*. [S.l.]: Cambridge University Press, 2009. v. 3.

MAYER, R. E. *The Cambridge Handbook of Multimedia Learning*. [S.l.]: Cambridge University Press, 2014. v. 1. 950 p.

MCCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, v. 5, n. 4, p. 115–133, 1943.

MediaPipe. *MediaPipe Face Mesh*. 2020. <https://google.github.io/mediapipe/solutions/face_mesh.html>. Accessed on February 28, 2022.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of Machine Learning, second edition*. [S.l.]: MIT Press, 2018. v. 1. 504 p.

MURATA, A. An attempt to evaluate mental workload using wavelet transform of eeg. *Human Factors*, v. 47, n. 3, p. 498–508, 2005.

MURPHY-CHUTORIAN, E.; TRIVEDI, M. M. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 31, p. 607–626, 2009.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. [S.l.]: MIT Press, 2012.

OLIVEIRA, R. G.; ALVES, E. dos S.; MALQUI, C. Redes neurais convolucionais aplicadas ‘a preens ao rob’otica. In: *XXXVII Congresso Brasileiro de Inform’atica na Educaç ao-CBIE*. [S.l.: s.n.], 2017.

PAAS, F.; RENKL, A.; SWELLER, J. *Cognitive Load Theory: A Special Issue of Educational Psychologist*. [S.l.]: Routledge, 2016. v. 3. 80 p.

PANTIC, M.; ROTHKRANTZ, L. J. M. Expert system for automatic analysis of facial expression. *Image and Vision Computing*, v. 18, 2000.

PAUL, A. et al. The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, v. 12, 2021. Accessed on: Feb 2, 2022. Disponível em: <https://www.frontiersin.org/article/10.3389/fpsyg.2021.702538>.

REVINA, I. M.; EMMANUEL, W. R. S. A survey on human face expression recognition techniques. *Journal of King Saud University - Computer and Information Sciences*, Elsevier, v. 33, n. 6, p. 619–628, 2021.

ROSSUM, G. v. *Python Reference Manual*. [S.l.], 1995.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, IBM, v. 3, p. 210–229, 1959.

SANDLER, M. et al. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. Disponível em: <http://arxiv.org/abs/1801.04381>.

SANTOS, L.; TAROUCO, L. *A contribuição dos princípios da teoria da carga cognitiva para uma educação mediada pela tecnologia*. Tese — Programa de Pós-Graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul (UFRGS), 2015.

SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, v. 45, p. 2673–2681, 1997.

SMAGULOVA, K.; JAMES, A. P. Overview of long short-term memory neural networks. In: ____. *Deep Learning Classifiers with Memristive Networks: Theory and Applications*. Cham: Springer International Publishing, 2020. p. 139–153. ISBN 978-3-030-14524-8. Disponível em: <https://doi.org/10.1007/978-3-030-14524-8_11>.

SWELLER, J. Cognitive load during problem solving: Effects on learning. *Cognitive science*, Wiley Online Library, v. 12, p. 257–285, 1988.

SWELLER, J.; AYRES, P.; KALYUGA, S. *Cognitive Load Theory*. [S.l.]: Springer Science & Business Media, 2011. v. 3. 274 p.

SZELISKI, R. *Computer Vision: Algorithms and Applications*. [S.l.]: Springer Nature, 2022. v. 1. 925 p.

TIANLONG, Z. et al. Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *Journal of Educational Psychology*, American Psychological Association, v. 112, n. 7, p. 1338–1352, 2020.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2001.

WELLER, J. J.; MERRIëNBOER, G. van; PAAS, F. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, v. 31, p. 261–292, 2019.

WU, Y.; JI, Q. Facial landmark detection: A literature survey. *International Journal on Computer Vision*, 2016.

YANG, M.; KRIEGMAN, D.; AHUJA, J. Detecting faces in images: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 1, p. 34–58, 2002.

ZAGERMANN, J.; PFEIL, U.; REITERER, H. Measuring cognitive load using eye tracking technology in visual computing. In: *PROCEEDINGS OF THE SIXTH WORKSHOP ON BEYOND TIME AND ERRORS ON NOVEL EVALUATION METHODS FOR VISUALIZATION*. [S.l.: s.n.], 2016.

ZHANG, K.; AL. et. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, v. 23, n. 10, p. 1499–1503, 2016.

ZHOU, B.; GHOSE, T.; LUKOWICZ, P. Expressure: Detect expressions related to emotional and cognitive activities using forehead textile pressure mechanomyography. *Sensors*, v. 3, p. 730, 2020.