

UNIVERSIDADE FEDERAL DO ABC
CENTRO DE ENGENHARIA, MODELAGEM E CIÊNCIAS SOCIAIS APLICADAS
ENGENHARIA DE INFORMAÇÃO

Mayara Maria Triveloni

SISTEMA COM INTERFACE GRÁFICA PARA ANÁLISE DE EMOÇÕES EM
TEMPO REAL

Santo André, SP

2025

MAYARA MARIA TRIVELONI

**SISTEMA COM INTERFACE GRÁFICA PARA ANÁLISE DE EMOÇÕES EM
TEMPO REAL**

Trabalho de Graduação apresentado ao curso de Engenharia de Informação da Universidade Federal do ABC como requisito parcial para obtenção do título de Bacharel em Engenharia de Informação.

Orientador: Prof. Dr. Celso Setsuo Kurashima

Coorientador: Prof. Dr. Fernando Teubl Ferreira

Santo André, SP

2025

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me conduzir e capacitar-me para a realização do curso de Engenharia de Informação e aos meus pais, que sempre me apoiaram para que eu alcançasse com êxito meus objetivos e sonhos em relação a minha vida pessoal, acadêmica e profissional. Ao meu irmão, pelo incentivo e companheirismo ao longo de todo o processo.

Aproveito para agradecer ao Prof. Dr. Celso Setsuo Kurashima por todo o apoio e orientação concedida durante a elaboração do Trabalho de Graduação. Desde o instante inicial me auxiliou e me acompanhou pontualmente para a elaboração do projeto, sempre compartilhando ensinamentos valiosos ao longo do ano.

Ao Prof. Dr. Fernando Teubl Ferreira, pela coorientação concedida durante a elaboração do Trabalho de Graduação e por todo o conhecimento transmitido.

Aos demais professores do curso de Engenharia de Informação pela dedicação, paciência e participação no processo de aprendizado, contribuindo ativamente para o meu progresso acadêmico.

RESUMO

Com a impossibilidade de deslocamento e o isolamento social causados pela pandemia de Covid-19 no ano de 2020, as plataformas de videoconferências passaram a ser utilizadas por diversos profissionais como psicólogos e recrutadores. Apesar da facilitação gerada, a análise da comunicação não verbal tende a ser um desafio em alguns casos. Neste contexto, este estudo busca avaliar a aplicação de técnicas para identificar emoções através de abordagens independentes para a detecção de expressões faciais e a detecção de sinais de voz com fala emotiva de um indivíduo. Inicialmente, para o desenvolvimento do Trabalho de Graduação e visando alcançar adequadamente as detecções para aplicação em uma Interface Web Multimídia, modelos de redes neurais convolucionais foram utilizados. Para o treinamento, a metodologia adotada envolveu o uso de dados de imagens faciais e de sinais de voz, etapa de entendimento e pré-processamento dos dados, treinamento dos modelos de redes neurais seguido pela conversão para um formato compatível com a execução posterior em um navegador. Ademais, para a avaliação de desempenho, as métricas acurácias, precisão e revocação foram consideradas. Em sequência, uma interface Web foi idealizada contendo a integração dos modelos de detecção na plataforma de videoconferência (construída com a aplicação da ferramenta *Web Real-Time Communications*) e uma página de análise de relatórios (emoções capturadas e armazenadas durante as reuniões) com visualização gráfica. Por fim, com a realização de uma sessão experimental e com a análise de métricas da plataforma, observou-se que o desempenho obtido em tempo real tende a indicar resultados compatíveis com o objetivo proposto.

Palavras-chaves: Reconhecimento de Expressões Faciais. Reconhecimento de Emoção em Voz. Interface Web Multimídia.

ABSTRACT

With the impossibility of travel and the social isolation caused by the Covid-19 pandemic in 2020, video conferencing platforms began to be used by various professionals such as psychologists and recruiters. Despite the facilitation generated, the analysis of nonverbal communication tends to be a challenge in some cases. In this context, this study seeks to evaluate the application of techniques to identify emotions through independent approaches for the detection of facial expressions and the detection of voice signals with emotive speech of an individual. Initially, for the development of the Graduation Project and to advance to improve the detections for application in a Multimedia Web Interface, convolutional neural network models were used. For training, the methodology involves the use of facial image data and of voice signals, a data understanding and pre-processing stage, training of neural network models followed by conversion to a format compatible with subsequent execution in a browser. In addition, for performance evaluation, accuracy, precision, and recall scores were considered. Subsequently, a web interface was designed integrating the detection models into the video conferencing platform (built using the Web Real-Time Communications tool) and a report analysis page (emotions captured and stored during meetings) with graphical visualization. Finally, after conducting an experimental session and analyzing platform metrics, it was possible to assess that the real-time performance achieved was overwhelming and adequate for the proposed objective.

Keywords: Facial Expression Recognition, Speech Emotion Recognition, Multimedia Web Interface.

LISTA DE ILUSTRAÇÕES

Figura 1 - Funcionamento da Tecnologia WebRTC	10
Figura 2 - Representação de Onda	11
Figura 3 - Representação de Espectrograma	12
Figura 4 - Ilustração de Redes Neurais de Múltiplas Camadas	14
Figura 5 - Ilustração da Matriz de Confusão (Classificação Binária)	15
Figura 6 - Diagrama Geral: WebRTC e Interface dos Usuários	17
Figura 7 - Diagrama Incluindo o Sistema de Análise Emocional	19
Figura 8 - Esboço para Implantação	20
Figura 9 - Esboço do Processo do CRISP-DM	22
Figura 10 - Ilustração dos Ambientes e Componentes Necessários	25
Figura 11 - Ilustração do Cenário Experimental	26
Figura 12 - Diagrama para a Realização das Sessões Experimentais	27
Figura 13 - Prototipagem da Interface Web	29
Figura 14 - Página Inicial da Interface	32
Figura 15 - Chamada de Vídeo	33
Figura 16 - Página de Relatórios e Análises dos Pacientes/Candidatos	34
Figura 17 - Relatórios e Análises do Paciente/Candidato (Linha do Tempo)	35
Figura 18 - Amostra de Imagens do dataset FER2013	36
Figura 19 - Amostra de Imagens do dataset AffectNet	37
Figura 20 - Pré-processamento da Imagem Capturada	39
Figura 21 - Arquitetura para Rede Neural Convolucional (CNN)	44
Figura 22 - Resumo do Modelo Desenvolvido	44
Figura 23 - Matriz de Confusão (FER2013)	46
Figura 24 - Testando o Modelo com Capturas da webcam	47
Figura 25 - Matriz de Confusão (AffectNet)	49
Figura 26 - Gráfico de Comparação dos Desempenhos	49
Figura 27 - Gráfico de Comparação das Precisões (Classes)	50
Figura 28 - Amostra dos Sinais de Voz para cada Classe	52
Figura 29 - Espectrogramas de STFT	53
Figura 30 - Espectrograma de MFCCs (Coeficiente Cepstral de Frequência Mel)	54
Figura 31 - Resumo do Modelo para Sinais de Voz	55
Figura 32 - Matriz de Confusão (RAVDESS)	57
Figura 33 - Logotipo do Sistema (Meet.Emotions)	58
Figura 34 - Página de Acesso ao Profissional	58
Figura 35 - Página Inicial da Plataforma	59
Figura 36 - Seção "Sobre" da Plataforma do Profissional	60
Figura 37 - Seção "Calendário" da Plataforma do Profissional	61

Figura 38 - Planilha de Agendamentos	62
Figura 39 - Sala de Reunião Virtual do Cliente (Permissões)	62
Figura 40 - Menu Suspenso dos Botões	63
Figura 41 - Sala de Reunião do Cliente (Câmera e Microfone Ativados)	63
Figura 42 - Plataforma do Profissional (Cliente Aguardando)	64
Figura 43 - Sala de Reunião Virtual do Profissional	65
Figura 44 - Sala de Reunião Virtual do Profissional (Comunicação Iniciada)	66
Figura 45 - Sala de Reunião Virtual do Profissional (Detecção Facial)	66
Figura 46 - Sala de Reunião Virtual do Profissional (Detecção dos Sinais de Voz)	69
Figura 47 - Página de Relatórios e Análises	70
Figura 48 - Página de Relatórios e Análises (Linha do Tempo)	70
Figura 49 - Página de Relatórios e Análises (Frequência das Emoções)	71
Figura 50 - Faixa Etária dos Participantes	74
Figura 51 - Nível de Importância das Análises de Expressões Faciais	74
Figura 52 - Nível de Importância das Análises da Tonalidade de Voz	75
Figura 53 - Uso de alguma ferramenta/tecnologia para monitoramento emocional	77
Figura 54 - Sessão Experimental com o Profissional	80
Figura 55 - Frequência da Detecção Facial durante a Sessão Experimental	82

LISTA DE TABELAS

Tabela 1 - Distribuição de Imagens por Classe (FER2013)	36
Tabela 2 - Distribuição de Imagens por Classe (AffectNet)	37
Tabela 3 - Detecção Obtida para Imagem Capturada	39
Tabela 4 - Acurácia com a Aplicação do DeepFace	40
Tabela 5 - Métricas/Performance para cada Classe (FER2013)	41
Tabela 6 - Métricas/Performance para cada Classe (AffectNet)	41
Tabela 7 - Acurácia com a Aplicação do RMN	42
Tabela 8 - Métricas/Performance para cada Classe (AffectNet)	43
Tabela 9 - Métricas das Classes (CNN) para FER2013	45
Tabela 10 - Métricas das Classes (CNN) para AffectNet	48
Tabela 11 - Distribuição de Imagens por Classe (RAVDESS)	52
Tabela 12 - Métricas das Classes após o Treinamento (RAVDESS)	56
Tabela 13 - Benefícios de um sistema de análise das expressões faciais e da tonalidade de voz na área de Recursos Humanos (RH)	78
Tabela 14 - Benefícios de um sistema de análise das expressões faciais e da tonalidade de voz na área de psicologia	78
Tabela 15 - Preocupações ao utilizar um sistema de análise/interpretação emocional	79
Tabela 16 - Funcionamento do Sistema e Interface	81
Tabela 17 - Experiência de Uso da Plataforma/Sistema Profissional	81

SUMÁRIO

1 INTRODUÇÃO	3
1.1 Objetivos	5
1.2 Motivações	6
2 FUNDAMENTAÇÃO TEÓRICA	8
2.1 Comunicações Verbais e Não Verbais	8
2.2 Emoções Universais	9
2.3 Web Real-Time Communications (WebRTC)	9
2.4 Sinais de Áudio	10
2.4.1 Coeficientes Cepstral de Frequência Mel (MFCC)	12
2.5 Redes Neurais Artificiais	13
2.5.1 Redes Neurais Convolucionais	14
2.6 Métricas de Avaliação de Modelo	15
3 METODOLOGIA	17
3.1 Visão Geral do Sistema	17
3.2 Levantamento de Opinião por Questionário	21
3.2.1 Caracterização do Estudo e Coleta de Dados	21
3.2.2 Tratamento dos Dados	21
3.3 Desenvolvimento da Modelagem de Detecção Facial e de Voz	22
3.3.1 Entendimento do Negócio	23
3.3.2 Coleta e Entendimento dos Dados	23
3.3.3 Preparação dos Dados	24
3.3.4 Modelagem	24
3.3.5 Validação	24
3.3.6 Implantação	24
3.4 Interface Web para o Profissional e para o Usuário	24
3.4.1 Componentes e Ambientes Necessários	24
3.4.2 Descrição do Cenário Experimental	26
3.4.3 Sessões Experimentais com Voluntário	26
3.5 Avaliação de Desempenho por Questionário	27
4 IMPLANTAÇÃO DO SISTEMA	29
4.1 Estudo Inicial para uma Interface Web Idealizada	29
4.1.1 Detalhamento das Telas Idealizadas (Protótipo)	31
4.2 Classificação de Emoções: Expressões Faciais	35
4.2.1 Modelos pré-treinados: DeepFace	38
4.2.2 Modelos pré-treinados: Residual Masking Network	42
4.2.3 Arquiteturas para Redes Convolucionais Neurais	43
4.3 Classificação de Emoções: Sinais de Voz	50
4.4 Resultado da Interface Web Implementada	57

4.4.1 Métricas de Desempenho da Interface Web	71
5 ANÁLISE DOS RESULTADOS	73
5.1 Levantamento de Opinião dos Profissionais: Importância da Análise Emocional	73
5.2 Levantamento de Opinião após a Realização de uma Sessão Experimental do Sistema	80
6 CONCLUSÕES	84
REFERÊNCIAS	86
APÊNDICE A	89
APÊNDICE B	92
APÊNDICE C	96

1 INTRODUÇÃO

Em 11 de março de 2020, a Organização Mundial de Saúde (OMS) classificou o coronavírus (Covid-19) como pandemia e, com o objetivo de evitar a propagação do vírus, algumas medidas e estratégias foram adotadas para identificar, proteger, isolar e tratar os casos existentes (BRASIL, 2020).

Com o limite e impossibilidade de deslocamento, somado às incertezas que existiam na sociedade, houve um crescimento de aproximadamente 80% nos quadros ansiosos, de acordo com um estudo realizado na Universidade do Estado do Rio de Janeiro. Consequentemente, a busca por suporte emocional de maneira remota, ou seja, “teleterapia” foi a alternativa encontrada (GLOBO, 2021).

A alta demanda gerada pelo isolamento durante a pandemia entre o período de 2020 e 2021, resultou em um aumento de 352% de psicólogos cadastrados no serviço *e-Psi* (que listava, de acordo com a regulamentação de 2018, os profissionais autorizados pelo Conselho de Psicologia a prestarem atendimentos psicológicos virtualmente) (GLOBO, 2021).

De acordo com um estudo realizado pela Starbem (*healthtech* que oferece soluções de saúde mental para pessoas e empresas), entre 305 pessoas diagnosticadas com ansiedade e depressão, 87% que aderiram ao tratamento psicológico online notaram uma melhora significativa ao longo de seus tratamentos e, entre os outros resultados obtidos, 82% relataram uma melhora no humor e estabilidade emocional (MEDICINA, 2025).

Entre os sites, plataformas e aplicativos que trabalham com o atendimento terapêutico existe o Zenklub¹, que oferece o tratamento por um dispositivo de videoconferência e disponibiliza diferentes materiais auxiliares; o Psitto², que apesar de realizar as consultas com o auxílio de outros softwares populares criptografados, como o WhatsApp³ (versão *mobile*), centraliza em sua plataforma, o processo de agendamento, busca de profissionais e outros serviços; o PSI Psicólogo e Terapia⁴ que dispõe de atendimento remoto e presencial em São Paulo, mas assim como o Psitto, faz uso de outros meios para as videochamadas.

¹ Zenklub. Disponível em: <<https://zenklub.com.br/>>.

² Psitto. Disponível em: <<https://www.psitto.com.br/>>.

³ WhatsApp. Disponível em < <https://www.whatsapp.com/> >.

⁴ PSI Psicólogo e Terapia. Disponível em: < <https://www.psicologoeterapia.com.br/> >.

Do mesmo modo, outro setor que precisou se adaptar diante das consequências da pandemia foi a área de Recursos Humanos. Em resumo, as dinâmicas e entrevistas realizadas para o processo seletivo em uma empresa passaram a ser realizadas através de diferentes plataformas de videochamadas, como Google Meet⁵, Zoom⁶, ConferênciaWeb⁷, Microsoft Teams⁸ e entre outras.

Apesar da mudança ser benéfica em termos de agilidade e democratização de acesso, de acordo com uma especialista em recolocação e carreira, Tais Targa, alguns desafios precisaram ser enfrentados pelos recrutadores e candidatos:

Conheço muitas empresas que antes da pandemia faziam todos os processos seletivos de forma presencial, queriam sentir a energia dos candidatos, com dinâmicas de grupo, com entrevista presencial. Os profissionais que conseguiram se reinventar e ir para o virtual sem tanto sofrimento, se destacaram (PARIS, 2021).

Além disso, outra desvantagem para ambas as situações (atendimento psicológico e processo seletivo remotos) está na limitação da análise da comunicação não verbal. De acordo com o psiquiatra e coordenador do serviço de psicologia e neuropsicologia do IPq, Antonio de Pádua Serafim, a possibilidade de analisar as expressões faciais, comportamento e postura são pontos que precisam ser aprimorados em uma videochamada (GLOBO, 2021).

Em um contexto geral, a Ademicon em conjunto com a startup OrganIA, implementou uma Inteligência Artificial generativa que identifica as emoções dos seus clientes pelo tom de voz, saudação e palavras. A ferramenta recebeu o nome de “Rose” e analisa as interações feitas por chamadas eletrônicas para captar possíveis insatisfações ou descontentamento com o atendimento. Outro estudo publicado na revista *Frontiers in Psychology* comparou três modelos para o reconhecimento de emoções (alegria, raiva, tristeza, medo, nojo, neutralidade) em amostras de áudios de diferentes tonalidades e com durabilidade de 1,5 segundos (DIEMERLING, 2024).

Já para a detecção de emoções em imagens, o Google anunciou o *PaliGemma 2*, que pode não apenas identificar, mas também gerar legendas

⁵ Google Meet. Disponível em: <<https://meet.google.com/landing?pli=1>>.

⁶ Zoom. Disponível em: <<https://www.zoom.com/pt>>.

⁷ ConferênciaWeb. Disponível em: <<https://conferenciaweb.rnp.br/>>.

⁸ Microsoft Teams. Disponível em: <<https://www.microsoft.com/pt-br/microsoft-teams/log-in>>.

detalhadas sobre as pessoas nas fotos; outra teoria chamada “Sistema de Codificação de Ações Faciais” (*Facial Action Coding System*, FACS) é amplamente utilizada em casos policiais e de publicidade buscando traduzir as emoções do interlocutor, através das expressões faciais. Considerando sua capacidade para estes casos, este programa “*Face Reading*”, passou a ser utilizado em entrevistas de emprego após a chegada do software no Brasil - em junho de 2010 -, como explicado pelo empresário responsável, Wanderson Castilho: “O programa de reconhecimento facial gera relatórios de acordo com as emoções captadas e o recrutador vai analisar se as emoções relatadas condizem com o que o candidato fala” (EXAME, 2013).

Segundo um estudo realizado por Christopher Pramerdorfer e Martin Kampel, “*Facial Expression Recognition using Neural Networks: State of Art*”, as diferenças e impactos de desempenho foram destacados ao utilizarem CNNs (Redes Neurais Convolucionais) para o reconhecimento de expressões faciais baseado em imagens (PRAMERDORFER, 2016). Em resumo, usando para teste e treinamento o conjunto de dados FER2013 (mais detalhes na Seção 3.3.2) obteve-se uma acurácia de aproximadamente 75,2%, sem a necessidade de dados auxiliares.

Apesar dos diversos estudos e técnicas já existentes atualmente, ainda não há a disponibilidade de aplicativos ou softwares que ofereçam tais serviços para as áreas abordadas nos parágrafos anteriores (atendimento psicológico e recrutamento virtual), com uma interface que permita ao usuário um fácil acesso a uma ferramenta de detecção de emoções através de sinais de voz e expressões faciais, com a possibilidade de gerenciar todo o processo e que garanta a privacidade de seus dados. Diante desse cenário, é importante compreender as necessidades e as possibilidades para a implementação de um produto que busca analisar e interpretar emoções.

1.1 Objetivos

Investigar a aplicação de técnicas para identificar emoções através de abordagens independentes para a detecção de expressões faciais e a detecção de sinais de voz com fala emotiva de um indivíduo, durante uma entrevista de emprego ou um atendimento terapêutico remoto.

Além disso, a seguir são apresentados os objetivos específicos do trabalho de graduação:

a) Desenvolver modelos de redes neurais avançadas que apresentem bom desempenho/performance, garantindo a acurácia da detecção das emoções ao usuário, através de processamento de imagens para expressões faciais e de processamento de voz para fala emotiva;

b) Implementar uma interface Web que garanta ao usuário o acesso de um sistema com serviço de qualidade, de baixo custo, com tempo de processamento rápido e que transmita em tempo real, o sinal de vídeo e voz;

c) Fornecer um sistema que assegure aos respectivos usuários (psicólogos, recrutadores, pacientes e candidatos em uma entrevista de emprego) a privacidade dos dados.

1.2 Motivações

Para fundamentar e compreender as necessidades dos respectivos usuários da ferramenta, foi realizado um levantamento de opinião com uma amostragem de profissionais das áreas de Recursos Humanos e Psicologia. A Seção 5.1 apresenta todas as informações coletadas no levantamento de opinião, bem como suas respectivas análises.

Em suma, em uma escala de 1 a 5, em que 1 é pouco importante e 5 muito importante, 88,9% dos participantes consideraram que a análise das expressões faciais dos candidatos/pacientes durante uma entrevista de emprego ou atendimento é um fator muito importante para auxiliar a interpretação em uma comunicação não verbal. A mesma opinião foi obtida na pergunta seguinte, que avaliava o nível de importância da análise da tonalidade de voz durante um atendimento ou entrevista de emprego, em que 77,8% dos participantes consideraram como muito importante.

Com isso, evidencia-se a importância dessas análises (da expressão facial e tonalidade de voz) para as duas áreas de atuações, ainda mais quando estes mesmos usuários também destacam as dificuldades que podem surgir durante a interpretação das emoções, geradas em muitos casos pela ansiedade e nervosismo, e amplificados com a realização de entrevistas e atendimentos remotos.

Diante dessa necessidade, este trabalho apresenta o desenvolvimento de uma interface Web que auxilie o profissional na análise emocional durante uma reunião virtual, utilizando técnicas de visão computacional, processamento de áudio e modelagem, para classificar os sinais de voz capturados e as expressões faciais do participante, no momento que o usuário desejar (habilitando ou desabilitando as funções, quando necessário). Além disso, visando a entrega de uma solução eficiente, outras funcionalidades estarão disponibilizadas na plataforma, como o acesso a arquivos com o registro das emoções capturadas e a possibilidade de análise dos dados com a possibilidade de visualização de relatórios e gráficos.

Esta monografia está estruturada em seis capítulos. No Capítulo 2 os principais conceitos, técnicas e tecnologias serão abordados a fim de embasar a proposta do trabalho de graduação. No Capítulo 3, uma visão geral do sistema é apresentada, assim como, a metodologia e o processo adotado para o seu desenvolvimento. No Capítulo 4 a interface Web é idealizada, do mesmo modo que os modelos para classificação das expressões faciais e sinais de voz são construídos e apresentados. No Capítulo 5 são discutidos os resultados obtidos com os levantamentos de opinião aplicados ao público-alvo do produto. Por fim, o Capítulo 6 apresenta as conclusões e possibilidades para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção aborda os principais conceitos, métodos e tecnologias que sustentam a temática proposta para o trabalho de graduação, permitindo assim, compreender as escolhas e as etapas da pesquisa. Entre as teorias e definições apresentadas, Comunicações Verbais e Não Verbais e a Teoria da Universalidade das Emoções são essenciais para a contextualização da ideia do sistema; *Web Real-Time Communications*, apresentará a tecnologia utilizada para a captura e transmissão da mídia dos usuários; a seção Sinais de Áudio, reunirá fundamentos e estudos relacionados aos sinais de voz; por fim, os modelos de Redes Neurais Artificiais e Redes Neurais Convolucionais serão introduzidos, assim como as principais Métricas de Avaliação utilizadas.

2.1 Comunicações Verbais e Não Verbais

Para compreender os métodos e abordagens que serão adotados no decorrer do desenvolvimento do projeto, é essencial explorar alguns conceitos e fundamentos teóricos que contribuem com a direção da pesquisa. Segundo estudos realizados pelo antropólogo Albert Mehrabian, as palavras representam apenas 7% da comunicação, enquanto 55% representam a comunicação corporal e 38% com a comunicação paraverbal (MEHRABIAN, 1971).

Resumidamente, a comunicação verbal pode ser definida como uma mensagem transmitida através de palavras, ou seja, em uma linguagem falada ou escrita, com o interlocutor podendo estar presente ou não.

Por sua vez, a comunicação não verbal envolve tudo que não faz uso das palavras, logo, a informação é transmitida por meio da linguagem corporal, expressões faciais e tonalidade de voz, por exemplo. Para estes casos, a atenção visual tende a ser de extrema importância e entre as vantagens é tido a transmissão de emoções e a conexão entre os interlocutores/emissores.

2.2 Emoções Universais

De acordo com Paul Ekman (2003), psicólogo americano e pioneiro no estudo das emoções, expressões faciais e criador da Teoria da Universalidade das Emoções, as emoções podem ser definidas como um processo.

As emoções são um processo, um tipo particular de avaliação automática influenciada pelo nosso passado evolutivo e pessoal, no qual sentimos que algo importante para o nosso bem-estar está ocorrendo, e um conjunto de mudanças psicológicas e comportamentos emocionais começa a lidar com a situação (EKMAN, 2003).

Em suma, durante a realização de seu estudo sobre indivíduos de uma tribo em Papua Nova Guiné, em 1972, Ekman definiu sete expressões faciais universais de emoções, sendo elas:

- Alegria;
- Raiva;
- Tristeza;
- Medo;
- Nojo;
- Surpresa;
- Neutro.

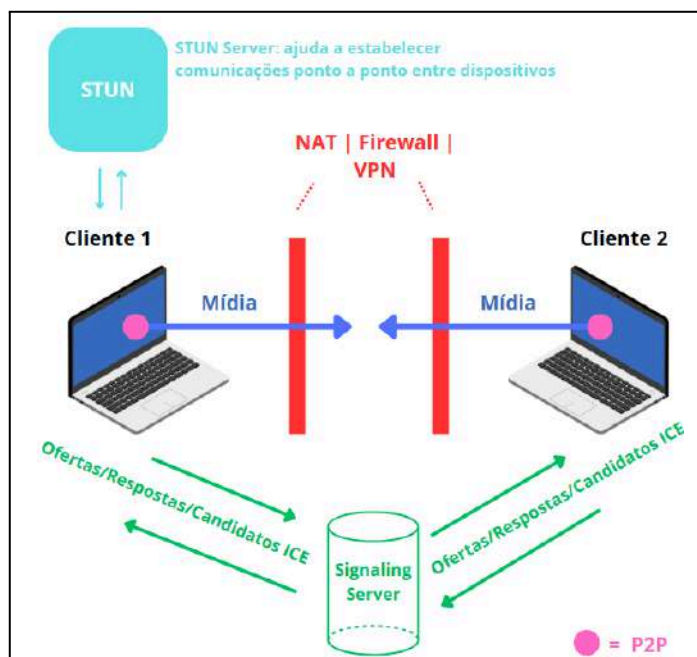
2.3 Web Real-Time Communications (WebRTC)

O WebRTC (*Web Real-Time Communications*) é uma tecnologia que permite que aplicativos e sites da Web capturem e, de maneira opcional, transmitam mídia de áudio e vídeo conferência, sem a necessidade de um intermediador durante uma possível troca de informações (MOZILLA, 2025a).

Basicamente, adota o mecanismo *peer-to-peer* (P2P – ponto a ponto) e este pode ser um tipo de arquitetura de rede distribuída em que cada nó - dispositivos conectados ao sistema -, cooperam e atuam também como servidores, diferente do que ocorre com o uso de cliente-servidor (que centraliza e concentra a informação) (MARCIANO *et al.*, 2018). Na Figura 1 é possível visualizar um esboço da estrutura

do WebRTC e na Seção 3.1 do trabalho, será descrito o funcionamento para o sistema.

Figura 1 - Funcionamento da Tecnologia WebRTC



Fonte: Elaborado pela autora.

Os motivos para o uso desta tecnologia podem ser compreendidos com os tópicos a seguir:

- Controle da ferramenta e flexibilidade;
- Segurança;
- Multiplataforma: compatibilidade com os principais navegadores da web;
- Custo: sem custos de licenciamento; incorporado ao *JavaScript*;
- Velocidade de processamento: o navegador se conectando diretamente para outro navegador.

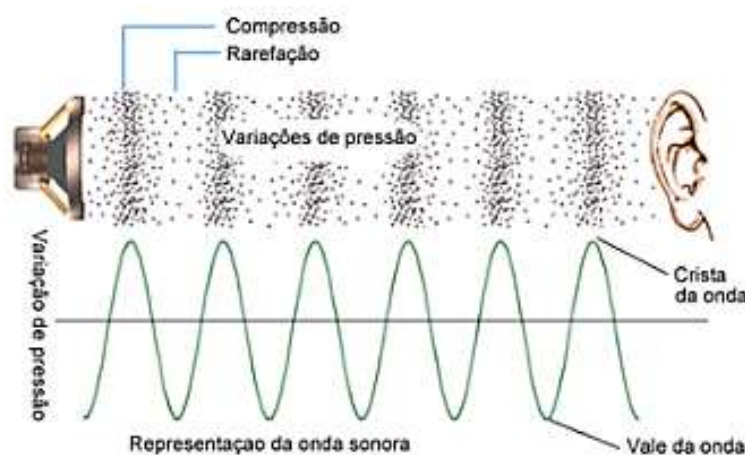
2.4 Sinais de Áudio

Resumidamente, de acordo com Müller, quando um objeto vibra as moléculas de ar tendem a oscilar para frente e para trás a partir de sua posição de repouso e transmitem sua energia para as moléculas vizinhas, com isso, se tem como

resultado a transmissão de energia de uma molécula para outra, que por sua vez, produz uma onda sonora.

Visualizando a Figura 2 é possível observar que as partes que o ar é empurrado para mais perto são chamados de compressões, e a situação contrária recebe o nome de rarefações. As ondas que atravessam o espaço utilizando compressões e rarefações recebem o nome de ondas longitudinais.

Figura 2 - Representação de Onda



Fonte: Adaptado de (CLARO, 2017).

Baseado na frequência é possível classificar os tipos de sons:

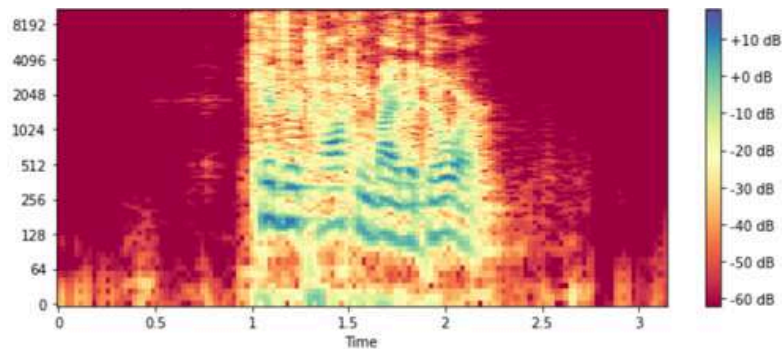
- Infrassom: Não perceptível ao ser humano. Sons com frequência abaixo de 20Hz;
- Ultrassom: Não perceptível ao ser humano. Sons com frequência acima de 20000Hz;
- Som audível: Perceptível ao ser humano. Sons com frequência entre 20Hz e 20000Hz.

Pensando na classificação de um som, o primeiro passo é extrair recursos/características que possam ser inseridos no algoritmo de aprendizado de máquina. Entre os recursos é tido o domínio do tempo, onde o sinal de áudio é representado pela amplitude em função do tempo, o que significa que as características são amplitudes registradas em diferentes intervalos, porém, sua

limitação é a ausência da informação sobre a taxa do sinal; por sua vez, no domínio de frequência o sinal é representado pela amplitude em função da frequência, logo, as características são as amplitudes registradas em diferentes frequências, porém, ignora a ordem ou a sequência do sinal, ou seja, sem informação do tempo.

Para obter as frequências dependentes do tempo, é possível utilizar um espectrograma, que pode ser definido como um gráfico 2D entre tempo e frequência, onde cada ponto do gráfico também representará a amplitude de uma determinada frequência em um determinado momento em termos de cores e intensidade, como apresentado na Figura 3.

Figura 3 - Representação de Espectrograma



Fonte: Elaborado pela autora.

2.4.1 Coeficientes Cepstral de Frequência Mel (MFCC)

A escala mel é uma escala de transformação não linear que transforma a faixa de frequência do áudio em uma faixa de valor diferente, como uma normalização dos dados, de modo que os sons de igual distância na escala Mel sejam percebidos como sendo de igual distância para os humanos (sendo o alcance da audição humana entre 20Hz e 20KHz) (MÜLLER, 2015).

A fórmula para a transformação da escala Hertz para a escala Mel pode ser visualizada na Eq. (2.1).

$$m = 1127 \log\left(1 + \frac{f(\text{Hz})}{700}\right) \quad (2.1)$$

O sinal de áudio é quebrado em quadros (janelas) sobrepostos, para cada quadro a transformada de Fourier será calculada (*Fast Fourier Transform*, FFT) e é gerado um banco de filtros (que se refere aos filtros mel), em seguida, um outro cálculo matemático é calculado (Transformada Discreta de Cosseno, TDC) e novamente, a Transformada de Fourier. Por fim, é tido como resultado os Coeficientes Cepstrais (MFCCs).

2.5 Redes Neurais Artificiais

Uma rede neural pode ser definida como um modelo de aprendizado de máquina que toma decisões de uma forma semelhante ao cérebro humano, com processos semelhantes a maneira que os neurônios biológicos trabalham juntos para o reconhecimento de padrões, como regressão (que tem como objetivo prever um valor numérico contínuo) e classificação que busca atribuir uma categoria a cada entrada (IBM, 2025a).

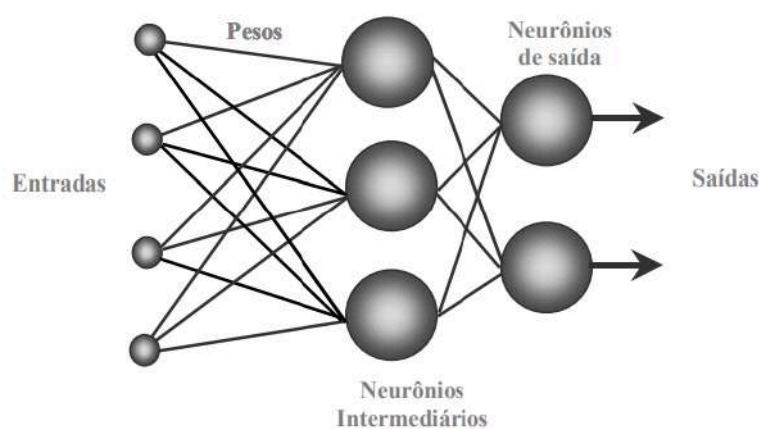
Basicamente, as redes neurais são formadas por neurônios e, sob a ótica computacional, é possível afirmar que em um neurônio é realizado o processamento sobre uma ou, geralmente, várias entradas, buscando gerar uma saída (FURTADO, 2019). Um dos maiores objetivos do treinamento de uma rede neural está em encontrar o menor conjunto de pesos para cada atributo, para assim, aproximar os resultados desejados.

Uma das primeiras e mais simples redes neurais artificiais criadas foram os Perceptrons, criada em 1958, por Frank Rosenblatt. No caso, ela é composta por apenas uma camada de entrada e uma de saída, contendo apenas um neurônio. Por sua vez, visando a resolução de situações mais complexas, os Perceptrons de Múltiplas Camadas (*Multilayer Perceptron*, MLP) são utilizados e, basicamente, os neurônios se dispõem em diversas camadas que podem ser classificadas em:

- Camada de entrada: onde os padrões são apresentados à rede;
- Camadas ocultas ou intermediárias: onde a maior parte do processamento é realizado;
- Camada de saída: onde a conclusão do processamento é apresentada.

Ao conectar-se em muitos neurônios buscando simular a captação de estímulos, uma série de saídas são geradas e, essas conexões transformam o sinal de saída de um neurônio em um sinal de entrada de outro. Visto que há a possibilidade de inúmeras conexões entre as camadas, diversas arquiteturas podem ser geradas (FURTADO, 2019). Na Figura 4 é possível observar uma ilustração representando uma rede neural de múltiplas camadas.

Figura 4 - Ilustração de Redes Neurais de Múltiplas Camadas



Fonte: Adaptado de (FURTADO, 2019).

2.5.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (*Convolutional Neural Network*, CNN) são extremamente utilizadas para aplicações em visão computacional e classificação de imagens, fala ou sinais de áudio, isso devido ao seu desempenho e capacidade de selecionar as melhores e essenciais características durante a extração.

De acordo com (HAYKIN, 2009), sobre a estrutura da rede para que o aprendizado apresente êxito se tem a extração de características, mapeamento de características e subamostragem. De modo geral, a camada de convolução necessita de dados de entrada, filtro e um mapa de funcionalidades. Ela é responsável pelo processo de adicionar cada elemento da imagem, por exemplo, para seus vizinhos, de maneira ponderada por um kernel (basicamente, aplicação de um filtro).

Em seguida, um mapa de características é gerado e, com isso, a dimensionalidade é reduzida e o processamento facilitado. Por sua vez, a próxima etapa/camada chamada *Pooling*, tende a enfatizar as características tidas como relevantes de cada mapa de característica e, conseqüentemente reduz ainda mais a dimensionalidade (a matriz gerada torna-se menor), diminui o custo computacional e reduz os ruídos desnecessários e o *overfitting* (situação que a rede neural se adapta demais aos dados de treinamento). Ao chegar na etapa de *Flattenning*, o formato da matriz é transformado em um vetor para então aplicar nas redes neurais.

2.6 Métricas de Avaliação de Modelo

Buscando medir o desempenho dos modelos de classificação a matriz de confusão foi utilizada. Em resumo, as colunas representam as classes reais, as linhas a classe prevista, os valores representam a frequência de ocorrência e, para gerá-la, basta utilizar a função disponível na documentação oficial do *Scikit-Learn* (SCIKIT-LEARN, 2025). Na Figura 5, uma ilustração é apresentada para um melhor entendimento do funcionamento da Matriz de Confusão para situações de classificações binárias, onde apenas duas classes são possíveis. Basicamente, a matriz permite visualizar como a rede neural treinada está classificando cada categoria por meio da aplicação de estatísticas de acertos e erros, facilitando a identificação de padrões. Na diagonal principal os acertos podem ser encontrados e nas demais células disponíveis os erros cometidos são disponibilizados.

Figura 5 - Ilustração da Matriz de Confusão (Classificação Binária)

		Classes Reais	
		+	-
Predições	+	TP	FP
	-	FN	TN

Fonte: Elaborado pela autora.

TP representa “Verdadeiro Positivo”, situação em que o classificador previu verdadeiro e a classe correta era verdadeiro; TN representa “Verdadeiro Negativo”, situação em que o modelo previu falso e a classe correta era falso; FP significa “Falso Positivo” e indica que o classificar previu verdadeiro, entretanto, a classe correta era falso; por fim FN ou “Falso Negativo”, quando o classificador previu falso e a classe correta é verdadeiro.

Além disso, algumas métricas são consideradas essenciais para a avaliação dos modelos como, por exemplo, a acurácia, precisão e *Recall*. Para a acurácia é possível defini-la como o percentual o qual os dados foram corretamente classificados. Para a realização do cálculo, a Eq. (2.2) pode ser observada.

$$Acurácia = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.2)$$

Por sua vez, a precisão considera a seguinte situação: “quando o modelo previu verdadeiro, o quanto a previsão estava correta?” Para o cálculo, a Eq. (2.3) é aplicada.

$$Precisão = \frac{TP}{TP+FP} \quad (2.3)$$

Já o *Recall* considera os valores FN, isto é: “quando a classe era verdadeira, o quanto o classificador estava correto?”. Para o cálculo, a Eq. (2.4) é utilizada.

$$Recall = \frac{TP}{TP+FN} \quad (2.4)$$

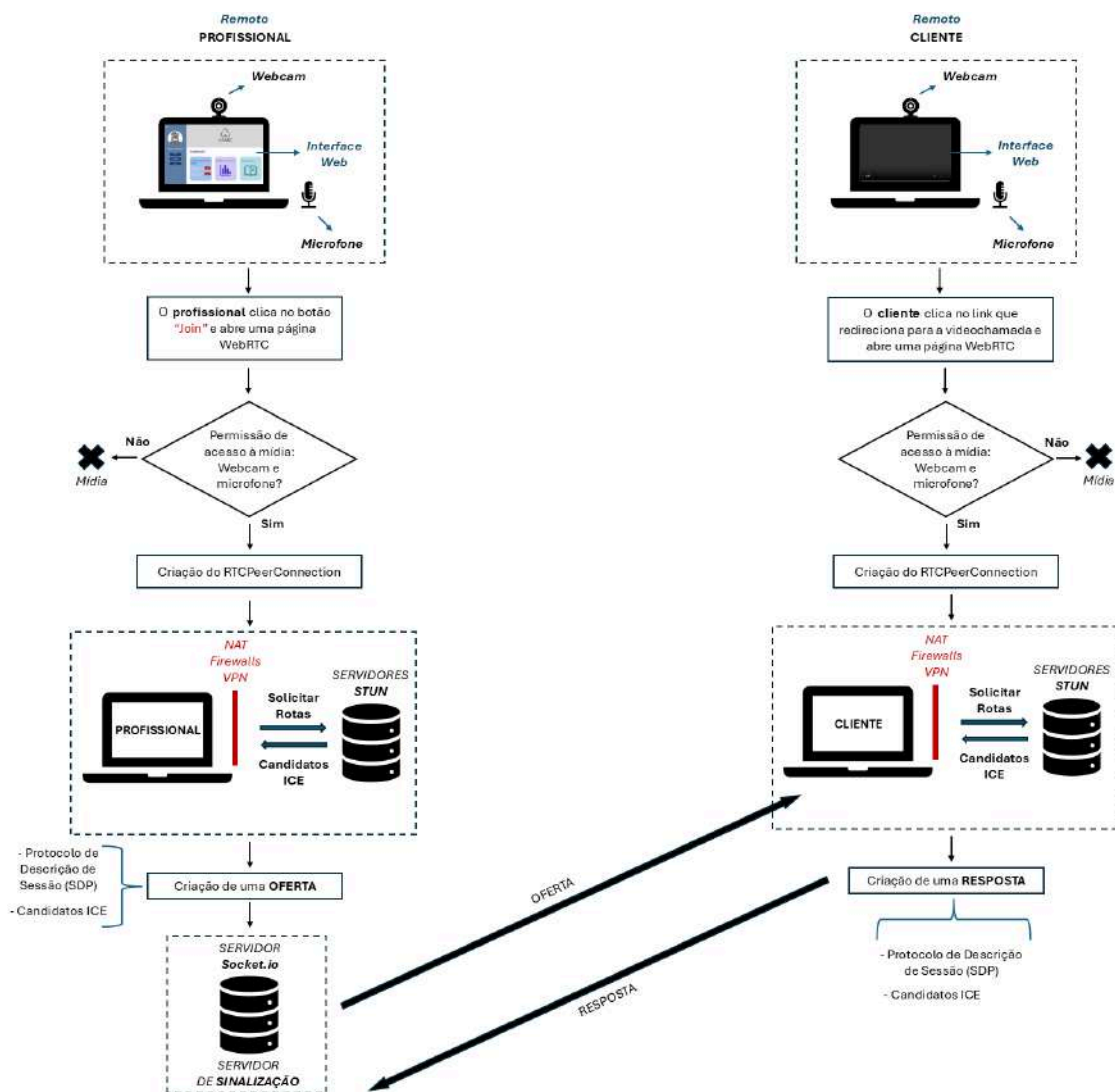
3 METODOLOGIA

Para o desenvolvimento do Trabalho de Graduação, foram previstas as seguintes atividades gerais para o projeto: (a) Levantamento de Opinião por Questionário; (b) Desenvolvimento da Modelagem de Detecção Facial e de Voz; (c) Interface Web para o Usuário; e (d) Avaliação de Performance por Questionário. Além disso, nos próximos parágrafos uma visão geral do sistema será apresentada.

3.1 Visão Geral do Sistema

Um diagrama apresentando o funcionamento da interface do usuário e do WebRTC a serem desenvolvidos é exibido na Figura 6.

Figura 6 - Diagrama Geral: WebRTC e Interface dos Usuários



Fonte: Elaborado pela autora.

Como é possível analisar, o profissional, após clicar no botão “Join” da Interface Web desenvolvida, abre uma página WebRTC. A Seção 4.1 apresenta os detalhes disponibilizados pelo protótipo deste trabalho. Em seguida, o navegador solicitará o acesso à mídia do usuário, isto é, a Webcam e o microfone. Caso o usuário permita o acesso, o *RTCPeerConnection*⁹ será criado.

O profissional precisa obter duas informações: 1) Onde o cliente está? e 2) O que está sendo enviado? Para isso, os servidores STUN (*Session Transversal Utilities for NAT*) são necessários para retornar algumas informações como endereço IP, status de conectividade e portas, mesmo com a existência de Firewalls, VPNs (*Virtual Private Network*) e NAT (*Network Address Translation*). Em sequência, serão retornados os candidatos ICE (*Interactive Connectivity Establishment*), que descrevem os protocolos e o roteamento necessário para que o WebRTC se comunique com um dispositivo remoto e, conseqüentemente, encontre a melhor conexão (MOZILLA, 2025b).

Em seguida, o profissional criará uma oferta contendo os fluxos de mídia, o SDP (Protocolo de Descrição de Sessão: que contém informações de conexão como codecs¹⁰/tipos de dados), e os candidatos ICE. Essa oferta é encaminhada para o servidor de sinalização, que basicamente irá armazenar e ajudar o navegador a encontrar e se comunicar com os clientes que estiverem disponíveis, por meio de um protocolo WebSocket (no caso, socket.io¹¹, visto que garante a segurança a transmissão da informação para o cliente).

Logo, as respostas para as questões levantadas anteriormente são: 1) Onde o cliente está? Resposta enviada: Candidatos ICE; 2) O que está sendo enviado? Resposta enviada: A oferta contendo o tipo e a informação.

Resumidamente, o cliente realizará o mesmo processo descrito anteriormente, mas ao invés de enviar uma oferta, enviará para o servidor uma resposta que terá um objeto/SDP contendo a informação. Com isso, o servidor transmitirá para ambos os lados (informações serão trocadas) e possibilitará a comunicação entre o profissional e cliente.

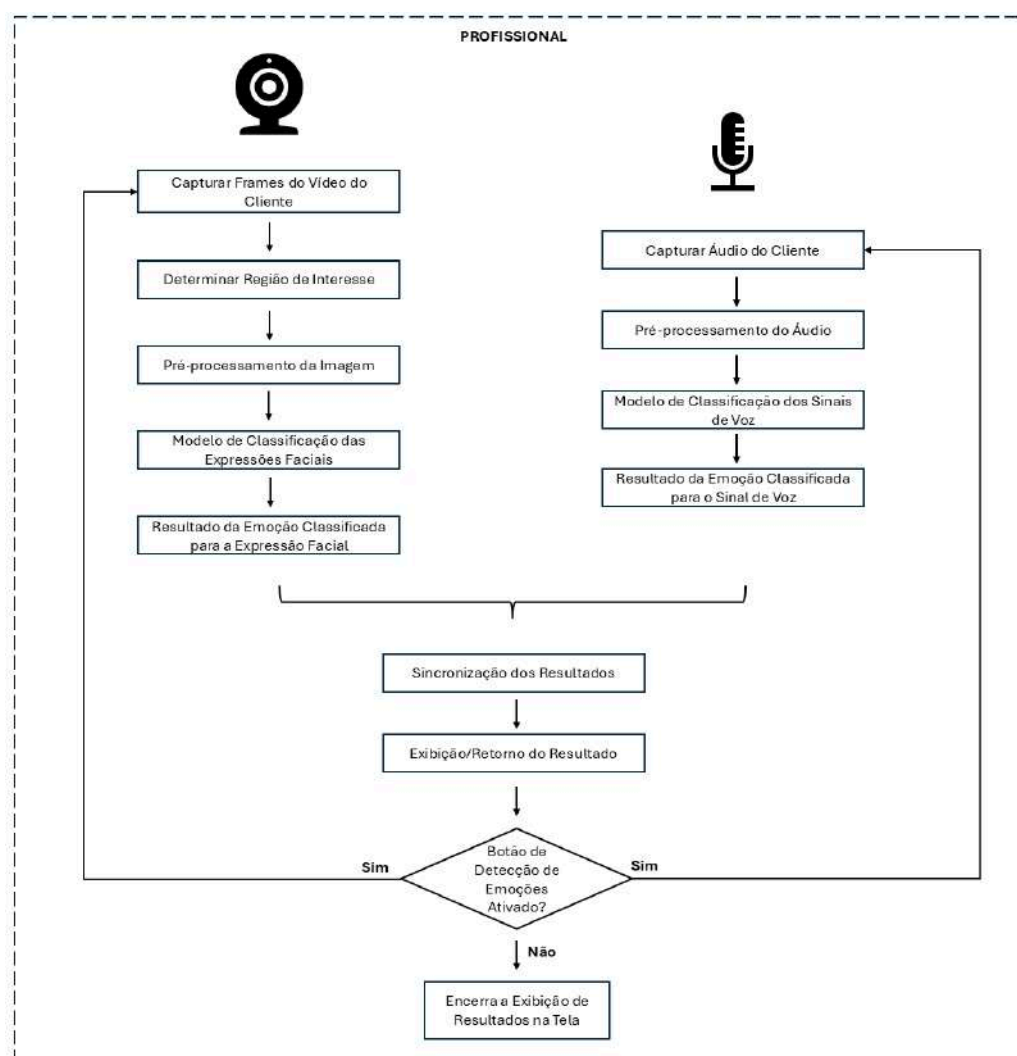
⁹ O *RTCPeerConnection* representa uma conexão WebRTC entre o computador local e um ponto remoto. Resumidamente, ela fornece métodos para se conectar a um ponto remoto, manter e monitorar a conexão, encerrando-a quando não for mais necessária (MOZILLA, 2025b).

¹⁰ Codecs podem ser definidos como algoritmos que codificam e decodificam mídia digital, como áudio e vídeo.

¹¹ Socket.io. Disponível em: <<https://socket.io/>>.

Na Figura 7 é apresentado um diagrama que inclui o sistema de análise emocional. As etapas para a detecção das emoções por meio da expressão facial e tonalidade da voz serão realizadas separadamente. Para a detecção das expressões faciais, os quadros do vídeo serão capturados, as regiões de interesse da imagem serão determinadas e, em seguida, um pré-processamento será realizado na imagem. Com isso será possível utilizar o modelo de classificação já treinado e então obter uma classificação para a respectiva emoção. O mesmo processo ocorrerá com os dados de áudio. Após a sincronização dos resultados, a classificação final da emoção será exibida na tela do profissional. Todas estas etapas serão possíveis enquanto o botão de detecção de emoções estiver habilitado na interface da videochamada.

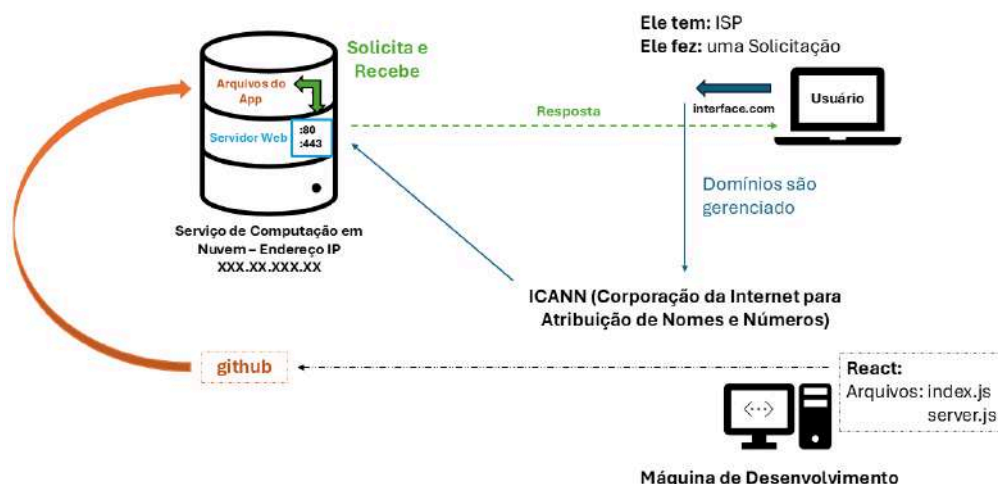
Figura 7 - Diagrama Incluindo o Sistema de Análise Emocional



Fonte: Elaborado pela autora.

Por sua vez, visando a implantação do sistema desenvolvido, a Figura 8 exemplifica a estrutura planejada/idealizada.

Figura 8 - Esboço para Implantação



Fonte: Elaborado pela autora.

Basicamente, o usuário – que possui um ISP (Provedor de Serviços de Internet) – enviará uma solicitação para acessar a interface (exemplo utilizado na imagem: interface.com). Em seguida, através do ISP é possível obter o acesso do site com o ICANN (Corporação da Internet para Atribuição de Nomes e Números), que gerencia os registros de domínios e permite descobrir o respectivo endereço IP.

Dado que é uma solicitação do navegador, será possível acessar uma das portas 80 ou 443, isto é, portas padrões para os protocolos HTTP (Protocolo de Transferência de Hipertexto) e HTTPS (Protocolo de Transferência de Hipertexto Seguro) de comunicação na web. Internamente, o servidor web irá acessar os arquivos (coletados anteriormente da máquina de desenvolvimento com o GitHub¹²). Tudo funcionará devido a instalação das dependências necessárias para a execução.

A solicitação será processada e, em seguida, uma resposta será enviada de volta para o servidor web que hospeda o site e aplicação. Logo, ao retornar para o navegador o site estará disponível para acesso.

¹² GitHub: Plataforma de Hospedagem de código fonte. Disponível em: <<https://github.com/>>.

3.2 Levantamento de Opinião por Questionário

3.2.1 Caracterização do Estudo e Coleta de Dados

Para um maior entendimento sobre as experiências, preferências, dificuldades e os problemas enfrentados por uma amostra de profissionais que atuam na área de Recursos Humanos e Psicologia, durante uma entrevista/processo seletivo com candidatos ou atendimento psicológico com pacientes, uma pesquisa de levantamento de caráter descritivo e abordagem quantitativa foi realizada.

A coleta de dados ocorreu no período de 24 de fevereiro a 14 de março, através da aplicação de um questionário estruturado fechado, criado com o auxílio da ferramenta online de gerenciamento de pesquisas “*Google Forms*”. O formulário foi divulgado em grupos da rede social *WhatsApp*, cujo integrantes são profissionais da área de Recursos Humanos e Psicologia. Com a divulgação e disponibilização do link, 18 profissionais foram alcançados.

Conforme disponibilizado no APÊNDICE A, o questionário criado para o levantamento de opinião contém oito perguntas, sendo cinco de múltipla escolha, duas de múltipla seleção e uma questão aberta.

3.2.2 Tratamento dos Dados

Finalizado o processo de coleta dos dados, a extração foi realizada para tabulação posterior, em formato compatível com o Microsoft Excel. Para uma análise segregada, isto é, com o conjunto de dados dividido pela categoria “áreas de atuação” (Recursos Humanos e Psicologia) e para uma visualização de dados detalhada, o Microsoft Power BI¹³ foi utilizado.

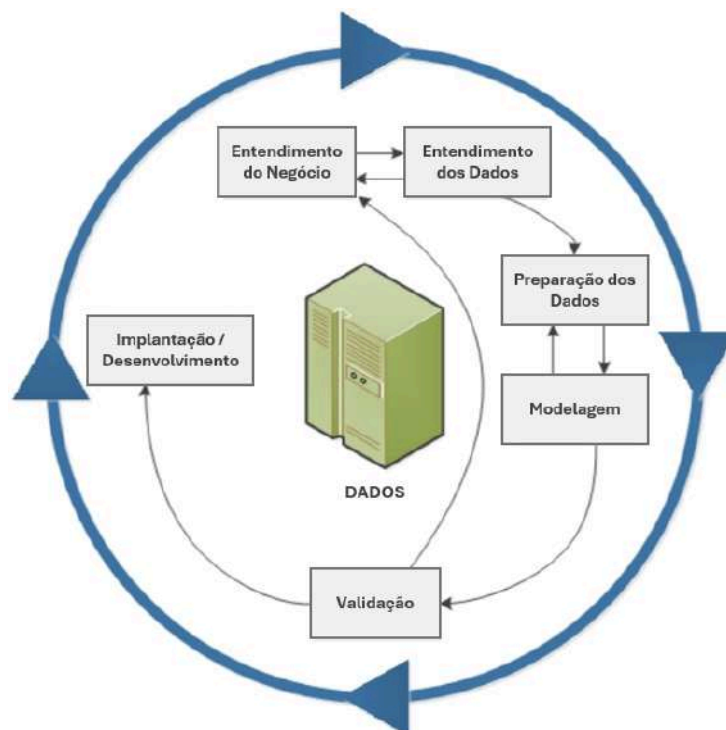
A fim de extrair compreensões e discernimentos, a estatística descritiva foi adotada e, portanto, a combinação das abordagens quantitativa e qualitativa se fez necessário para uma análise baseada em métricas, cálculos, ideias e experiências dos participantes.

¹³ Microsoft Power BI. Disponível em:
<<https://www.microsoft.com/pt-br/power-platform/products/power-bi>>.

3.3 Desenvolvimento da Modelagem de Detecção Facial e de Voz

Nesta seção será introduzido o CRISP-DM (*Cross Industry Standard Process for Data Mining*), que é uma metodologia altamente aplicada em projetos baseados em Ciência de Dados, a fim de se obter insights e garantir uma estrutura flexível e com etapas padronizadas durante o gerenciamento de projetos ágeis. De modo geral, o processo de mineração é dividido em seis fases. Na Figura 9 é possível analisar o ciclo com as setas indicando os avanços de acordo com o resultado obtido nas etapas anteriores, visto que é permitido a adaptação conforme os requisitos se alteram (WIRTH *et al.*, 2000).

Figura 9 - Esboço do Processo do CRISP-DM



Fonte: Adaptado de IBM (2025b).

Resumidamente, o CRISP-DM é dividido em seis blocos. A etapa inicial é composta pelo entendimento do negócio, seguido pelo entendimento dos dados, preparação dos dados, modelagem, validação e *deploy*/implantação do produto gerado.

3.3.1 Entendimento do Negócio

Inicialmente, será fundamental um entendimento sobre a situação dos usuários envolvidos: recrutadores e psicólogos. De modo geral, um levantamento bibliográfico acerca dos temas que a fundamentam, a determinação dos objetivos, requisitos e recursos entrará nesta etapa.

3.3.2 Coleta e Entendimento dos Dados

Esta etapa consistirá na coleta e entendimento dos conjuntos de dados que contém sinais de voz e imagens de expressões faciais. Resumidamente, os dados de voz serão coletados do *dataset* RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*), que disponibiliza registros de 24 atores profissionais, sendo 12 homens e 12 mulheres, falando expressões em entonações calmas, felizes, tristes, irritadas, com medo, surpresa e desgosto, além de registros cantados em entonações calmas, felizes, tristes, raivosas e temerosas (LIVINGSTONE, 2018).

Para a detecção de expressões faciais, um dos conjuntos de dados que será coletado chama-se FER2013 (*Facial Expression Recognition 2013*), que é composto por 28709 registros de imagens 48x48 em escala cinza, com as seguintes expressões disponíveis: raiva, tristeza, desgosto, medo, felicidade, surpresa, neutralidade (GOODFELLOW *et al.*, 2013).

O outro conjunto que será utilizado é o *AffectNet*, que disponibiliza registros de expressões faciais demonstrando felicidade, tristeza, medo, surpresa, nojo, desprezo, braveza e neutralidade. Este é considerado um dos maiores bancos de dados de expressões, sendo composto por imagens coloridas retiradas da Internet após a consulta em três dos principais mecanismos de busca (MOLLAHOSSEINI *et al.*, 2019).

Realizada a coleta, para um melhor entendimento dos dados, uma análise exploratória será efetuada, aplicando técnicas quantitativas, visualizações gráficas - se necessário -, utilizando a linguagem de programação *Python* e o editor de código fonte *Visual Studio Code*¹⁴.

¹⁴ Visual Studio Code. Disponível em: <<https://code.visualstudio.com/>>

3.3.3 Preparação dos Dados

Considerando os insights identificados na etapa anterior, um pré-processamento será aplicado aos conjuntos de dados de voz e imagens (aplicando diferentes técnicas e utilizando a linguagem de programação *Python*).

3.3.4 Modelagem

Na etapa de modelagem, serão exploradas diferentes ferramentas, técnicas computacionais e técnicas de modelagem a fim de encontrar e obter os resultados de classificações desejados.

3.3.5 Validação

Através da análise de métricas como, por exemplo, Precisão, Acurácia, *Recall*, F1 Score (média harmônica entre a Precisão e *Recall*) os modelos e suas performances serão comparados/avaliados. Dependendo do desempenho obtido, ajustes e parâmetros serão otimizados.

3.3.6 Implantação

Finalizado e escolhido o modelo que apresentar melhor desempenho, a etapa final consistirá em integrá-lo no sistema/interface desenvolvido para o projeto.

3.4 Interface Web para o Profissional e para o Usuário

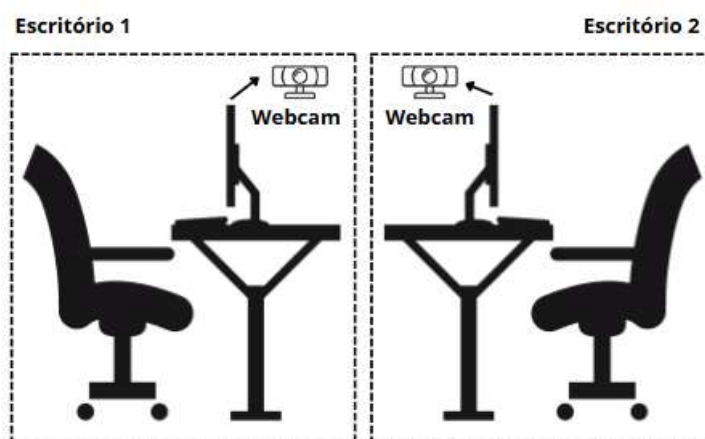
3.4.1 Componentes e Ambientes Necessários

Resumidamente, ambos participantes (o profissional e o cliente) estarão em seus respectivos escritórios/ambientes de trabalho e participarão de maneira remota, através de uma plataforma de videoconferência que permite que o usuário (de

maneira online) se comunique e hospede suas reuniões com o compartilhamento de um link para o acesso.

Os componentes necessários para cada participante são: um computador com acesso à rede, uma Webcam, um microfone, um alto-falante/fone de ouvido. Na Figura 10, um esboço do cenário é ilustrado.

Figura 10 - Ilustração dos Ambientes e Componentes Necessários



Fonte: Elaborado pela autora.

Visando a criação de uma interface interativa e que facilite a navegação do usuário e a manutenção (caso necessário), as linguagens de programação e ferramentas utilizadas serão: CSS (*Cascading Style Sheets*) para o estilo personalizado, *JavaScript*, *React*¹⁵, que torna possível a criação de interfaces de usuários a partir de componentes, *Redux*¹⁶, que pode ser definido como uma biblioteca de gerenciamento de estado para aplicação e seus componentes, o *socket.io*¹⁷ que permite a comunicação bidirecional em tempo real entre clientes e servidores, sendo composto por dois componentes (servidor e cliente *Node.js*) e, adicionado a isso, o WebRTC, previamente apresentado nas Seções 2.3 e 3.1. Ademais, o editor de código fonte utilizado será o *Visual Studio Code*¹⁸. A idealização da interface Web do profissional está descrita em detalhes na Seção 4.1.

¹⁵ React. Disponível em: <<https://react.dev/>>.

¹⁶ Redux. Disponível em: <<https://redux.js.org/>>.

¹⁷ Socket.io. Disponível em: <<https://www.npmjs.com/package/socket.io>>.

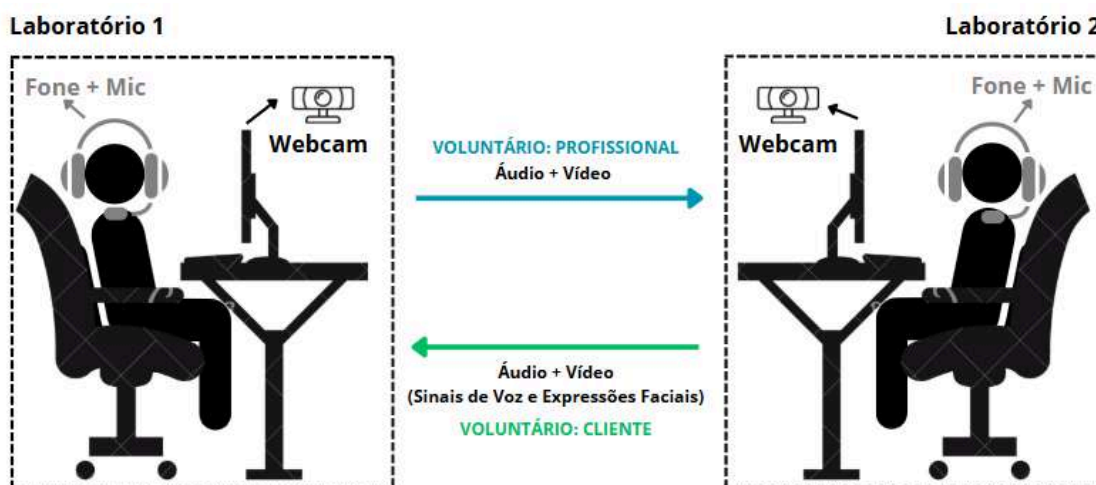
¹⁸ Visual Studio Code. Disponível em: <<https://code.visualstudio.com/>>.

3.4.2 Descrição do Cenário Experimental

As sessões experimentais serão realizadas em dois laboratórios localizados na Universidade Federal do ABC (UFABC), campus de São Bernardo do Campo. No caso, o cenário simulará os dois ambientes descritos na Seção 3.4.1 (Escritório 1 e Escritório 2) e ambos os espaços contarão com os mesmos componentes: um computador, uma Webcam, microfone e fone de ouvido. Em resumo, isso permitirá que a situação de vídeo conferência remota ocorra e seja simulada e, ainda sim, exista o controle sobre as possíveis variáveis.

Na Figura 11 o cenário experimental é ilustrado. No Laboratório 1, o voluntário (no caso, o profissional) interagirá e se comunicará com o indivíduo do outro cenário (Laboratório 2) através do uso da interface de vídeo chamada.

Figura 11 - Ilustração do Cenário Experimental



Fonte: Elaborado pela autora.

3.4.3 Sessões Experimentais com Voluntário

Para a realização da sessão experimental com voluntário, a Figura 12 foi elaborada a fim de organizar e visualizar o processo geral.

Figura 12 - Diagrama para a Realização das Sessões Experimentais



Fonte: Elaborado pela autora.

Em resumo, o voluntário participará de uma simulação de atendimento ou entrevista de emprego e atuará de acordo com a sua respectiva profissão. Esta simulação de videochamada será conduzida pelo pesquisador do estudo que também participará do processo. Em média, a duração de cada sessão experimental será de aproximadamente, 45 minutos.

Para os casos de simulação de recrutamento é desejado que o diálogo ocorra de maneira descontraída, mas espera-se que perguntas como, “Por que você quer trabalhar nesta empresa?”, “Quais são os seus pontos fortes e fracos?”, “Pode me falar um pouco sobre você?”, “Quais são os seus objetivos?”, sejam formuladas pelo profissional voluntário, buscando reações naturais e respostas autênticas do entrevistado. Finalizada a entrevista, será aplicado um questionário com o intuito de avaliar a interatividade, performance e o desempenho da interface, assim como o impacto do sistema de detecção para a condução de sua entrevista e tomadas de decisão.

Por sua vez, para a simulação de atendimento psicológico é esperado que a comunicação seja assertiva, com um diálogo eficaz e acolhedor. Finalizado o atendimento, também será aplicado um questionário com o intuito de avaliar a interatividade, performance e o desempenho da interface, assim como o impacto do sistema de detecção para a condução de seu atendimento.

3.5 Avaliação de Desempenho por Questionário

Por fim, do mesmo modo que foi descrito na Seção 3.2, um novo levantamento para avaliar o uso, o desempenho e o funcionamento da interface Web será realizado. Nesta etapa, após o profissional (psicólogo ou recrutador) testar as

funcionalidades do interpretador emocional, eles responderão um questionário estruturado fechado, criado com o auxílio da ferramenta online de gerenciamento de pesquisas “*Google Forms*”.

Resumidamente, este questionário apresentará questões envolvendo questões de múltipla escolha (para avaliar se a plataforma, a interatividade, detecção das emoções, a velocidade da transmissão de dados foram boas ou ruins, em uma escala de 1 a 5, por exemplo), além de questões abertas, para que o usuário disserte sobre suas impressões, dificuldades enfrentadas durante o acesso e sugestões de melhorias visando uma otimização do serviço. O esboço do formulário pode ser visualizado no APÊNDICE B.

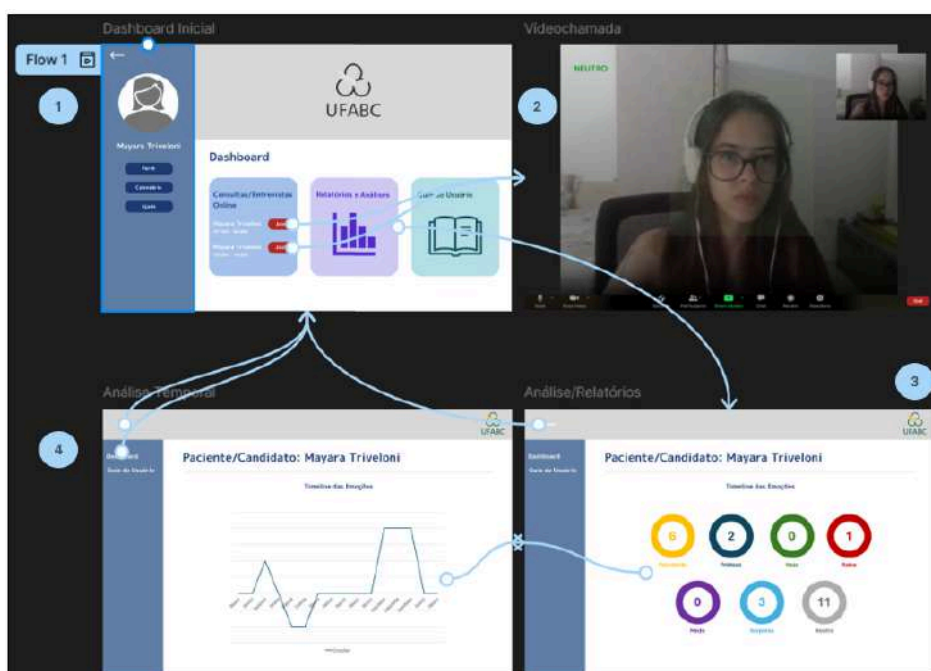
4 IMPLANTAÇÃO DO SISTEMA

Visando a implantação do sistema, um protótipo foi idealizado com o auxílio de uma ferramenta de projetos e, em paralelo, a modelagem para a classificação de emoções por meio da detecção facial e dos sinais de voz foi desenvolvida. Devido a algumas dificuldades enfrentadas e preferências visuais/estéticas durante a codificação, o produto implementado apresentou um visual diferente da proposta inicial. Estas etapas e alterações foram descritas e apresentadas nas seções a seguir.

4.1 Estudo Inicial para uma Interface Web Idealizada

Para o desenvolvimento do protótipo da Interface, o Figma¹⁹, editor gráfico de prototipagem de projetos de design baseado em navegador Web foi utilizado. A estrutura foi organizada em quatro quadrantes e as respectivas posições adotadas podem ser visualizadas na Figura 13. Além disso, o nome de cada tela está apresentado no canto superior esquerdo de cada quadrante.

Figura 13 - Prototipagem da Interface Web



Fonte: Elaborado pela autora.

¹⁹ Figma. Disponível em: <<https://www.figma.com/login>>.

Nesta seção, o termo “profissionais” será utilizado para referenciar os psicólogos e recrutadores. Por sua vez, o termo “cliente” será utilizado para referenciar os entrevistados e pacientes. Em resumo, cada quadrante representa uma tela da interface principal, interligada por setas que indicam os fluxos esperados após a interação do usuário.

- Quadrante 1 (Flow 1) – *Dashboard* Inicial: Duas das setas que partem dessa tela, foram configuradas com ação interativa “*on click*” nos botões vermelhos “*Join*”. Neste caso, essas setas direcionam o usuário para a tela Videochamada. Por sua vez, a terceira seta (também configurada com ação interativa “*on click*”) parte do ícone roxo chamado “Relatórios e Análises”, que encaminha o profissional para a tela Análise/Relatórios;

- Quadrante 2 – Videochamada: nesta tela é possível visualizar uma representação de como será a transmissão de vídeo e áudio, assim como, as funções disponíveis para o usuário: detecção gerada pela análise emocional e botões para habilitar/desabilitar o compartilhamento de áudio, vídeo, tela e detecção. Assim como botão para o encerramento da videochamada;

- Quadrante 3 – Análise/Relatórios: tela que apresenta o relatório emocional do cliente após o encerramento da videochamada. A seta que sai deste quadrante direciona o usuário para a tela Análise Temporal. No caso, essa transição foi configurada de modo que permite que o conteúdo se mova verticalmente (simulando a funcionalidade de rolagem do scroll do mouse para baixo). Já a outra seta retorna para o Dashboard Inicial, com o clique do usuário na seta/botão de retorno disponível no canto superior esquerdo;

- Quadrante 4 – Análise Temporal: tela que apresenta o gráfico temporal das emoções do cliente após o encerramento da videochamada. Como no quadrante anterior, uma seta foi configurada de modo a permitir que o conteúdo se mova verticalmente (simulando a funcionalidade de rolagem do scroll do mouse para cima) para retornar ao Quadrante 3. A outra seta que sai dessa tela retorna para o Dashboard Inicial, com o clique do usuário na seta/botão de retorno disponível no canto superior esquerdo.

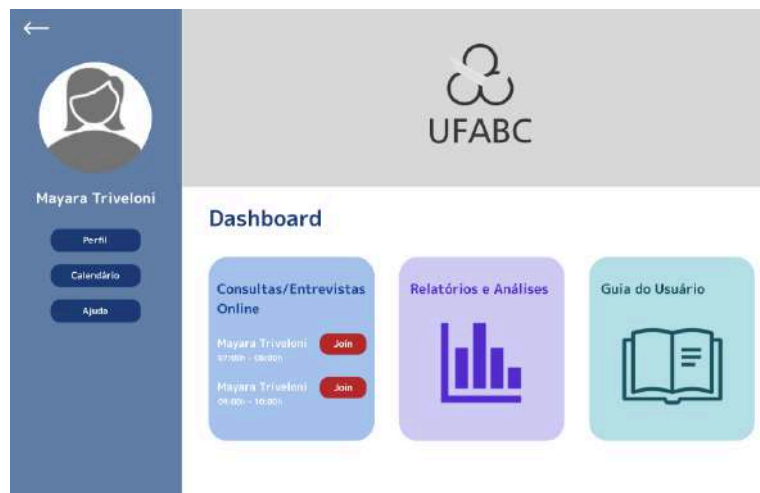
4.1.1 Detalhamento das Telas Idealizadas (Protótipo)

A tela inicial disponibiliza as principais funções e o *Dashboard* para o profissional, sendo possível visualizar na Figura 14. Na lateral esquerda, uma faixa de tonalidade azulada separa a seção que contém informações do usuário. Ainda nesta região, na parte superior, uma foto pode ser adicionada (no formato circular) e logo abaixo, o nome do profissional é apresentado. Além disso, existem três botões (na cor azul escuro) que permitem que em apenas um clique, o usuário edite seu perfil (alterando o nome, foto e outros dados pessoais), visualize o calendário (visando os possíveis agendamentos de entrevista ou atendimento) e possa obter ajuda (as dúvidas frequentes que possam existir sobre a interface – FAQ). A seta no canto superior esquerdo efetua o *log out* do usuário.

Na região localizada à direita da tela inicial, na parte superior (com o fundo cinza claro) um modelo de logotipo da UFABC está centralizado. Logo abaixo, na região do Dashboard (com fundo na tonalidade branca) três ícones retangulares com bordas arredondadas representam as principais funções da interface.

O primeiro ícone “Consultas/Entrevistas Online” disponibiliza os principais agendamentos do dia, com o nome do cliente, o respectivo horário da chamada de vídeo e um botão vermelho “Join” que direciona para a sala de reunião. O segundo ícone “Relatórios e Análises” direciona para uma página que apresenta análises e gráficos obtidos após a finalização da chamada de vídeo (encerramento da entrevista/consulta). Por fim, ao clicar no terceiro ícone “Guia do Usuário”, o profissional é direcionado para uma página que contém a documentação e guia contendo as funcionalidades da interface.

Figura 14 - Página Inicial da Interface

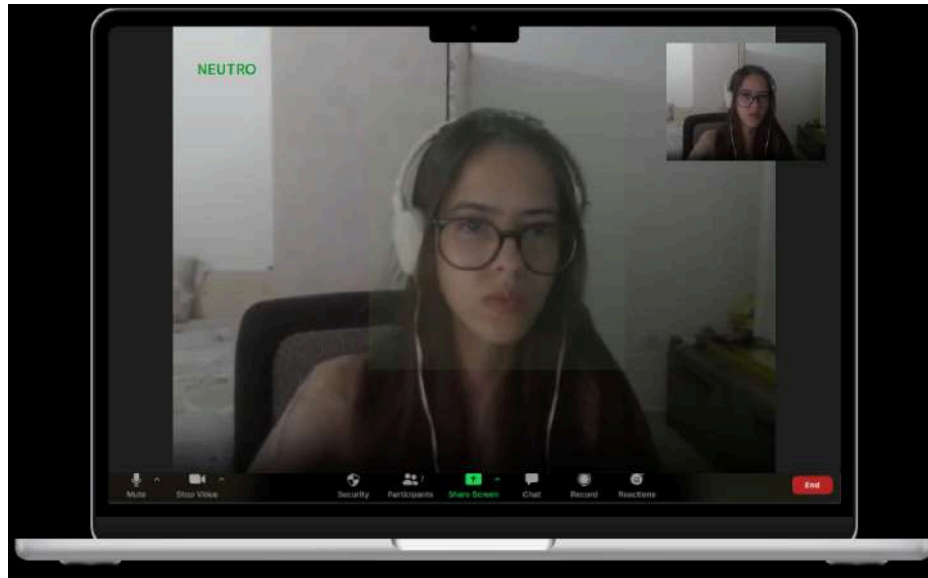


Fonte: Elaborado pela autora.

Caso o profissional clique no botão “Join” presente no ícone “Consultas/Entrevistas”, ele será direcionado para uma nova página e, automaticamente, “abrirá” a respectiva sala de reunião da chamada de vídeo. Como é possível observar na Figura 15, a tela do dispositivo é preenchida pela transmissão do vídeo do cliente e a detecção das expressões/tonalidade de voz será sinalizada no canto superior esquerdo.

No canto superior direito, o próprio vídeo captado pela Webcam do profissional é entregue em um retângulo de tamanho reduzido. Na parte inferior da tela, as funções básicas/botões permitirão que o usuário ative/desative o áudio e/ou vídeo, compartilhe a tela, altere as configurações de transmissão de áudio e/ou vídeo e encerre a chamada.

Figura 15 - Chamada de Vídeo



Fonte: Elaborado pela autora.

Por sua vez, se o profissional clicar no gráfico roxo presente no ícone “Relatórios e Análises” (disponibilizado na página inicial), ele será direcionado para uma nova aba que contém o nome do cliente na parte superior e, logo abaixo (separado por uma linha horizontal cinza), a “*Timeline* das Emoções” que contém os resultados/contabilização das emoções captadas durante a vídeo chamada. Nesta região será possível visualizar o resultado numérico nos sete círculos de cores diferentes, a fim de representar cada emoção:

- Amarelo: Felicidade;
- Azul Escuro: Tristeza;
- Verde Musgo: Nojo;
- Vermelho: Raiva;
- Roxo: Medo;
- Azul Claro: Surpresa;
- Cinza: Neutro.

Mais uma vez, na lateral esquerda há uma faixa de tonalidade azulada que separa uma seção que contém duas opções: um botão chamado “*Dashboard*”, que retorna para a página inicial e o botão “Guia do Usuário”, que direciona o profissional para a documentação da interface. A seta no canto superior esquerdo, desloga o profissional. Esta tela pode ser visualizada na Figura 16.

Figura 16 - Página de Relatórios e Análises dos Pacientes/Candidatos



Fonte: Elaborado pela autora.

Ainda na mesma página, com o *scroll* do mouse o profissional poderá visualizar um gráfico contendo uma linha do tempo de todas as emoções registradas na última chamada, fornecendo assim, a oportunidade de uma melhor análise das oscilações que ocorreram no decorrer da entrevista/consulta. Esta tela pode ser visualizada na Figura 17.

Para a construção do classificador de emoções, a metodologia CRISP-DM foi seguida e, como apresentado na Seção 3.3.2, entre os *datasets* utilizados para o treinamento se tem o *Facial Expression Recognition 2013 - FER2013*. Este conjunto de dados é composto por 28709 registros de imagens para treinamento e 7178 registros de imagens para teste, com dimensão de 48x48 pixels, em escala cinza, com sete expressões disponíveis: raiva, tristeza, desgosto/nojo, medo, felicidade, surpresa, neutralidade (GOODFELLOW *et al.*, 2013).

Na Figura 18 é possível visualizar uma amostra de sete imagens devidamente rotuladas de cada classe disponível.

Figura 18 - Amostra de Imagens do *dataset* FER2013



Fonte: Adaptado de FER2013 (GOODFELLOW *et al.*, 2013).

Na Tabela 1 é possível visualizar a distribuição de imagens para as sete classes. Percebe-se que a distribuição não é uniforme, caracterizando um desbalanceamento entre os dados. Como exemplo, existe uma diferença considerável entre as classes felicidade, contendo 7215 imagens e de desgosto, com apenas 436 imagens.

Tabela 1 - Distribuição de Imagens por Classe (FER2013)

Emoções	Quantidade de Imagens Treino	Quantidade de Imagens Teste
Raiva	3995	958
Desgosto	436	111
Medo	4097	1024
Felicidade	7215	1774
Neutralidade	4965	1233
Tristeza	4830	1247
Surpresa	3171	831

Fonte: Elaborado pela autora.

O outro conjunto de dados utilizado é o *AffectNet*, que disponibiliza registros de expressões faciais demonstrando felicidade, tristeza, medo, surpresa, desgosto, desprezo, raiva e neutralidade. Este conjunto de dados é composto por 16108 registros para treinamento e 14518 registros para teste, com dimensão de 96x96 pixels e escala colorida (MOLLAHOSSEINI *et al.*, 2019).

Na Figura 19 é possível visualizar uma amostra de oito imagens rotuladas de cada classe disponível.

Figura 19 - Amostra de Imagens do *dataset AffectNet*



Fonte: Adaptado de *AffectNet* (MOLLAHOSSEINI *et al.*, 2019).

Na Tabela 2 é possível visualizar a distribuição de imagens para as oito classes disponíveis. Mais uma vez, percebe-se que a distribuição não é uniforme, caracterizando um desbalanceamento entre os dados.

Tabela 2 - Distribuição de Imagens por Classe (*AffectNet*)

Emoções	Quantidade de Imagens Treino	Quantidade de Imagens Teste
Raiva	1500	1718
Desprezo	1559	1312
Desgosto	1229	1248
Medo	1512	1664
Felicidade	2340	2704
Neutralidade	2758	2368
Tristeza	3091	1584
Surpresa	2119	1920

Fonte: Elaborado pela autora.

De modo geral, o treinamento será dividido em dois processos de variação: treinamento com o *dataset* desbalanceado, isto é, mantendo a quantidade real de imagens coletadas e treinamento com o *dataset* balanceado, que através do *undersample*, realiza-se uma redução no número de dados disponíveis nas classes maiores e define-se uma quantidade que iguale as classes menores.

A modelagem será conduzida em duas etapas principais:

- Teste de modelos pré-treinados (*DeepFace* e *RMN*) para testes preliminares e obter uma avaliação inicial de desempenho para ambas as bases de dados, FER2013 e *AffectNet*;
- Implementação e teste com arquiteturas de Redes Neurais Convolucionais (CNNs) para a classificação de expressões faciais.

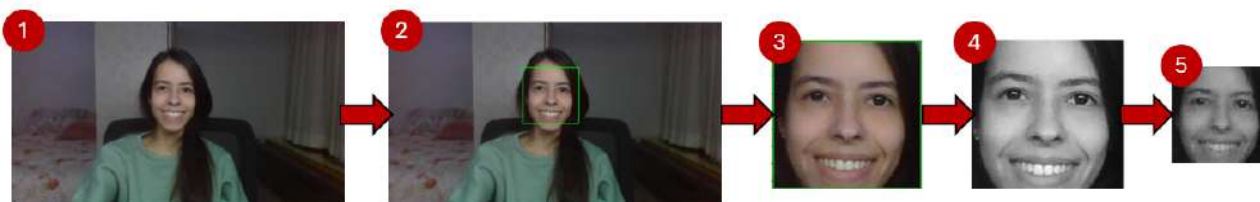
4.2.1 Modelos pré-treinados: *DeepFace*

O *DeepFace* pode ser definido como um framework híbrido de reconhecimento facial e análise de idade, gênero, emoção e etnia. Basicamente, apresenta um pipeline com cinco etapas: detecção, alinhamento, normalização, representação e verificação. As sete emoções disponíveis são: raiva, medo, neutralidade, tristeza, nojo, felicidade e surpresa (SERENGIL, 2021).

Para um melhor entendimento da ferramenta, este modelo pré-treinado foi testado com fotos específicas que retratavam emoções como por exemplo, felicidade e que foram retiradas com o auxílio de uma webcam. As etapas deste processo foram listadas a seguir e na Figura 20 é possível visualizar também os resultados obtidos:

- 1) Captura de imagem com o auxílio de uma webcam;
- 2) Detecção da região de interesse, ou seja, a face do indivíduo por meio do modelo *Haar cascade*, da biblioteca *OpenCV*, proposto por Paul Viola e Michael Jones em seu artigo "*Rapid Object Detection using a Boosted Cascade of Simple Features*", em 2001;
- 3) Extração da região de interesse;
- 4) Conversão de escala colorida para a tonalidade cinza;
- 5) Redimensionamento das dimensões da imagem para 48x48 pixels.

Figura 20 - Pré-processamento da Imagem Capturada



Fonte: Elaborado pela autora.

Após a aplicação da função *DeepFace.analyze()* para a análise da imagem e reconhecimento de emoção, as probabilidades foram retornadas com o apontamento para uma emoção dominante de maneira correta (Felicidade), como destacado na Tabela 3.

Tabela 3 - Detecção Obtida para Imagem Capturada

Emoções	Quantidade de Imagens Treino
Raiva	0,00%
Desgosto	0,00%
Medo	0,00%
Felicidade	99,52%
Neutralidade	0,00%
Tristeza	0,00%
Surpresa	0,48%

Fonte: Elaborado pela autora.

Utilizando agora o modelo *DeepFace* com os dados de treinamento FER2013 desbalanceados, a acurácia obtida foi satisfatória e de aproximadamente, 91,41% com um tempo médio de processamento por imagem de $0,014 \pm 0,004$ segundos. Após uma investigação para compreender o motivo por trás do excelente desempenho obtido, foi descoberto que entre as diversas bases de dados utilizadas para o treinamento do respectivo modelo, a FER2013 foi inclusa. Portanto, a avaliação para este caso é comprometida e incorreta, visto que impede a capacidade de generalização do modelo.

Por sua vez, para os dados de treinamento *AffectNet*, a acurácia obtida foi inferior e igual a 32,08% e o tempo médio de processamento por imagem de aproximadamente, $0,015 \pm 0,005$ segundos.

Considerando agora o balanceamento do FER2013, a redução no número de dados disponíveis nas classes foi realizada para alcançar 436 registros (foi definida uma quantidade aleatória que se igualasse as classes menores). Mais uma vez, obteve-se uma boa performance e semelhante ao resultado obtido com a base de dados desbalanceada anteriormente. Neste caso, a acurácia foi de 90,79% e um tempo de processamento por imagem de $0,013 \pm 0,009$ segundos.

Por fim, para o balanceamento do *AffectNet*, a redução no número de dados disponíveis nas classes foi realizada para alcançar 1229 registros e se manteve próximo à performance alcançada anteriormente.

Na Tabela 4, todos os resultados descritos foram registrados.

Tabela 4 - Acurácia com a Aplicação do *DeepFace*

Conjunto de Dados	Dados Desbalanceados		Dados Balanceados	
	DeepFace	Tempo de Processamento	DeepFace	Tempo de Processamento
FER2013	91,41%	$0,014 \pm 0,004$ seg	90,79%	$0,013 \pm 0,009$ seg
AffectNet	32,08%	$0,015 \pm 0,005$ seg	29,37%	$0,014 \pm 0,003$ seg

Fonte: Elaborado pela autora.

Além disso, outras métricas como, *Precisão* e *Recall*, também foram avaliadas para cada uma das classes disponíveis, conforme apresentado na Tabela 5. Para a base de dados FER2013 desbalanceada, o modelo apresentou um excelente resultado para todas as emoções, com uma performance acima de 85% e com ênfase para a felicidade, onde o algoritmo conseguiu identificar 97% das faces felizes, com uma precisão de 96%. Percentuais próximos foram obtidos também com os dados balanceados, logo, dado os resultados obtidos, o balanceamento pode ser uma técnica descartada para os casos analisados.

Tabela 5 - Métricas/Performance para cada Classe (FER2013)

Emoções	Dados Desbalanceados		Dados Balanceados	
	Precisão	Recall	Precisão	Recall
Raiva	93%	85%	92%	87%
Desgosto	96%	91%	100%	91%
Medo	89%	90%	86%	90%
Felicidade	96%	97%	92%	95%
Neutralidade	85%	94%	82%	95%
Tristeza	88%	86%	88%	84%
Surpresa	97%	93%	97%	92%

Fonte: Elaborado pela autora.

Do mesmo modo, na Tabela 6 está registrado todos os desempenhos alcançados ao aplicar o *DeepFace* no *AffectNet*. Os percentuais foram próximos para a base de dados balanceada e desbalanceada, portanto, para este caso, o balanceamento pode ser uma técnica descartada. Além disso, percebe-se que as métricas se mostraram favoráveis para algumas emoções como, felicidade e neutralidade, onde o algoritmo conseguiu identificar 63% das faces felizes, com uma precisão de 52% e conseguiu identificar 40% das faces neutras, com uma precisão de 25%, apesar disso, os demais resultados se mostram insatisfatórios e com baixos percentuais.

Tabela 6 - Métricas/Performance para cada Classe (*AffectNet*)

Emoções	Dados Desbalanceados		Dados Balanceados	
	Precisão	Recall	Precisão	Recall
Raiva	24%	36%	29%	36%
Desgosto	45%	01%	56%	01%
Medo	21%	21%	27%	21%
Felicidade	52%	63%	48%	64%
Neutralidade	25%	40%	19%	40%
Tristeza	33%	27%	24%	26%
Surpresa	47%	18%	39%	17%

Fonte: Elaborado pela autora.

4.2.2 Modelos pré-treinados: *Residual Masking Network*

O *Residual Masking Network* é um modelo voltado para o reconhecimento de expressões faciais: felicidade, tristeza, raiva, medo, desgosto, surpresa e neutralidade. Resumidamente, sua arquitetura é baseada em Redes Neurais Convolucionais (CNNs) utilizando blocos residuais no treinamento (PHAM *et al.*, 2021). Dado que o conjunto de dados FER2013 foi uma das bases utilizadas para o seu treinamento, ela não será considerada nesta seção.

Portanto, utilizando agora o modelo RMN apenas com os dados de treinamento *AffectNet* desbalanceados, alcança-se uma acurácia igual a 41,04% e com um tempo médio de processamento por imagem de $1,1 \pm 0,2$ segundos. Para os dados balanceados não ocorreu uma mudança significativa, mas nota-se que o tempo de processamento por imagem reduziu para $1,0 \pm 0,2$ segundos, assim como a acurácia, com aproximadamente 37,93%. Na Tabela 7 estas informações serão apresentadas.

Tabela 7 - Acurácia com a Aplicação do RMN

Conjunto de Dados	Dados Desbalanceados		Dados Balanceados	
	RMN	Tempo de Processamento	RMN	Tempo de Processamento
AffectNet	41,04%	$1,1 \pm 0,2$ seg	37,93%	$1,0 \pm 0,2$ seg

Fonte: Elaborado pela autora.

Mais uma vez, outras métricas como, Precisão e *Recall*, também foram avaliadas para cada uma das classes disponíveis, conforme apresentado na Tabela 8. Nenhuma mudança notável com o balanceamento para estas situações.

Tabela 8 - Métricas/Performance para cada Classe (*AffectNet*)

Emoções	Dados Desbalanceados		Dados Balanceados	
	Precisão	Recall	Precisão	Recall
Raiva	20%	38%	24%	37%
Desgosto	40%	14%	53%	11%
Medo	26%	33%	31%	33%
Felicidade	73%	77%	68%	77%
Neutralidade	47%	44%	38%	42%
Tristeza	44%	38%	32%	39%
Surpresa	34%	25%	36%	26%

Fonte: Elaborado pela autora.

4.2.3 Arquiteturas para Redes Convolucionais Neurais

Conforme apresentado anteriormente, os treinamentos foram realizados utilizando duas bases de dados: FER2013 e *AffectNet*. Por meio da linguagem Python, os modelos foram criados e a abordagem de *data augmentation* (aumentar a variação dos dados) foi aplicada ao gerar a base de treinamento ao rotacionar a imagem, variar o zoom e inversão de linhas e colunas horizontalmente. Além disso, para o pré-processamento foi realizada a normalização dos dados, um redimensionamento padrão de 48x48 pixels para a base de dados completa e foi definido que os dados serão enviados a cada 16 imagens (valor de *batch size*). Essa etapa foi realizada com o auxílio de um módulo presente no TensorFlow/Keras²⁴. Para a base de testes as imagens apenas foram normalizadas

Basicamente, a rede neural desenvolvida consiste em quatro blocos. Em cada bloco, existem duas camadas de convolução seguidas por uma camada de *BatchNormalization*, para transformar e manter a saída média próxima de 0 e o desvio padrão de saída próximo a 1, de acordo com a documentação da *API Keras* 3. Além disso, o bloco é seguido de uma camada de *MaxPooling* e uma de *Dropout*, a fim de selecionar o maior/principal valor de cada parte da imagem (reduzindo assim as dimensões, mas mantendo características importantes) e em sequência, prevenir o *overfitting*, isto é, não permitir que o algoritmo se adapte muito aos dados de treinamento (no caso, o valor 0.20 foi adotado para zerar 20% dos neurônios).

²⁴ TensorFlow Core: API do TensorFlow para criar e treinar modelos de Deep Learning. Disponível em: <<https://www.tensorflow.org/guide/keras?hl=pt-br>>

Finalizado os quatro blocos, uma camada *Flatten* é adicionada para transformar as saídas (matrizes) em um vetor, mantendo o seu respectivo tamanho. Por fim, duas camadas escondidas são adicionadas com 64 neurônios cada mais uma camada de saída objetivando gerar uma probabilidade para cada classe, através da função de ativação '*softmax*'. Na Figura 21 é possível visualizar a arquitetura desenvolvida e na Figura 22 para visualizar o resumo das camadas e parâmetros adotados.

Figura 21 - Arquitetura para Rede Neural Convolucional (CNN)

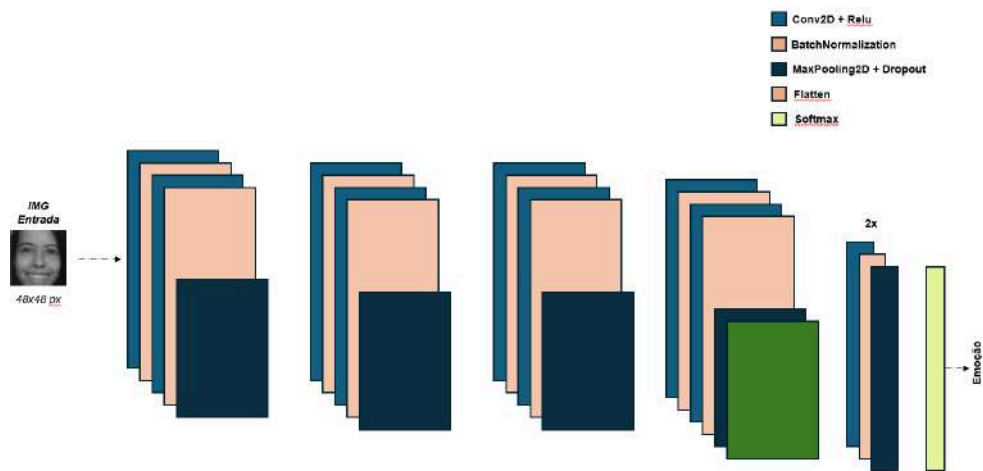


Figura 22 - Resumo do Modelo Desenvolvido

Layer (type)	Output Shape	Param #
conv2d_16 (Conv2D)	(None, 48, 48, 32)	896
batch_normalization_20 (Batch Normalization)	(None, 48, 48, 32)	128
conv2d_17 (Conv2D)	(None, 48, 48, 32)	9,248
batch_normalization_21 (Batch Normalization)	(None, 48, 48, 32)	128
max_pooling2d_8 (Max Pooling 2D)	(None, 24, 24, 32)	0
dropout_12 (Dropout)	(None, 24, 24, 32)	0
conv2d_18 (Conv2D)	(None, 24, 24, 64)	18,496
batch_normalization_22 (Batch Normalization)	(None, 24, 24, 64)	256
conv2d_19 (Conv2D)	(None, 24, 24, 64)	36,928
batch_normalization_23 (Batch Normalization)	(None, 24, 24, 64)	256
max_pooling2d_9 (Max Pooling 2D)	(None, 12, 12, 64)	0
dropout_13 (Dropout)	(None, 12, 12, 64)	0
conv2d_20 (Conv2D)	(None, 12, 12, 128)	73,856
batch_normalization_24 (Batch Normalization)	(None, 12, 12, 128)	512
conv2d_21 (Conv2D)	(None, 12, 12, 128)	147,584
batch_normalization_25 (Batch Normalization)	(None, 12, 12, 128)	512
max_pooling2d_10 (Max Pooling 2D)	(None, 6, 6, 128)	0
dropout_14 (Dropout)	(None, 6, 6, 128)	0
conv2d_22 (Conv2D)	(None, 6, 6, 256)	295,168
batch_normalization_26 (Batch Normalization)	(None, 6, 6, 256)	1,024
conv2d_23 (Conv2D)	(None, 6, 6, 256)	590,080
batch_normalization_27 (Batch Normalization)	(None, 6, 6, 256)	1,024
max_pooling2d_11 (Max Pooling 2D)	(None, 3, 3, 256)	0
dropout_15 (Dropout)	(None, 3, 3, 256)	0
flatten_1 (Flatten)	(None, 2304)	0
dense_6 (Dense)	(None, 64)	147,520
batch_normalization_28 (Batch Normalization)	(None, 64)	256
dropout_16 (Dropout)	(None, 64)	0
dense_7 (Dense)	(None, 64)	4,100
batch_normalization_29 (Batch Normalization)	(None, 64)	256
dropout_17 (Dropout)	(None, 64)	0
dense_8 (Dense)	(None, 7)	455

Fonte: Elaborado pela autora.

Após realizar o treinamento do modelo com 70 épocas, a acurácia obtida foi de aproximadamente, 75,84% com uma perda de 0,6613. Ao avaliar o modelo treinado com o conjunto de dados de teste, a acurácia resultante foi igual a 67,4%.

Na Tabela 9 é possível visualizar o desempenho alcançado com o treinamento. Para a classe Felicidade, o algoritmo conseguiu identificar 89% das faces felizes com uma precisão de 83%; para a expressão facial de surpresa, o algoritmo pôde identificar 81% das faces com uma precisão de 76% e a neutralidade apresentou como desempenho a identificação de 66% das faces com uma precisão de 61%. Em resumo, a emoção/classe que apresentou um pior desempenho foi a expressão de medo, onde o algoritmo conseguiu identificar apenas 45% das respectivas faces, com uma precisão de 55%.

Tabela 9 - Métricas das Classes (CNN) para FER2013

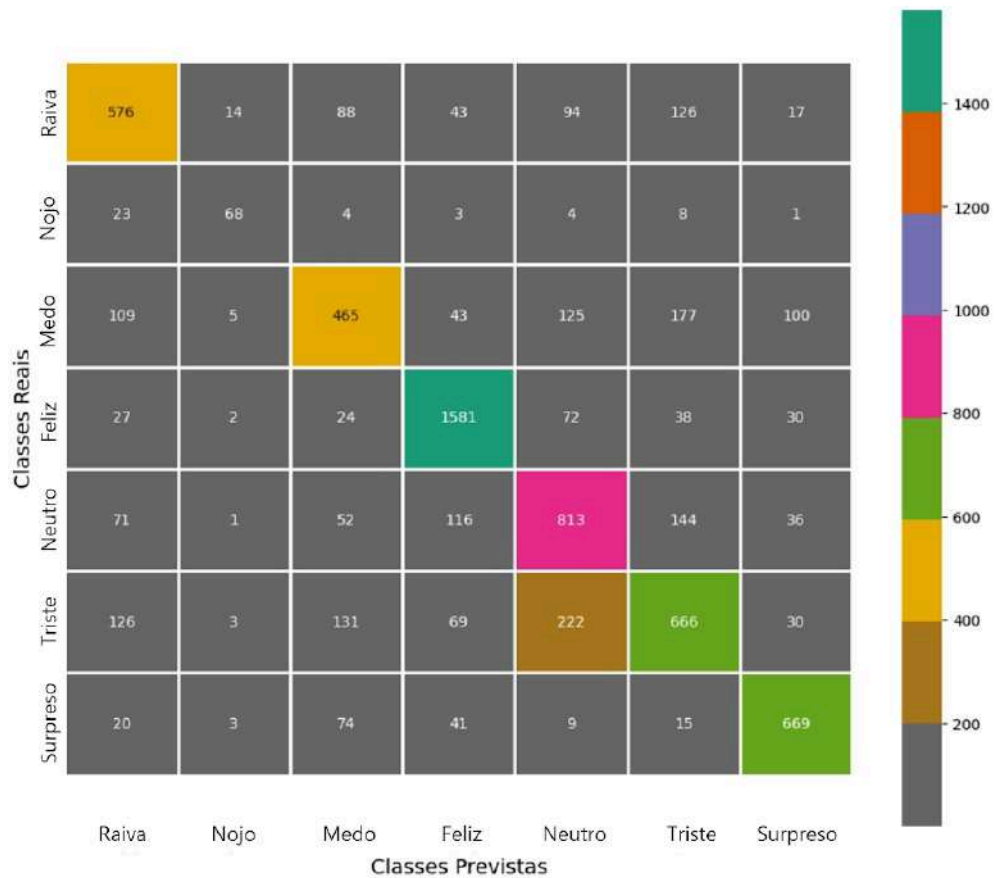
FER2013		
Emoções	Precisão	Recall
Raiva	61%	60%
Desgosto	71%	61%
Medo	55%	45%
Felicidade	83%	89%
Neutralidade	61%	66%
Tristeza	57%	53%
Surpresa	76%	81%

Fonte: Elaborado pela autora.

A matriz de confusão também foi utilizada para avaliar o desempenho do modelo, o que tornou possível a análise da distribuição dos acertos e erros por classe. É importante ressaltar que o desbalanceamento influencia no resultado da matriz, visto que o modelo tende a identificar as classes majoritárias. Como apresentado na Figura 23, o modelo acertou 1581 instâncias das expressões faciais felizes e, como é possível observar, as classificações incorretas para este caso podem ser consideradas mínimas. O mesmo ocorreu com as classes de surpresa, neutralidade e raiva onde a maior concentração de valores manteve-se em sua diagonal principal. Por outro lado, a classe tristeza foi confundida com a expressão

de neutralidade com uma frequência mais considerável, do mesmo modo que a expressão de nojo foi classificada incorretamente como raiva em alguns casos.

Figura 23 - Matriz de Confusão (FER2013)



Fonte: Elaborado pela autora.

O procedimento realizado anteriormente na Seção 4.2 de captura de imagem com a webcam foi aplicado novamente com a finalidade de teste do modelo. As mesmas etapas de pré-processamento foram seguidas e a detecção se mostrou correta ao identificar as expressões: feliz, neutra e surpresa. Por sua vez, ao realizar o teste com a captura de uma face triste, o modelo classificou incorretamente como neutro. As capturas utilizadas podem ser visualizadas na Figura 24.

Figura 24 - Testando o Modelo com Capturas da webcam



Fonte: Elaborado pela autora.

Agora utilizando o conjunto de dados *AffectNet* para o treinamento do modelo com 70 épocas, a acurácia obtida foi de aproximadamente, 75,18% com uma perda de 0,6934. Ao avaliar o modelo treinado com o conjunto de dados de teste, a acurácia resultante foi igual a 66,47%. Na Tabela 10 são apresentadas as métricas por classe/emoção identificada. A emoção que apresentou melhor desempenho foi a felicidade, onde o algoritmo conseguiu identificar 90% das faces felizes com uma precisão de 89%; a neutralidade, em que pôde identificar 93% das respectivas expressões com uma precisão de 72% e a que apresentou pior performance foi a surpresa, com um percentual de 39% para identificação das faces e uma precisão de 51%.

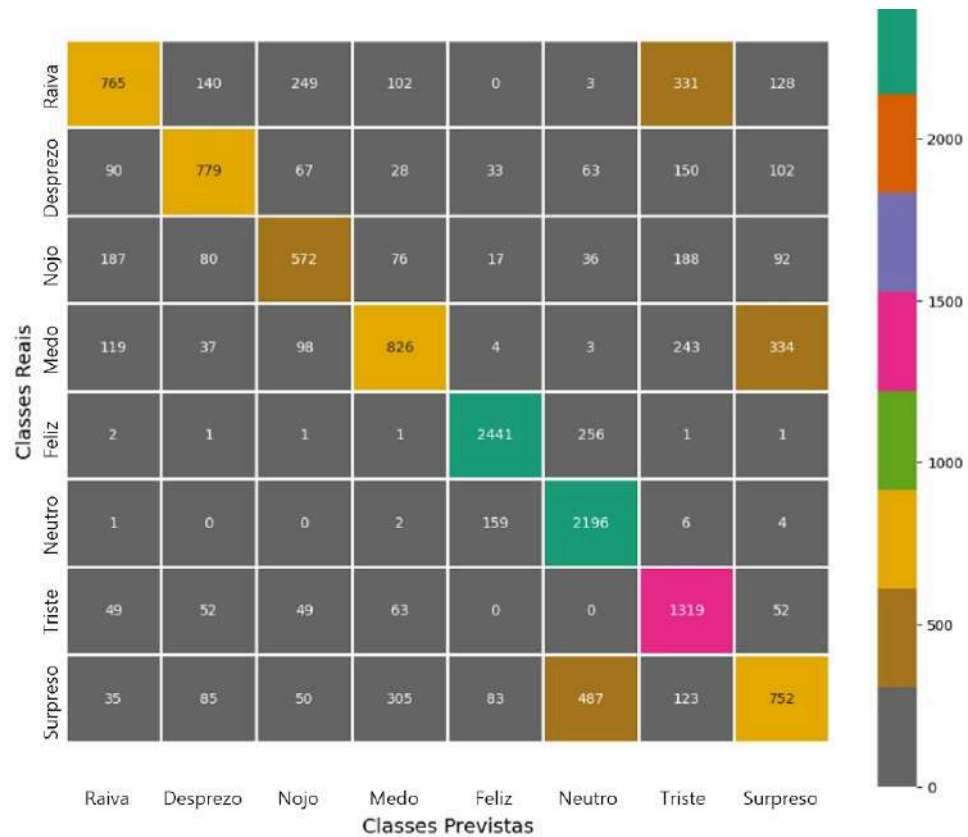
Tabela 10 - Métricas das Classes (CNN) para **AffectNet**

AffectNet		
Emoções	Precisão	Recall
Raiva	61%	45%
Desprezo	66%	59%
Desgosto	53%	46%
Medo	59%	50%
Felicidade	89%	90%
Neutralidade	72%	93%
Tristeza	56%	83%
Surpresa	51%	39%

Fonte: Elaborado pela autora.

Mais uma vez, a matriz de confusão foi utilizada para avaliar o desempenho do modelo, o que tornou possível a análise da distribuição dos acertos e erros por classe. Como apresentado na Figura 25, o modelo acertou 2441 instâncias das expressões faciais felizes e, como é possível observar, as classificações incorretas para este caso podem ser consideradas mínimas novamente.

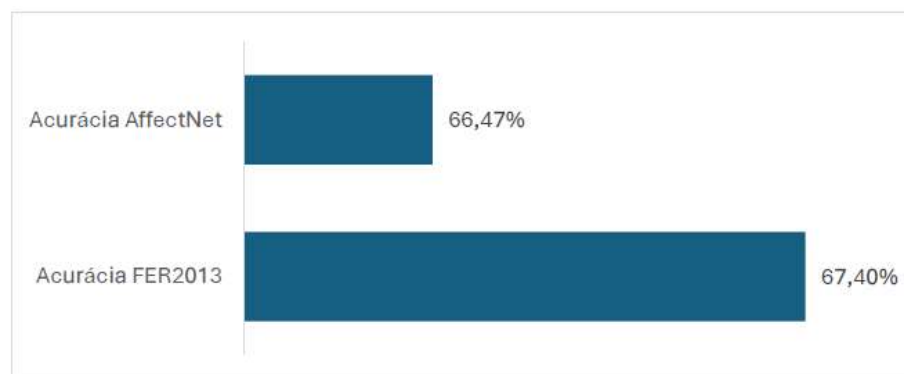
O mesmo ocorreu com as classes de neutralidade (2196 acertos) e tristeza (1319) onde a maior concentração de valores manteve-se em sua diagonal principal. Por outro lado, na expressão facial de surpresa houve confusão com a classe de neutralidade, do mesmo modo que a expressão de medo foi classificada incorretamente e de modo recorrente, como surpresa ou tristeza; por fim, a expressão de raiva ocasionou em tristeza e nojo para algumas classificações.

Figura 25 - Matriz de Confusão (*AffectNet*)

Fonte: Elaborado pela autora.

Em resumo, com todos os testes realizados, é possível observar o gráfico na Figura 26 obtido para a acurácia dos dois modelos. A acurácia do modelo treinado com a base de dados FER2013 indicou um melhor desempenho.

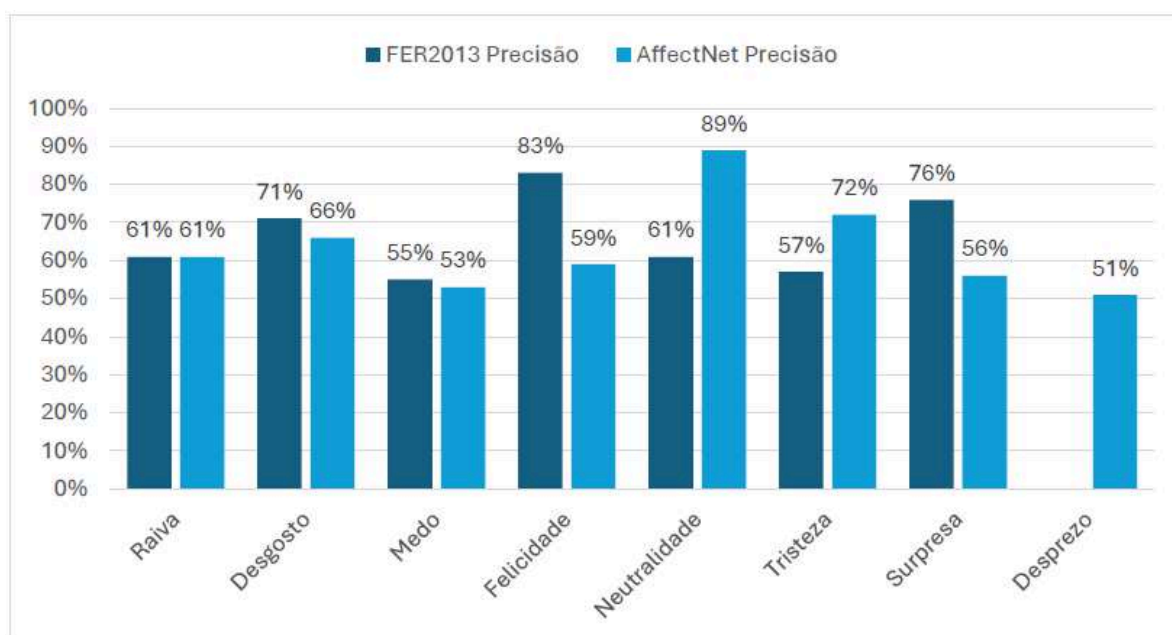
Figura 26 - Gráfico de Comparação dos Desempenhos



Fonte: Elaborado pela autora.

Outro gráfico foi gerado para a análise comparatória entre as precisões obtidas para cada classe, em cada modelo treinado. De modo geral, com exceção das classes “Neutro”, “Tristeza” e “Desprezo” (essa última emoção não está disponível em ambos os conjuntos de dados), o desempenho com o FER2013 foi superior ao do *AffectNet*. Isso pode ser observado na Figura 27.

Figura 27 - Gráfico de Comparação das Precisões (Classes)



Fonte: Elaborado pela autora.

4.3 Classificação de Emoções: Sinais de Voz

Para a construção do classificador de emoções por meio dos sinais de voz, a metodologia CRISP-DM será seguida e, como apresentado na Seção 3.3.2, entre os *datasets* utilizados para o treinamento se tem o RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*), que disponibiliza registros de 24 atores profissionais, sendo 12 homens e 12 mulheres, falando expressões em entonações calmas, felizes, tristes, irritadas, com medo, surpresa e desgosto, além de registros cantados em entonações calmas, felizes, tristes, raivosas e temerosas (LIVINGSTONE, 2018).

Inicialmente, foi necessário compreender a estrutura dos arquivos disponibilizados e as informações contidas. Em resumo, os arquivos (1440 sinais de voz) apresentavam os seguintes identificadores/rótulos:

- Modalidade (03 = apenas áudio);
- Canal vocal (01 = fala, 02 = música);
- Emoção:
 - o 01 = neutro;
 - o 02 = calmo;
 - o 03 = feliz;
 - o 04 = triste;
 - o 05 = zangado;
 - o 06 = medo;
 - o 07 = nojo;
 - o 08 = surpresa.
- Intensidade emocional:
 - o 01 = normal;
 - o 02 = forte.
- Frase:
 - o 01 = *"Kids are talking by the door"*;
 - o 02 = *"Dogs are sitting by the door"*.
- Repetição (01 = 1ª repetição, 02 = 2ª repetição):
 - o Ator (01 a 24):
 - Números Pares = Mulheres;
 - Números Ímpares = Homens.

Ao analisar a distribuição dos dados disponíveis, nota-se que a base consiste em 720 sinais de voz de homens e 720 sinais de voz de mulheres. Os áudios possuem uma taxa de amostragem média de 48 kHz (48 000 amostras por segundo) e na Tabela 11, a distribuição de imagens para as oito classes disponíveis é fornecida. Com exceção da classe neutra, percebe-se que a distribuição é uniforme, caracterizando um conjunto de dados balanceado.

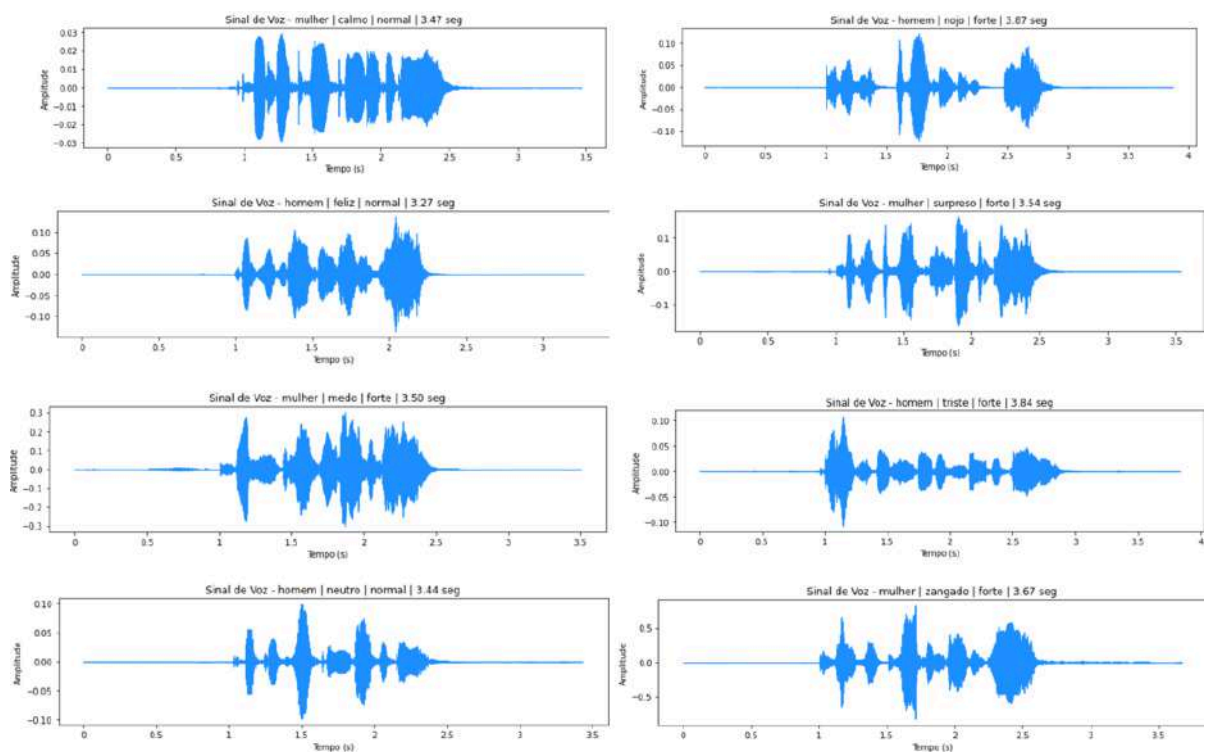
Tabela 11 - Distribuição de Imagens por Classe (RAVDESS)

Emoções	Quantidade de Imagens
Raiva	192
Calmo	192
Desgosto	192
Medo	192
Felicidade	192
Neutralidade	96
Tristeza	192
Surpresa	192

Fonte: Elaborado pela autora.

Para um melhor entendimento, durante a análise exploratória dos dados, os gráficos de onda foram plotados buscando visualizar a intensidade/amplitude da onda sonora ao longo do tempo. Na Figura 28 é possível observar uma amostra de oito sinais selecionados aleatoriamente para representar cada emoção e demonstrar as diferenças existentes para cada situação.

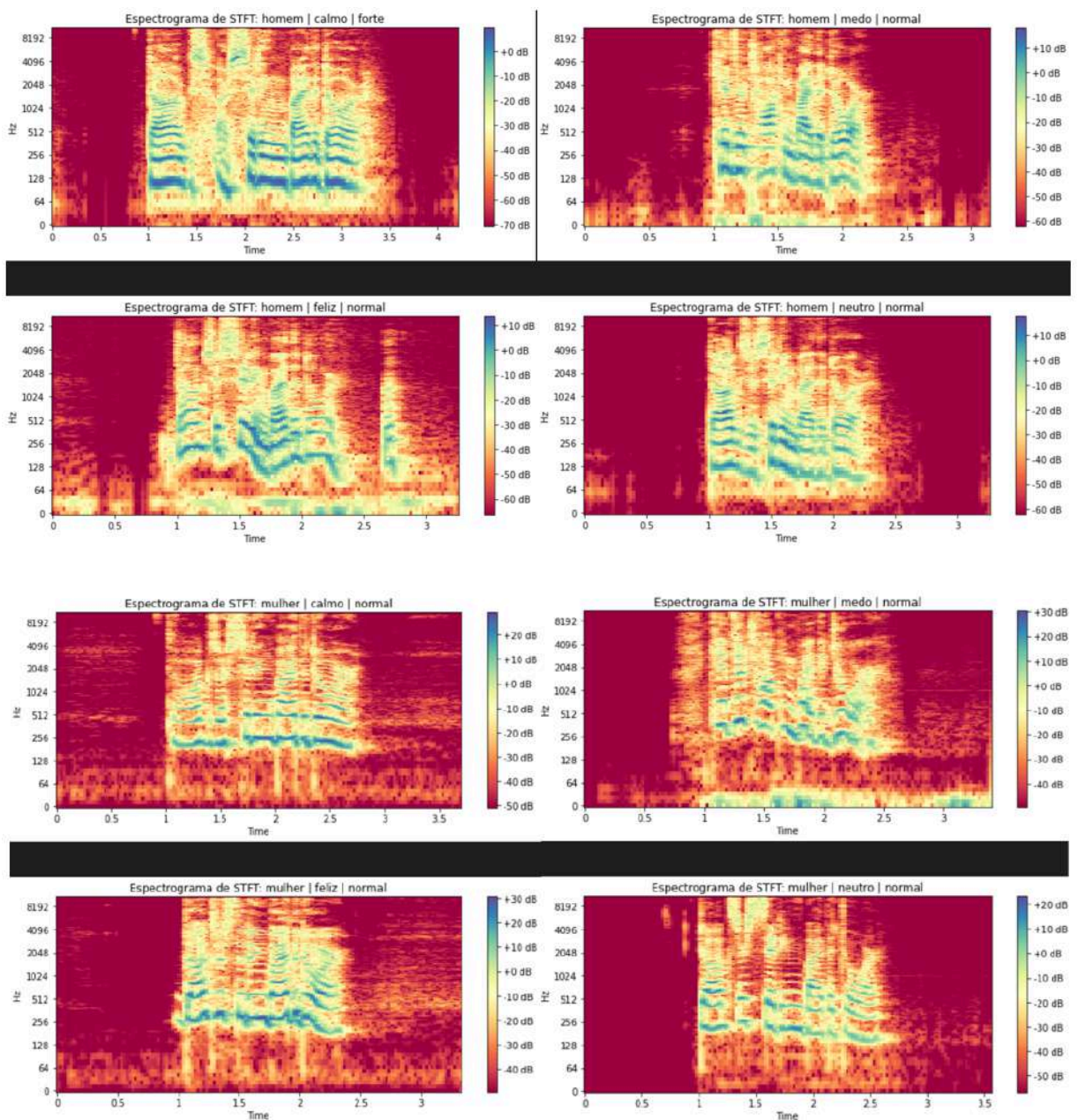
Figura 28 - Amostra dos Sinais de Voz para cada Classe



Fonte: Elaborado pela autora.

Em sequência, visando uma representação visual da Transformada de Fourier de Curto Prazo (STFT), o sinal é dividido em segmentos menores (trabalhando em janelas), foi aplicado a Transformada de Fourier (FFT) a cada segmento, uma matriz foi criada e os resultados combinados para gerar uma representação tempo-frequência do sinal, ou seja, um mapa de calor. Na Figura 29 é possível analisar os resultados obtidos para sinais selecionados e plotados aleatoriamente. Em resumo, cada ponto do gráfico representa a amplitude de uma determinada frequência em um determinado momento em termos de cores e intensidade.

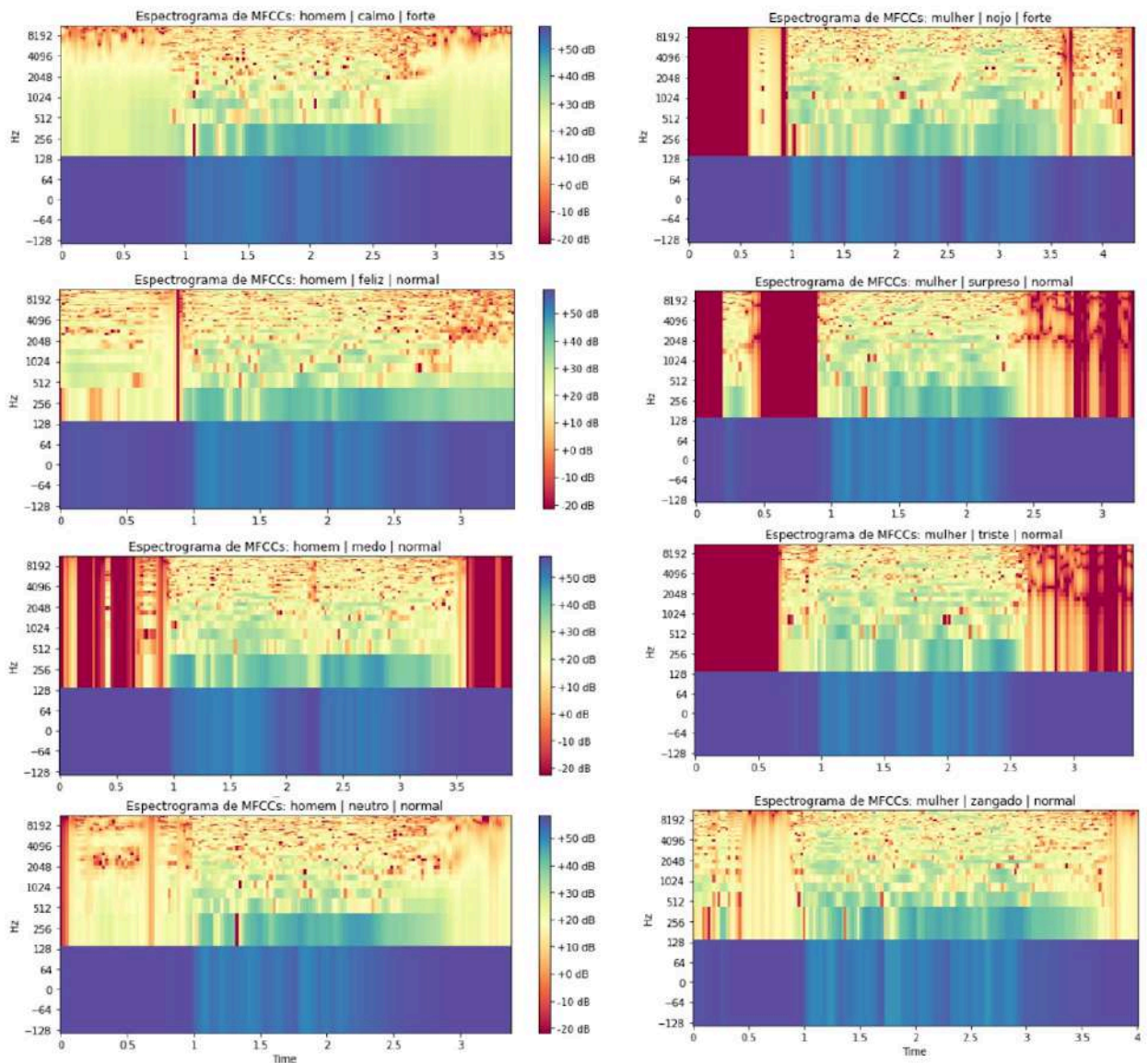
Figura 29 - Espectrogramas de STFT



Fonte: Elaborado pela autora.

Dado que para a modelagem a extração de recursos/features é fundamental ao se trabalhar com sinais de voz, foi importante analisar também os espectrogramas de MFCCs (Coeficiente Cepstral de Frequência Mel), que basicamente realiza uma normalização nos dados com a aplicação da função *mfcc* da biblioteca *librosa*, capturando as características perceptíveis do sinal de voz, como por exemplo, onde está concentrada a energia do sinal. Para as plotagens geradas na Figura 30, foram calculados 40 coeficientes MFCC por frame de análise para um sinal de voz feminino e um masculino, como tonalidade forte (apesar de representarem emoções totalmente opostas: raiva e calma).

Figura 30 - Espectrograma de MFCCs (Coeficiente Cepstral de Frequência Mel)



Fonte: Elaborado pela autora.

Após a extração das características MFCCs dos sinais de voz, para a modelagem os dados foram divididos em treinamento, validação e teste. Basicamente, 80% dos dados foram destinados para o treino da rede neural e, dos 20% restantes, 40% será utilizado para fazer o teste após o treino da rede neural, e o restante para validação.

A rede neural desenvolvida consiste em duas camadas de convolução seguidas por uma camada de *MaxPooling* e uma de *Dropout*, a fim de selecionar o maior/principal valor de cada parte da imagem (reduzindo assim as dimensões, mas mantendo características importantes) e em sequência, adicionar uma nova camada de convolução e *MaxPooling* novamente. Por fim, para prevenir o *overfitting*, isto é, não permitir que o algoritmo se adapte muito aos dados de treinamento é adicionado o *Dropout*, uma camada *Flatten* para transformar as saídas (matrizes) em um vetor. Na Figura 31 para visualizar o resumo das camadas e parâmetros adotados.

Figura 31 - Resumo do Modelo para Sinais de Voz

Layer (type)	Output Shape	Param #
conv1d_6 (Conv1D)	(None, 36, 64)	384
conv1d_7 (Conv1D)	(None, 36, 128)	41,088
max_pooling1d_4 (MaxPooling1D)	(None, 7, 128)	0
conv1d_8 (Conv1D)	(None, 7, 256)	164,096
max_pooling1d_5 (MaxPooling1D)	(None, 1, 256)	0
dropout_2 (Dropout)	(None, 1, 256)	0
flatten_2 (Flatten)	(None, 256)	0
dense_4 (Dense)	(None, 64)	16,448
dense_5 (Dense)	(None, 8)	520
activation_2 (Activation)	(None, 8)	0

Fonte: Elaborado pela autora.

Finalizado o treinamento do modelo com 50 épocas, a acurácia obtida foi de aproximadamente, 97,17% com uma perda de 0,1134. Ao avaliar o modelo treinado com o conjunto de dados de teste, a acurácia resultante foi igual a 70,35%. Na Tabela 12 é possível visualizar o desempenho alcançado com o treinamento.

Tabela 12 - Métricas das Classes após o Treinamento (RAVDESS)

RAVDESS		
Emoções	Precisão	Recall
Calmo	71%	87%
Felicidade	60%	88%
Medo	58%	61%
Neutralidade	46%	55%
Desgosto	85%	79%
Surpreso	82%	61%
Tristeza	68%	68%
Raiva	92%	57%

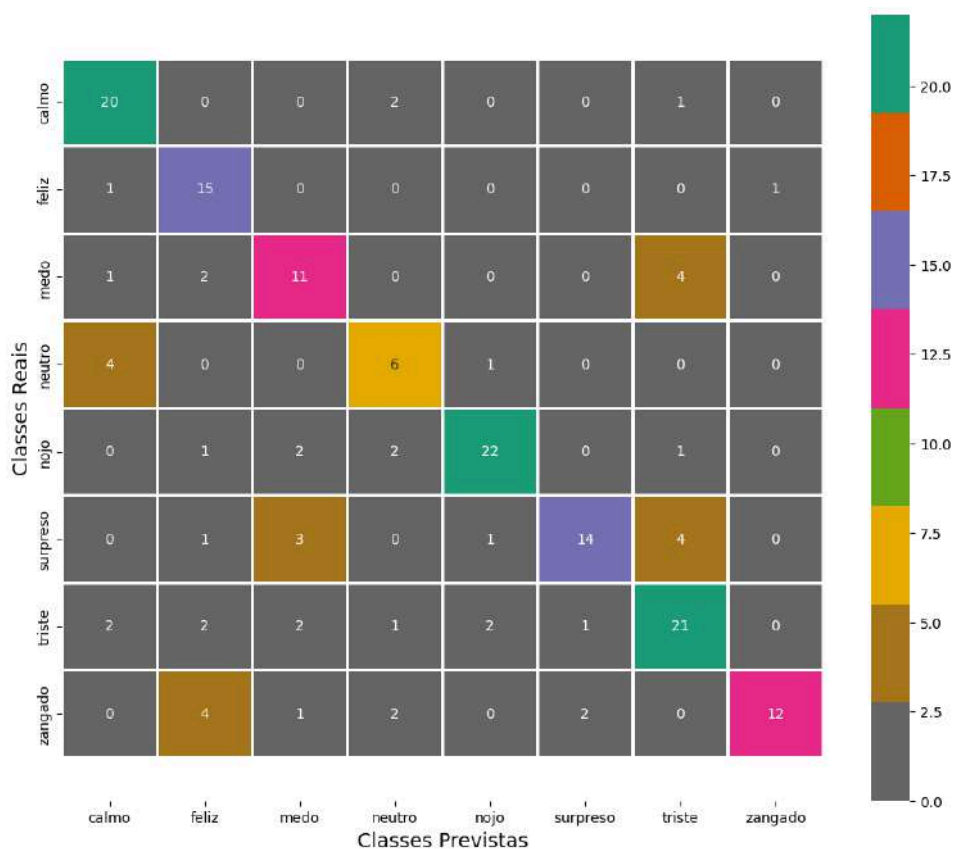
Fonte: Elaborado pela autora.

Para a classe Felicidade, o algoritmo conseguiu identificar 88% dos sinais de voz felizes com uma precisão de 60%; para a classe calmo, o algoritmo pôde identificar 87% dos respectivos sinais com uma precisão de 71%. Por outro lado, a emoção/classe que apresentou um desempenho inferior foi o sinal de voz que expressa a neutralidade, onde o algoritmo conseguiu identificar apenas 55% das respectivas vozes neutras, com uma precisão de 46%.

Novamente, a matriz de confusão foi utilizada para avaliar o desempenho do modelo, tornando possível a análise da distribuição dos acertos e erros por classe. Como apresentado na Figura 32, o modelo obteve êxito na classificação dos sinais de voz que expressavam calma, felicidade, tristeza e nojo.

Por sua vez, a classe de surpresa foi confundida em algumas situações com tristeza e medo, e do mesmo modo, a entonação de medo ocasionou em algumas classificações incorretas como tristeza. Além disso, a neutralidade foi classificada incorretamente como calma (o que de certo modo é possível considerar como um erro aceitável, visto as similaridades entre as duas classes).

Figura 32 - Matriz de Confusão (RAVDESS)



Fonte: Elaborado pela autora.

4.4 Resultado da Interface Web Implementada

Como comentado nas seções anteriores, a interface e sua estética previamente planejadas sofreram algumas mudanças conforme o processo de criação foi ocorrendo.

Primeiramente, um nome foi atribuído ao sistema/plataforma de videoconferência: “*Meet.Emotions*”. De acordo com o *Cambridge Dictionary*²⁵, a palavra em inglês *Meet* significa encontrar, podendo ser usada para esperar em algum lugar pela chegada de alguém ou estar relacionado com a ideia de ver e/ou falar com alguém pela primeira vez. Por sua vez, a palavra *Emotions* significa em português sentimentos ou emoções como, por exemplo, amor, raiva, felicidade, tristeza, entre outros. Com isso, a combinação das duas palavras entrega

²⁵ Cambridge Dictionary. Disponível em:
<<https://dictionary.cambridge.org/pt/dicionario/ingles-portugues/>>

exatamente a intenção e o objetivo por trás da criação desta plataforma e suas respectivas funcionalidades.

Para o logotipo, a cor azul foi selecionada para o fundo e as letras na tonalidade branca foi escolhida para um maior destaque ao nome do sistema. Entre os elementos adicionados, o quadrado busca representar a caixa delimitadora que é utilizada para determinar a área de detecção e região de interesse dos frames durante a detecção facial. Com a finalidade de representar a detecção de sinais de voz, uma marca d'água contendo esse elemento foi adicionada ao fundo da palavra *Emotions*. Na Figura 33 é possível observar o resultado.

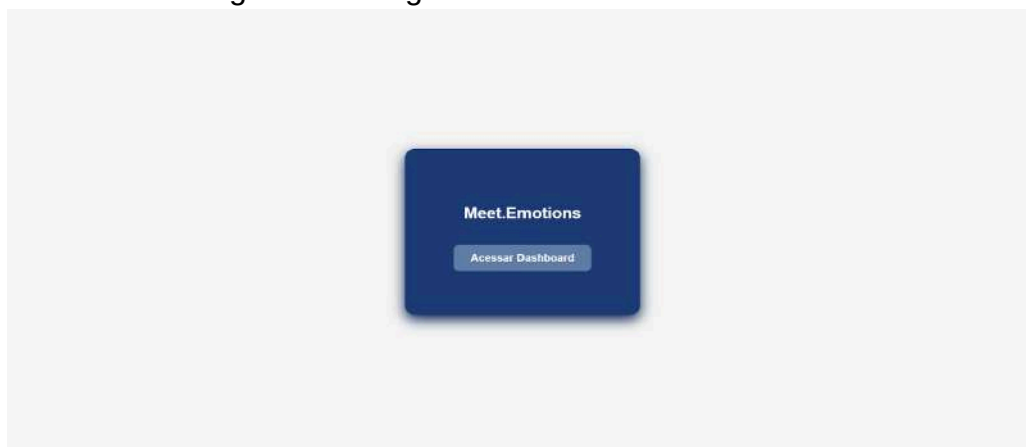
Figura 33 - Logotipo do Sistema (*Meet.Emotions*)



Fonte: Elaborado pela autora.

Inicialmente, o profissional recebe um link de acesso a plataforma *Meet.Emotion*. Como se pode visualizar na Figura 34, a página consiste em uma caixa retangular azul sombreada posicionada ao centro, contendo o nome do sistema e um botão em azul claro que ao clicá-lo permite o acesso ao dashboard.

Figura 34 - Página de Acesso ao Profissional



Fonte: Elaborado pela autora.

Ao clicar no botão “Acessar *Dashboard*”, o profissional será encaminhado para uma outra página, que basicamente é a plataforma do respectivo profissional. Como comentado no início desta seção, algumas mudanças foram necessárias por fins estéticos e de funcionalidade. Na região localizada à direita da tela inicial, na parte superior, o que antes seria um fundo cinza claro com um modelo de logotipo da UFABC centralizado, foi alterado para um azul escuro com o logotipo *Meet.Emotions* no centro.

Do mesmo modo, a lateral esquerda mantém a proposta inicial, entretanto, dois botões foram alterados: o botão de edição de perfil foi substituído pelo botão de “*Dashboard*”, que leva o usuário em um clique para a respectiva seção (apresentada na Figura 35 – área em branco que contém três blocos). A alteração se deu devido a dificuldades enfrentadas para o desenvolvimento e a viabilidade no estágio atual do projeto, para uma possível atualização de informações como os nomes e as fotos dos usuários e a necessidade do uso de um banco de dados para estes casos.

Figura 35 - Página Inicial da Plataforma



Fonte: Elaborado pela autora.

O outro botão com funcionamento alterado é o “Sobre”, que leva o profissional para a seção que contém informações sobre o objetivo do sistema, instituição de ensino e responsáveis pela criação, como apresentado na Figura 36. O motivo da mudança é a existência de um botão para acesso ao guia do usuário na área de

Dashboard, logo, a presença de dois botões apresentando a mesma funcionalidade tornava-se desnecessário.

Figura 36 - Seção "Sobre" da Plataforma do Profissional



Fonte: Elaborado pela autora.

Ao clicar no botão calendário ou simplesmente usar o scroll do mouse para baixo, a página a ser visualizada é apresentada na Figura 37. Para a criação do calendário, a biblioteca *open-source react-calendar*²⁶ foi adicionado ao projeto após executar o comando `npm install react-calendar`. Ao importá-lo torna-se possível selecionar os dias, meses e anos, diversos idiomas são suportados e basta chamar no código a linha `<Calendar />` e `onChange` para adquirir novos valores, conforme recomendado em sua documentação. Além disso, visando a estética definida para a plataforma, o estilo pôde ser personalizado com o auxílio da seguinte importação: `import 'react-calendar/dist/Calendar.css'`.

²⁶ React-Calendar. Documentação Disponível em: <<https://www.npmjs.com/package/react-calendar>>

Figura 37 - Seção "Calendário" da Plataforma do Profissional

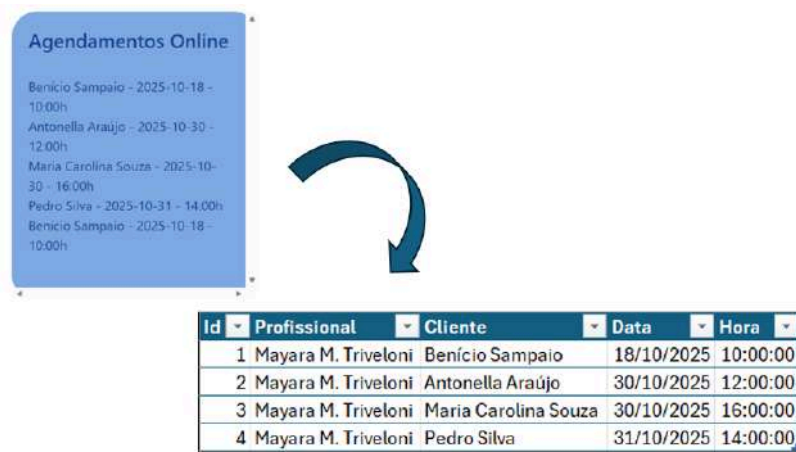


Fonte: Elaborado pela autora.

Além dos detalhes descritos anteriormente, ao selecionar uma data do calendário, os possíveis agendamentos são disponibilizados logo abaixo. Estes agendamentos são capturados de uma planilha de agendamentos preenchida pelo profissional e atualmente armazenado na pasta *public* do *front-end*, o mesmo ocorre para os agendamentos do ícone “Agendamentos Online”, posicionado na seção de *Dashboard*. A planilha preenchida pode ser visualizada na Figura 38. No caso o profissional preenche as colunas ID, Profissional, Cliente, Data e Hora. Com o auxílio da biblioteca JavaScript *Papa Parse* ²⁷ que analisa arquivos no formato CSV (*Comma-Separated Values*) localmente ou pela rede e transforma-os em uma estrutura de *arrays*/lista de objetos.

²⁷ Papa Parse. Documentação disponível em: <<https://www.npmjs.com/package/papaparse>>

Figura 38 - Planilha de Agendamentos

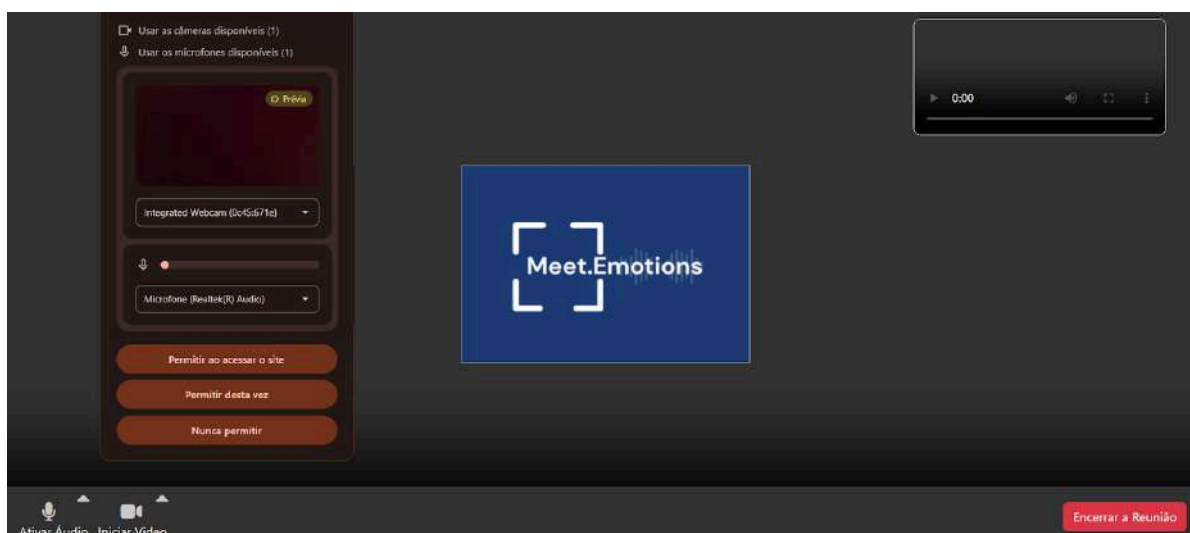


Id	Profissional	Cliente	Data	Hora
1	Mayara M. Triveloni	Benício Sampaio	18/10/2025	10:00:00
2	Mayara M. Triveloni	Antonella Araújo	30/10/2025	12:00:00
3	Mayara M. Triveloni	Maria Carolina Souza	30/10/2025	16:00:00
4	Mayara M. Triveloni	Pedro Silva	31/10/2025	14:00:00

Fonte: Elaborado pela autora.

Um link para o acesso do cliente à sala de conferência é gerado para o profissional e, com isso, ele deve enviar ao cliente pelo meio de comunicação desejado (via Email, mensagem ou outras mídias sociais). Quando o cliente acessa o link recebido, ele automaticamente entra na sala de reunião virtual e permanece em uma “sala de espera” até que o cliente conceda as permissões de uso de suas mídias (câmera e microfone), como apresentado na Figura 39. Caso o cliente queira sair da sala de reunião, basta clicar no botão vermelho “Encerrar a Reunião”.

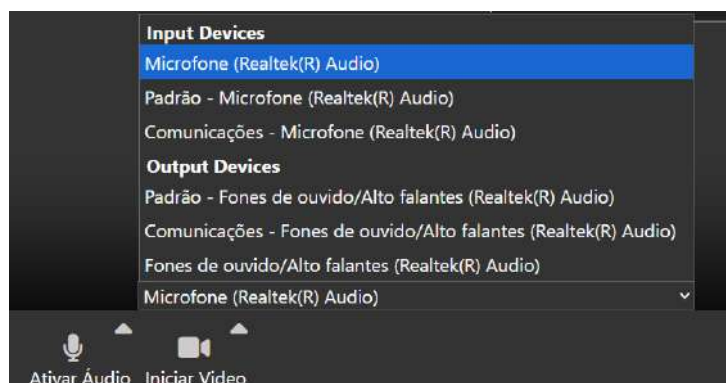
Figura 39 - Sala de Reunião Virtual do Cliente (Permissões)



Fonte: Elaborado pela autora.

Em seguida, há a necessidade de ativação dos dois botões posicionados no canto inferior esquerdo da tela: “Ativar Áudio” e “Iniciar Vídeo”. Para os dois botões, existe a possibilidade de escolha dos dispositivos de entrada e saída do microfone e a Webcam ao clicar no triângulo do canto superior direito, como apresentado na Figura 40.

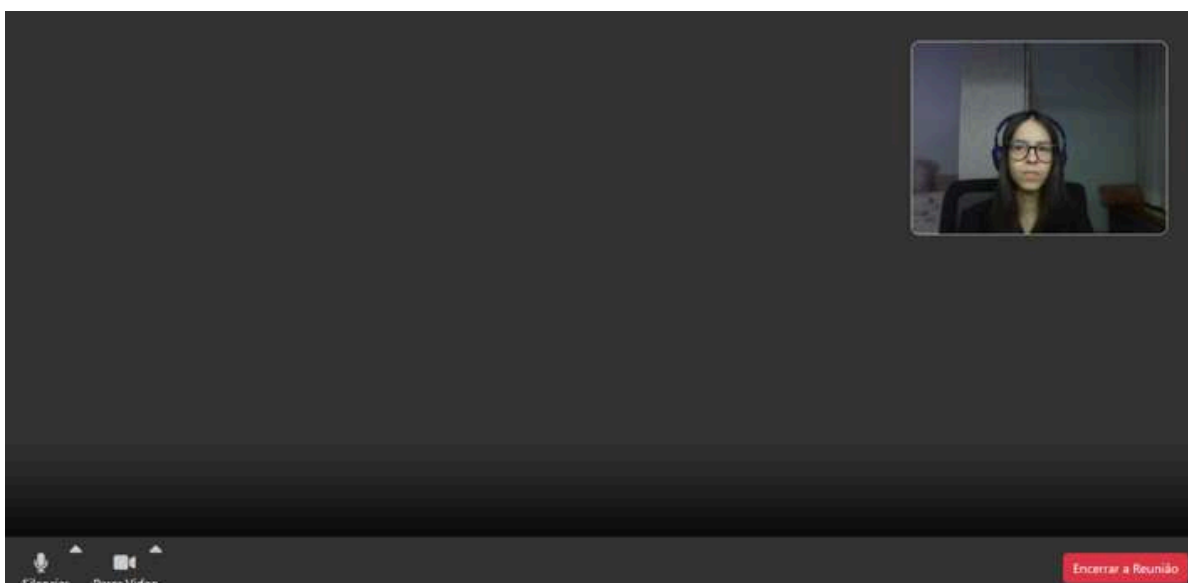
Figura 40 - Menu Suspensão dos Botões



Fonte: Elaborado pela autora.

Com as escolhas definidas e os botões ativados, no canto superior direito o vídeo capturado pela Webcam é transmitido, como apresentado na Figura 41.

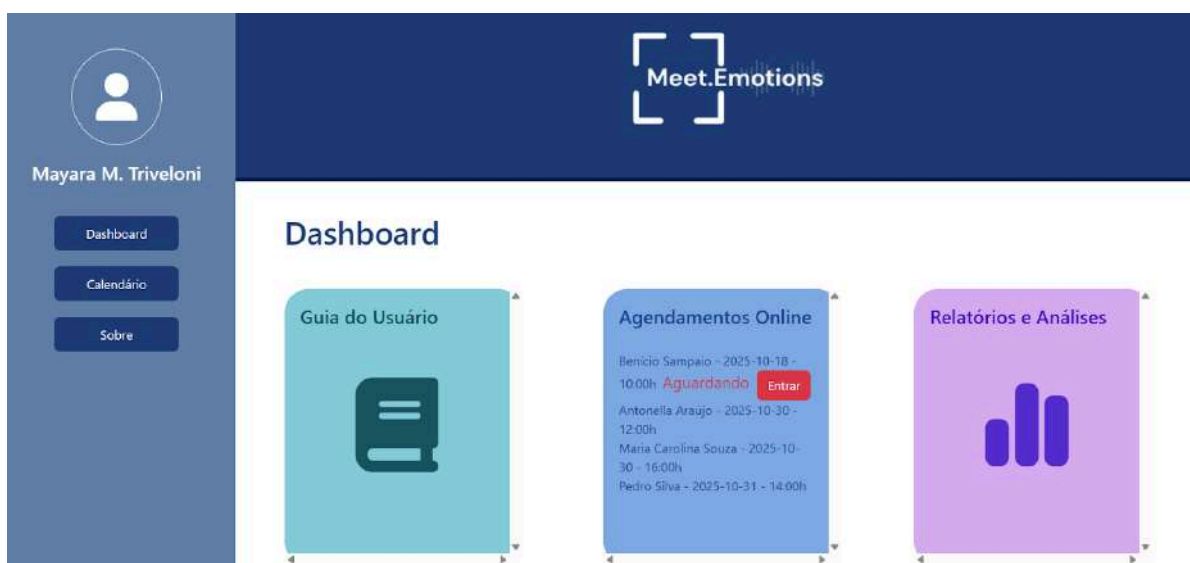
Figura 41 - Sala de Reunião do Cliente (Câmera e Microfone Ativados)



Fonte: Elaborado pela autora.

Uma vez que o cliente está na sala de reuniões do *Meet.Emotions*, a plataforma do profissional será atualizada e no bloco “Agendamentos Online” um botão vermelho escrito “Entrar” aparecerá, assim como a mensagem “Aguardando”. Na Figura 42 é possível observar essa mudança.

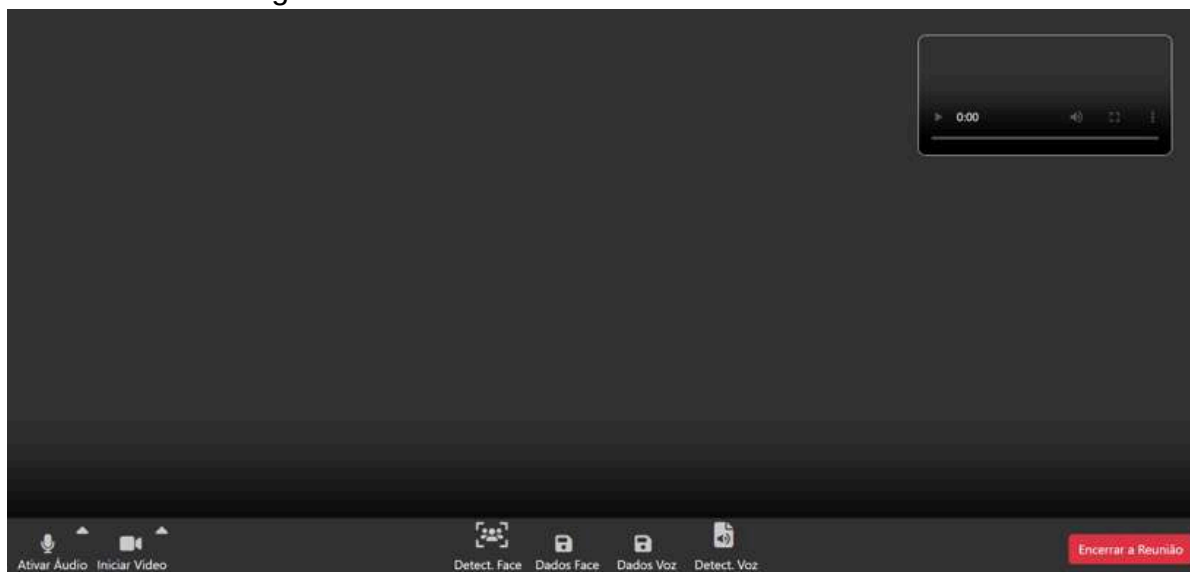
Figura 42 - Plataforma do Profissional (Cliente Aguardando)



Fonte: Elaborado pela autora.

Se o profissional clicar no botão “Entrar”, ele será direcionado para a sala de reunião em que o cliente se encontra, entretanto, para o respectivo profissional mais funcionalidades serão disponibilizadas, como é possível observar os novos botões na parte inferior da Figura 43. Mais uma vez, até que as permissões de compartilhamento do microfone e Webcam sejam concedidas, o profissional estará em uma “sala de espera” com a mensagem “{nome do cliente} está aguardando na sala de espera. Ative o vídeo e o áudio para iniciar a reunião.”. No caso, o nome do cliente utilizado é fictício.

Figura 43 - Sala de Reunião Virtual do Profissional



Fonte: Elaborado pela autora .

Concedida as permissões de uso de suas mídias (câmera e microfone) o vídeo capturado pela Webcam do profissional será transmitido em um retângulo no canto superior direito e, automaticamente, o vídeo do cliente será transmitido na tela central em tempo real (o mesmo ocorre na situação oposta: para o cliente, o vídeo transmitido do profissional será transmitido em sua tela central). Tudo isso pode ser visualizado na Figura 44, sendo importante ressaltar o uso de duas Webcams diferentes para os testes.

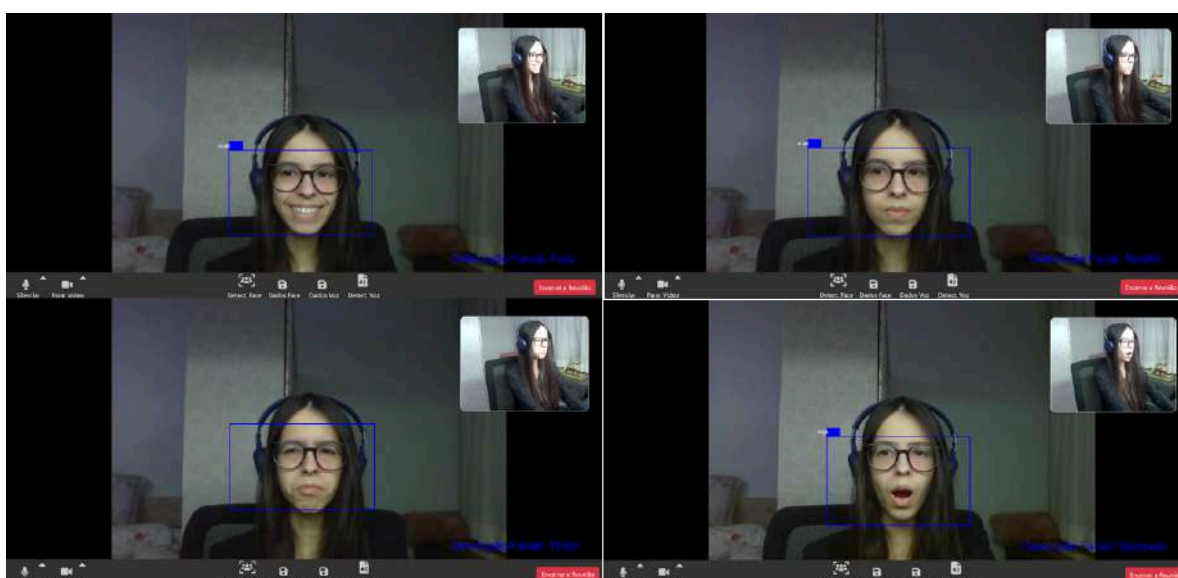
Figura 44 - Sala de Reunião Virtual do Profissional (Comunicação Iniciada)



Fonte: Elaborado pela autora.

Logo, já é possível realizar uma comunicação entre o cliente e o profissional. Caso seja do interesse do profissional, ele pode optar por clicar no botão “Detect. Face” para ativar a função de detecção facial em tempo real e consequentemente auxiliar em sua condução de reunião. Ao clicar nesse botão, a detecção da emoção é impressa no canto inferior direito da tela, podendo ser observado o resultado de quatro emoções na Figura 45, enquanto a região de interesse é desenhada sob a face do cliente até que a função seja desativada.

Figura 45 - Sala de Reunião Virtual do Profissional (Detecção Facial)



Fonte: Elaborado pela autora.

Como comentado anteriormente na Seção 4.2, o modelo de classificação de imagens faciais foi treinado na linguagem Python com a aplicação da API (*Application Programming Interface*) do TensorFlow `tf.keras`²⁸. Basicamente, o TensorFlow²⁹ é uma biblioteca de código aberto desenvolvido em 2015 pelo *Google Brain Team*, que buscava facilitar e flexibilizar a criação de modelos de *Machine Learning*.

No caso, o modelo de rede neural foi salvo no formato *Hierarchical Data Format* versão 5 (HDF5³⁰), que permite armazenar grandes quantidades de dados na forma de matrizes multidimensionais. Devido a necessidade de aplicação React/JavaScript, o modelo precisou ser convertido para um formato que possibilitasse a interpretação ao navegador. A tentativa inicial foi a conversão do modelo para o formato TensorFlow.js, uma biblioteca para *Machine Learning* no JavaScript que tem como funcionalidade a execução de modelos existentes ou a conversão de modelos TensorFlow em Python para que sejam executados no navegador ou pelo Node.js.

Em um notebook do Google Colab, a biblioteca *tensorflowjs* foi instalada e após utilizar o comando `tfjs.converters.save_keras_model()` a conversão resultou em um arquivo no formato JSON (*JavaScript Object Notation*) e em dois arquivos no formato BIN (Binário). Feito isso, a pasta gerada foi movida para a pasta *public* do projeto e, em sequência foi possível carregar o modelo com o comando `tf.loadLayersModel()`.

Aparentemente, o modelo estava sendo carregado corretamente, mas ao tentar realizar as predições com os frames capturados, pré-processados (imagem única, redimensionado para 48x48px, pixels normalizados e nos padrões de entrada necessários para escala de cor/quantidade de canais) e aplicar a função `tf.browser.fromPixels()` responsável por criar um Tensor com os valores de pixels da região de interesse, alguns erros foram gerados. Após investigar e buscar por justificativas nas documentações, foi possível concluir que o modelo foi convertido com uma versão Keras/TensorFlow incompatível com o conversor utilizado.

Visando corrigir esse problema, um novo ambiente foi criado para utilizar versões do TensorFlow inferiores a 2.15. Ao retreinar o modelo e salvá-lo com essas

²⁸ Keras. Documentação disponível em: <<https://www.tensorflow.org/guide/keras?hl=pt-br>>

²⁹ TensorFlow. Disponível em: <<https://www.tensorflow.org/?hl=pt-br>>

³⁰ HDF5. Documentação disponível em: <<https://docs.fileformat.com/pt/misc/h5/>>

novas condições, novamente a função para conversão era aplicada e novos arquivos eram gerados. Mesmo com essa nova tentativa, o erro detectado permanecia (com o surgimento de outras inconsistências de estrutura do novo arquivo carregado). Considerando tudo isso, a busca por uma nova solução foi necessária.

A alternativa encontrada foi a conversão do modelo para ONNX³¹ (*Open Neural Network Exchange*) que evita as incompatibilidades das versões treinadas e funciona com modelos Keras, por exemplo, independente do framework. Além disso, contém as camadas e as conexões definidas (grafo computacional) e parâmetros da rede em seu arquivo. Considerando tudo isso, novamente com o auxílio do Google Colab, a versão 1.16.1 do *tf2onnx* foi instalada e a função *tf2onnx.convert.from_keras()* foi aplicada para a conversão do modelo previamente carregado.

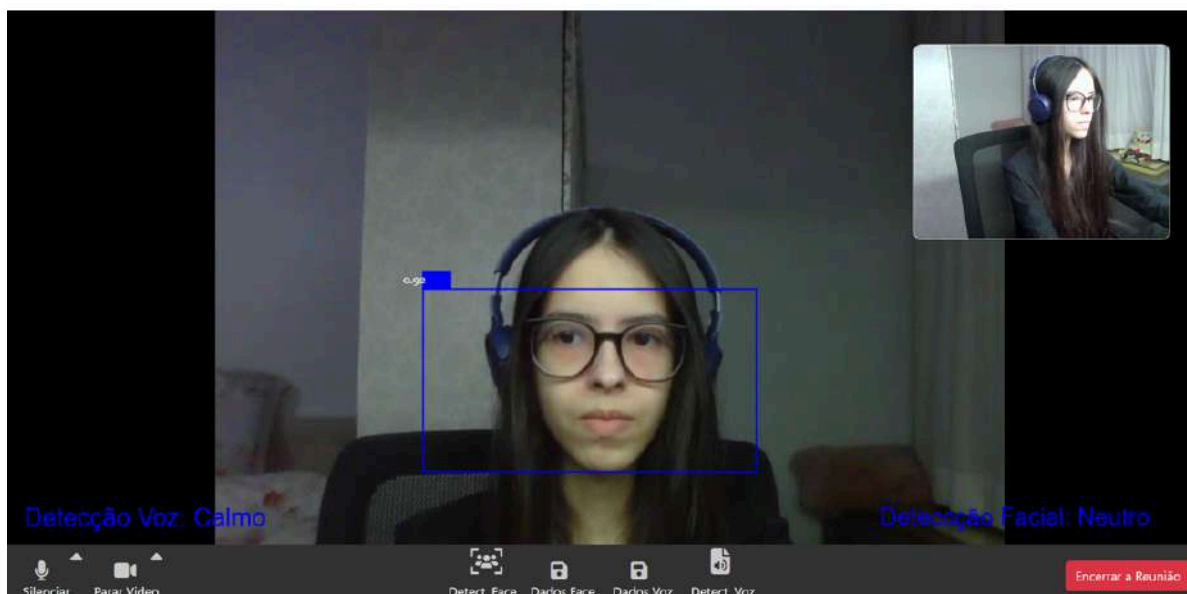
Com o modelo gerado no novo formato, a função utilizada para o carregamento do modelo foi: *ort.InferenceSession.create()*, que gerencia o grafo do modelo e aloca os recursos necessários após a execução. A resposta obtida com a predição foi satisfatória e solucionou o problema antes enfrentado.

Além disso, é importante ressaltar como a detecção facial pôde ser realizada nesta aplicação. Para todos os testes realizados durante o treino do modelo o *Haar Cascade* foi a opção escolhida, entretanto, para o funcionamento adequado foi necessário recorrer a uma API *JavaScript* para o uso no navegador que se chama *face-api.js*. O responsável pela criação foi Vicente Mühler que implementou um *Tiny Face Detector* otimizado que utiliza convoluções separáveis em profundidade ao invés de convoluções regulares, o que o torna um detector de faces muito mais rápido, porém menos preciso. Além disso, as respectivas redes retornam as caixas delimitadoras do rosto com a probabilidade de cada região de interesse realmente mostrar um rosto (MÜHLER, 2018).

Retornando para o funcionamento da plataforma de vídeo chamada, caso o profissional opte pela análise dos sinais de voz do cliente, basta clicar no botão “Detect. Voz” e, automaticamente, no canto inferior esquerdo a respectiva emoção será impressa, como analisado na Figura 46.

³¹ ONNX. Documentação disponível em: <<https://onnx.ai/>>

Figura 46 - Sala de Reunião Virtual do Profissional (Detecção dos Sinais de Voz)



Fonte: Elaborado pela autora.

O mesmo processo de conversão se fez necessário para a aplicação do modelo de detecção dos sinais de voz, ou seja, a conversão do modelo para ONNX foi fundamental para o funcionamento sem empecilhos.

É importante ressaltar que durante o processo de habilitação/desabilitação dos botões de detecção, as emoções, o dia e a hora da captura são armazenadas em uma lista. Com estas informações coletadas, a criação do arquivo no formato CSV (*Comma Separated Values*) é possibilitada. Portanto, caso seja de preferência do profissional ele conseguirá realizar o Download do relatório da reunião clicando nos botões “Dados Face” ou “Dados Voz”, no momento que desejar.

Em seguida, para a análise dos dados coletados basta retornar para a página inicial da plataforma do profissional e clicar no ícone roxo “Relatórios e Análises” que levará o usuário para a página ilustrada na Figura 47.

Figura 47 - Página de Relatórios e Análises



Fonte: Elaborado pela autora.

Clicando no botão “Escolher Arquivo”, o profissional poderá escolher o arquivo desejado para análise e, uma vez carregado, o gráfico contendo a linha do tempo das emoções detectadas será impressa na tela e a frequência das emoções detectadas poderá ser observado conforme apresentado na Figura 48 e Figura 49. Para a criação dos gráficos, a biblioteca open-source *Recharts*³² foi utilizada em conjunto de alguns dos seus componentes como, *LineChart* e *Line* para o gráfico de linha e *Pie* e *PieChart* para o gráfico de pizza/rosca. Resumidamente, a entrada recebida foi em formato de matriz de objetos, conforme especificado na documentação.

Figura 48 - Página de Relatórios e Análises (Linha do Tempo)



Fonte: Elaborado pela autora.

³² Recharts. Documentação disponível em: <<https://recharts.github.io/en-US/examples/>>

Figura 49 - Página de Relatórios e Análises (Frequência das Emoções)

Frequência das Emoções Detectadas



Fonte: Elaborado pela autora.

4.4.1 Métricas de Desempenho da Interface Web

Visando obter o intervalo necessário para o processamento dos modelos, de modo que a entrada processada seja recebida e, em seguida, retorne a classificação adequada, a função *performance.now()* foi utilizada. Com isso, foi possível obter a marcação temporal em milissegundos do momento anterior e posterior a predição do modelo, sendo o cálculo da diferença dos dois valores o tempo de inferência. Em resumo, essa métrica reflete o desempenho computacional do modelo durante a execução em tempo real e demonstra a eficiência da integração entre o modelo e a aplicação *React*. Resumidamente, o tempo médio obtido para os dois modelos gerados (facial e de voz) foi de aproximadamente 27,4 ms e 28,6 ms, apresentando um desempenho satisfatório e quase que em tempo real.

Além disso, com o auxílio da ferramenta *Google Lighthouse*³³, utilizada para avaliar a qualidade de páginas web, verificar desempenho, acessibilidade e SEO (otimização para mecanismos de busca), foi possível verificar métricas voltadas para carregamento e interatividade do profissional com a plataforma web criada.

³³ Google Lighthouse. Disponível em:
<<https://developer.chrome.com/docs/lighthouse/overview?hl=pt-br>>

Para as categorias de “Acessibilidade”, “Boas Práticas” e “SEO”, os resultados foram satisfatórios, sendo 92, 96 e 100% respectivamente, e estando de acordo com os padrões esperados. Para o tempo do primeiro elemento visível exibido ao usuário o valor obtido foi de 0.9 segundos, indicando também um carregamento inicial rápido, logo, o resultado foi satisfatório para a estabilidade visual do layout (*Cumulative Layout Shift*), comprovando a sua usabilidade.

Entretanto, a métrica *Total Blocking Time (TBT)* foi de 330 ms, indicando que houve um bloqueio moderado de processamento devido a execução de scripts, o que de certo modo, impede a interação do usuário por esse período. Isso pode estar relacionado com as operações de carregamento dos modelos, por exemplo. O que trouxe também uma métrica de “Desempenho” moderada de aproximadamente, 68%.

Analisando também a interface da sala de reuniões virtuais, o componente principal levou apenas 0,2 ms para ser renderizado, e os botões de áudio e voz aproximadamente, 0,1 ms, comprovando que a página se encontra otimizada e os componentes principais estão renderizando de maneira rápida.

5 ANÁLISE DOS RESULTADOS

Este capítulo apresenta análises e discussões sobre as opiniões de profissionais obtidas com o levantamento através de entrevistas e formulários elaborados especificamente para esta finalidade de avaliar a importância do uso de reconhecimento de emoções pela expressão facial e pela característica da voz falada por entrevistados e pacientes em sistemas de vídeo conferências.

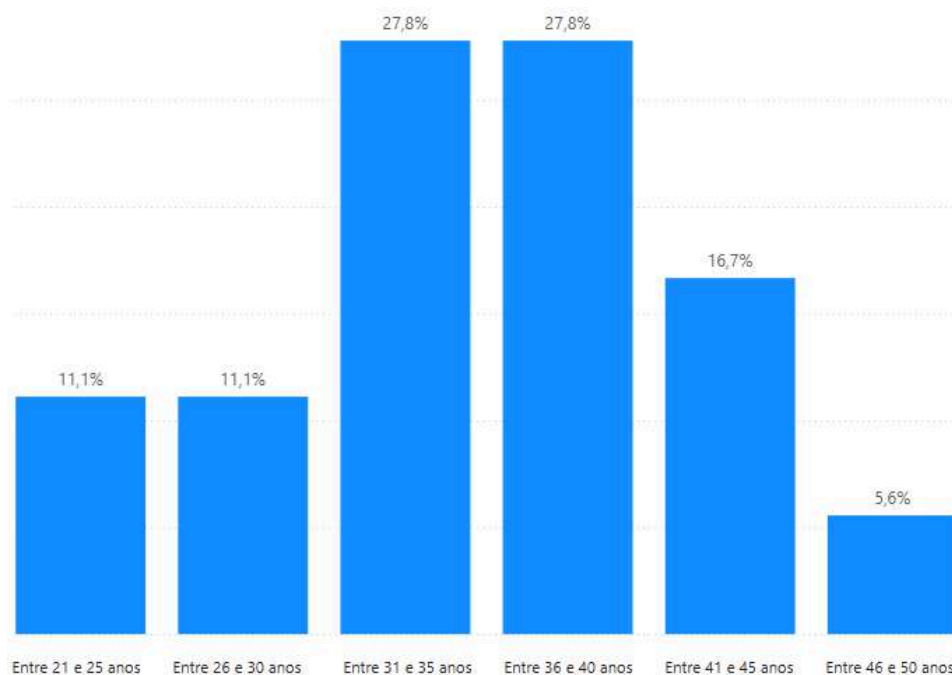
A Interface gráfica Web desenvolvida neste projeto junto com o sistema completo foi submetida a uma sessão experimental realizada por um profissional voluntário, de acordo com o cenário experimental descrito na Seção 3.4, e submetido à avaliação prevista na Seção 3.5. A análise das opiniões obtidas foram realizadas na segunda parte deste presente capítulo.

5.1 Levantamento de Opinião dos Profissionais: Importância da Análise Emocional

Para validar a relevância do tema abordado, um levantamento foi realizado por meio da aplicação de um questionário estruturado fechado, divulgado para o público-alvo do estudo e criado com o auxílio da ferramenta online de gerenciamento de pesquisas “*Google Forms*” (os detalhes do método adotado está disponível na Seção 3.2). Em resumo, 18 profissionais foram alcançados.

Dentre os participantes da amostra, 66,7% atuam na área de recursos humanos e 33,3% correspondem à área de psicologia. Além disso, também foram coletadas informações de característica pessoal durante a pesquisa: na Figura 50 é possível observar que 27,8% correspondem às faixas etárias entre “31 e 35 anos” e “36 e 40 anos”, por sua vez, 16,7% de “41 a 45 anos”.

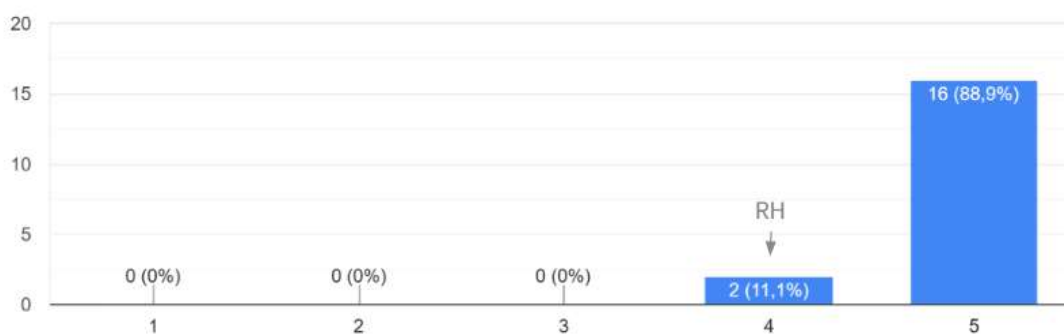
Figura 50 - Faixa Etária dos Participantes



Fonte: Elaborado pela autora.

Em uma escala de 1 a 5, em que 1 é pouco importante e 5 muito importante, 16 recrutadores e psicólogos responderam que a análise das expressões faciais dos candidatos e/ou pacientes durante uma entrevista de emprego ou atendimento é muito importante (uma pequena divergência ocorreu apenas entre os profissionais da área de recursos humanos, onde dois participantes classificaram como 4). A Figura 51 ilustra a distribuição obtida para o nível de importância das análises de expressões faciais.

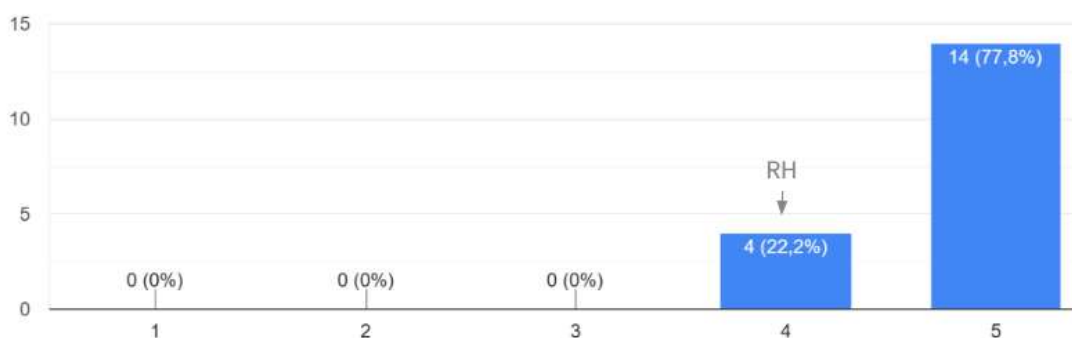
Figura 51 - Nível de Importância das Análises de Expressões Faciais



Fonte: Elaborado pela autora.

A mesma pergunta foi realizada visando o nível de importância da análise da tonalidade de voz dos candidatos e/ou pacientes durante uma entrevista ou atendimento terapêutico. Novamente, em uma escala de 1 a 5, em que 1 é pouco importante e 5 muito importante, 14 recrutadores e psicólogos responderam que a análise da tonalidade de voz dos candidatos durante uma entrevista de emprego ou atendimento é muito importante (mais uma vez, as quatro divergências ocorreram apenas entre os profissionais da área de recursos humanos, que classificaram como 4). Esta distribuição pode ser observada na Figura 52.

Figura 52 - Nível de Importância das Análises da Tonalidade de Voz



Fonte: Elaborado pela autora.

De modo geral, estes resultados demonstraram que estes profissionais (de ambas as áreas) consideram a análise das expressões faciais e tonalidade de voz fatores essenciais para as tomadas de decisões durante uma possível contratação ou para o andamento das sessões/atendimentos psicológicos.

A fim de compreender as dificuldades encontradas durante a interpretação emocional de um paciente ou candidato para uma vaga de emprego, os respectivos profissionais das áreas responderam a uma quinta pergunta qualitativa. Após análise das respostas coletadas, pôde-se constatar que alguns termos foram recorrentes, como nervosismo, ansiedade, neutralidade, expressões e emoções.

Em uma das respostas fornecidas por um profissional da área de Recursos Humanos (RH), “muitas vezes o candidato está nervoso e ansioso devido a entrevista e para não causar/impactar negativamente, ele tenta se manter neutro e contido em suas reações, expressões faciais e corporais”, já outro recrutador

pontuou que “assertividade na identificação e nervosismo do candidato” são alguns fatores que dificultam seu processo de análise e escolha durante o processo seletivo.

Considerando as expressões que o candidato pode esboçar durante a entrevista, um profissional indicou que “a ansiedade do candidato em alguns momentos da entrevista pode interferir em suas ações e ou expressões (de forma indesejada)”, logo, os indivíduos tendem a se conter em suas reações na maioria dos casos. De maneira semelhante, outro candidato ressaltou “que o candidato pode estar extremamente nervoso e pode não refletir necessariamente ao seu verdadeiro estado emocional ou intenção. As expressões têm múltiplos, até mesmo culturais”.

Apesar da tonalidade de voz não ter sido destaque nas respostas obtidas, um respondente apontou que “o candidato pode naturalmente falar um pouco mais baixo ou estar nervoso na entrevista (influenciando na tonalidade apenas naquele momento)”, sendo necessário considerar estas situações para dar continuidade na entrevista e assim, tomar uma decisão final no processo seletivo. Logo, desenvolver um sistema que consiga distinguir e detectar a neutralidade das expressões faciais dos candidatos é de extrema importância, assim como, garantir a qualidade na transmissão do áudio, visto o caso da tonalidade de voz citado por um dos respondentes.

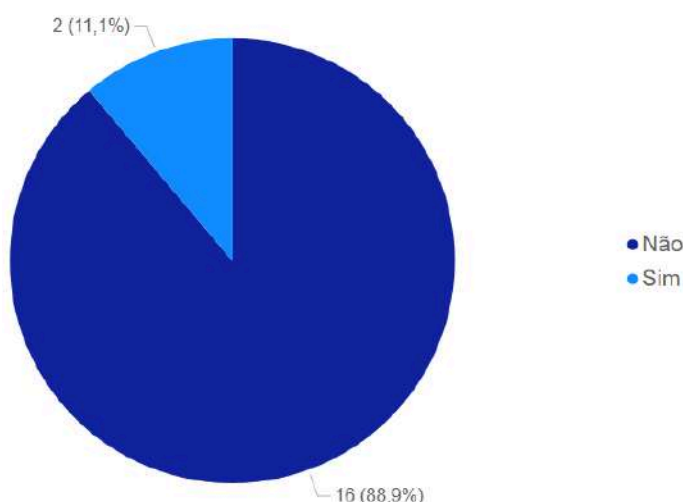
Analizando as respostas e dificuldades enfrentadas pelos psicólogos durante um atendimento, um dos profissionais pontuou que

Interpretar a comunicação não verbal do paciente é essencial para conduzir o atendimento conforme suas emoções e assim, se criar uma conexão. Uma das maiores dificuldades está no fato de que a manifestação das emoções é diferente em cada pessoa.

Além disso, outro psicólogo respondeu que “construir um vínculo com o paciente que apresenta maior resistência em demonstrar sentimentos e emoções durante as consultas” é uma grande dificuldade na análise emocional no decorrer das sessões, assim como a possibilidade de ambiguidade dos sentimentos ressaltados por um psicólogo: “dependendo do paciente, algumas emoções podem ser ambíguas ou neutras, o que dificulta na interpretação. Além disso, é necessário considerar o contexto em muitas situações”.

Na Figura 53 é possível visualizar os respondentes que já utilizaram alguma ferramenta/tecnologia para o monitoramento emocional e, no caso, apenas dois profissionais (um da área de recursos humanos e um psicólogo) relataram que já tiveram alguma experiência. Portanto, aproximadamente 89% da amostra nunca pôde explorar e aplicar uma ferramenta como esta, comprovando que a criação de uma interface que esteja dentro do conceito de acessibilidade digital, pode impactar de maneira positiva, uma vez que tais profissionais recebam a oportunidade de utilizá-los e testá-los em suas análises no dia a dia.

Figura 53 - Uso de alguma ferramenta/tecnologia para monitoramento emocional



Fonte: Elaborado pela autora.

Com o resultado obtido na análise anterior e visando a idealização/desenvolvimento de uma ferramenta que atenda a necessidade de todos, buscou-se identificar e compreender os possíveis benefícios que um sistema de análise das expressões faciais e da tonalidade de voz pode gerar para os doze respondentes, que atuam na área de Recursos Humanos.

No caso, na Tabela 13 é possível visualizar que onze participantes (aproximadamente, 92%) afirmaram “Identificação de emoções não verbalizadas pelo candidato”, sete participantes (64%) afirmaram “Otimização de *feedbacks*”, cinco optaram pela “Redução de interpretações equivocadas na entrevista” e apenas um para “Auxiliar na tomada de decisões durante a entrevista/atendimento”.

Tabela 13 - Benefícios de um sistema de análise das expressões faciais e da tonalidade de voz na área de Recursos Humanos (RH)

Benefícios	Qtde. de Respostas
Identificação de emoções não verbalizadas pelo candidato	11
Otimização de feedbacks	7
Redução de interpretações equivocadas na entrevista	5
Auxiliar na tomada de decisões durante a entrevista	1

Fonte: Elaborado pela autora.

A mesma análise foi realizada nas respostas coletadas com os psicólogos envolvidos na pesquisa. Na Tabela 14 observa-se que para todos os respondentes, isto é, para os seis psicólogos a “Identificação de emoções não verbalizadas pelo candidato” é um dos grandes benefícios que um sistema de análise/interpretação emocional pode proporcionar. Diferentemente dos resultados obtidos com os recrutadores, cinco participantes (aproximadamente, 83%) afirmaram “Personalização de tratamentos ao longo das sessões (durante atendimento psicológico)” e apenas um selecionou “Auxiliar na tomada de decisões durante o atendimento”.

Tabela 14 - Benefícios de um sistema de análise das expressões faciais e da tonalidade de voz na área de psicologia

Benefícios	Qtde. de Respostas
Identificação de emoções não verbalizadas pelo candidato	6
Personalização de tratamentos ao longo das sessões	5
Redução de interpretações equivocadas no atendimento	1

Fonte: Elaborado pela autora.

Outro aspecto importante que foi questionado aos profissionais/participantes da pesquisa, foi em relação às suas preocupações ao utilizar um sistema de análise/interpretação emocional no ambiente de trabalho. Novamente, as respostas obtidas são distintas quando analisadas separadamente (por área de atuação).

Como é possível observar na Tabela 15, entre os recrutadores respondentes, onze demonstraram preocupação com a “Acurácia na Detecção” do sistema, isto é,

com o percentual de acertos da nova ferramenta e, conseqüentemente, a confiabilidade dos resultados. A outra preocupação foi o “Custo”, entre os doze participantes, nove priorizam este requisito. Na sequência, quatro se atentam para o “Tempo de Processamento rápido/lento” e três participantes pensam na “Dificuldade no manuseio” e “Privacidade dos Dados” dos respectivos candidatos.

Entre os psicólogos respondentes, todos demonstraram preocupação com a “Acurácia na Detecção” do sistema, o que ressalta a importância de um sistema que garanta a confiabilidade dos resultados gerados. A outra preocupação foi a “Privacidade dos Dados”, onde novamente, todos priorizam este requisito. Quanto às demais preocupações, três profissionais se atentam para o “Tempo de Processamento rápido/lento”, dois participantes pensam na “Dificuldade no manuseio” e apenas um psicólogo se preocupa com o “Custo” do sistema.

Tabela 15 - Preocupações ao utilizar um sistema de análise/interpretação emocional

Preocupações	Qtde. de Respostas Recursos Humanos	Qtde. de Respostas Psicologia
Acurácia na Detecção	11	6
Custo	9	1
Dificuldade no manuseio	3	2
Privacidade dos Dados	3	6
Tempo de processamento rápido/lento	4	3

Fonte: Elaborado pela autora.

De modo geral, para que se tenha uma maior aceitação do sistema de análise emocional, as preocupações dos respectivos usuários devem ser contempladas durante o desenvolvimento. Para uma implementação eficaz, é fundamental garantir aos profissionais uma ferramenta que forneça os resultados com êxito e confiabilidade, reforce a privacidade dos dados de todos os envolvidos (usuários, pacientes e candidatos), apresente uma interface intuitiva, de baixo custo e com tempo de processamento rápido das informações.

5.2 Levantamento de Opinião após a Realização de uma Sessão Experimental do Sistema

Uma sessão experimental foi realizada com um voluntário da área de Recursos Humanos na faixa etária entre 31 e 35 anos. Como descrito na Seção 3.4.3, foi solicitado que o mesmo conduzisse o diálogo de maneira descontraída e seguisse o roteiro de entrevista de emprego ao qual ele estivesse habituado em suas experiências prévias de recrutamento.

Antes do início da simulação, a plataforma foi apresentada e todas as funcionalidades explicadas ao profissional. Além disso, a pessoa entrevistada foi a pesquisadora/desenvolvedora do projeto. Uma vez iniciada a videochamada, o profissional pediu para que a pessoa entrevistada se apresentasse e falasse sobre a sua experiência profissional e pessoal, o início da comunicação foi mais leve e descontraído. Em seguida, perguntas como “Você conhece a nossa empresa?”, “Por que você gostaria de trabalhar conosco?”, “O que te motiva na sua vida profissional?” foram realizadas durante a conversa. Na Figura 54 é possível visualizar um registro dos participantes (recrutador e entrevistado) durante a realização da sessão experimental, já na sala de reunião virtual.

Figura 54 - Sessão Experimental com o Profissional



Fonte: Elaborado pela autora.

Finalizada a entrevista, o voluntário testou as funcionalidades disponíveis na plataforma e analisou os resultados obtidos na detecção. Em seguida, o questionário

disponível no APÊNDICE B foi aplicado com o auxílio do “Google Forms” para que o profissional avaliasse a interatividade, performance e o desempenho da interface.

Em uma escala de 1 a 5, em que 1 é “Discordo Totalmente” e 5 “Concordo Totalmente”, oito perguntas foram realizadas visando avaliar a performance e qualidade do sistema de detecção, da plataforma e das demais funcionalidades disponíveis. A avaliação foi extremamente positiva, como é possível observar na Tabela 16. Além disso, em uma das perguntas, o tempo de processamento do sistema durante a troca de informações foi avaliada como satisfatório.

Tabela 16 - Funcionamento do Sistema e Interface

Em uma escala de 1 a 5:	CONCORDO TOTALMENTE
Me senti confiante ao usar o sistema.	X
O sistema funcionou sem erros.	X
Todas as funcionalidades da ferramenta estavam ativas/disponíveis para o uso.	X
A performance do sistema atendeu às minhas expectativas.	X
A interface é de fácil aprendizado.	X
O áudio e o vídeo apresentaram excelente qualidade durante a interação entre os participantes na sessão experimental.	X

Fonte: Elaborado pela autora.

Para as perguntas dicotômicas (Sim/Não), o voluntário relata que esta foi a sua primeira experiência com uma tecnologia de monitoramento emocional e a avaliação foi mais uma vez positiva, como é possível observar na Tabela 17.

Tabela 17 - Experiência de Uso da Plataforma/Sistema Profissional

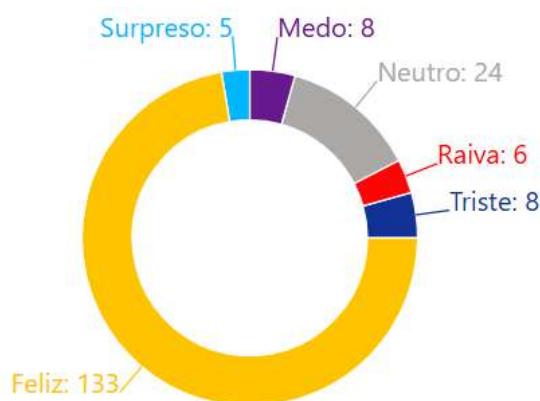
Perguntas	SIM	NÃO
Você está satisfeito(a) com a experiência de uso?	X	
Esta foi a sua primeira experiência com alguma ferramenta/tecnologia para monitoramento emocional?	X	
Recomendaria o sistema para outras pessoas?	X	
O sistema de análise das expressões faciais e dos sinais de voz foi uma ferramenta que te auxiliou em algum momento, no processo de identificação de emoções não verbalizadas pelo candidato?	X	
O sistema de análise das emoções te auxiliou em algum momento, no processo de tomada de decisões durante a entrevista de emprego?	X	

Fonte: Elaborado pela autora.

Por sua vez, para a pergunta “O que você achou do sistema de detecção emocional? Descreva a sua experiência e impressões (positivas ou negativas)”, o voluntário respondeu: “Gostei muito! É comum durante uma entrevista, o candidato transitar entre vários momentos, demonstrando sentimentos diferentes, como nervosismo, medo, segurança ou insegurança e ter isso registrado e principalmente, transformado em dados, é valioso nos dias atuais e nos ajuda a avaliar o desempenho geral do candidato na entrevista. Em relação a melhorias, enxergo uma oportunidade de aperfeiçoamento na detecção de voz”.

Ao ser questionado sobre as suas impressões gerais sobre a interface Web utilizada a resposta foi: “Achei a experiência de navegação fluida e bastante intuitiva e os atalhos ajudam o profissional a localizar rapidamente as principais funcionalidades”. Do mesmo modo, na pergunta seguinte ele confirmou que a ferramenta de detecção de emoções apresentou resultados próximos ou semelhantes as suas próprias percepções/impressões durante a conversa: “Fiquei muito surpreso, pois de fato, a detecção de emoções condiz com as expressões do candidato, principalmente a detecção facial”, inclusive após a entrevista ele comentou o fato da entrevistada sorrir bastante durante a sessão experimental, isso foi comprovado após visualizar o gráfico de frequência, como apresentado na Figura 55.

Figura 55 - Frequência da Detecção Facial durante a Sessão Experimental



Fonte: Elaborado pela autora.

Por fim, o profissional confirmou que a ferramenta facilitou o processo de recrutamento: “Ajudou demais! Achei a ferramenta inovadora e estou certo de que tem muito a contribuir com nosso trabalho, como profissionais de RH, para que nossa tomada de decisão no processo de recrutamento e seleção seja cada vez mais assertivo”.

6 CONCLUSÕES

Este trabalho de graduação propôs inicialmente, a criação de uma interface Web de videoconferência que permitisse aos usuários a identificação das emoções por meio da detecção de expressões faciais e sinais de voz, durante uma entrevista de emprego ou um atendimento terapêutico. Além disso, buscava-se garantir o acesso a um sistema com serviço de qualidade, tempo de processamento satisfatório, navegação intuitiva e que de algum modo, auxiliasse o profissional em interpretações durante comunicações verbais e/ou não verbais e em suas possíveis tomadas de decisões nas reuniões.

Os dados obtidos com a aplicação do questionário após a realização de uma sessão experimental com um profissional da área de Recursos Humanos, foram analisados de forma descritiva e confirmaram a usabilidade da plataforma, a satisfação e confiança do profissional ao testar o sistema, a qualidade na entrega e transmissão da mídia, isto é, do áudio e vídeo, além do feedback positivo de que a ferramenta tem muito a contribuir e facilitar com o seu trabalho de recrutamento, visto que durante uma entrevista de emprego, é comum o candidato transitar entre vários momentos, demonstrando sentimentos diferentes. Resumidamente, as métricas obtidas com o uso da ferramenta *Google Lighthouse*, que busca avaliar a qualidade de páginas web (verificando o desempenho, acessibilidade, entre outras), foram também satisfatórias.

Como citado anteriormente, as plataformas de agendamento e videoconferência similares existentes não contemplam todas as funcionalidades propostas neste trabalho de graduação. Dessa forma, a implementação do sistema contribui de maneira significativa e torna-se uma ferramenta auxiliar para as respectivas áreas.

Em relação aos resultados obtidos durante a simulação, a reação do usuário foi de surpresa, pois de fato, a detecção de emoções condizia em grande parte com as expressões do respectivo candidato, principalmente as detecções faciais de emoções como neutralidade, felicidade, surpresa e tristeza. Basicamente, modelos pré-treinados e arquiteturas de redes neurais convolucionais personalizadas foram testadas para avaliar e identificar a abordagem/configuração que apresentasse

melhor performance. Do mesmo modo, algumas técnicas de pré-processamento foram utilizadas durante as tentativas.

Para o modelo de detecção facial selecionado, os resultados indicaram uma acurácia adequada para o propósito do estudo de, aproximadamente, 70%, observando-se ainda, que em algumas classes específicas o desempenho de classificação foi significativamente superior, ultrapassando 80% (ênfase para detecção de felicidade, surpresa e neutralidade) e mostrando-se de acordo com as percepções relatadas pelo voluntário no teste do sistema.

Por sua vez, para a modelagem de sinais de voz a extração de recursos/*features* é uma etapa fundamental, logo, foi importante analisar também os espectrogramas de MFCCs (Coeficiente Cepstral de Frequência Mel) dos dados. Com a estrutura de rede neural construída e treinada, mais uma vez o modelo apresentou uma performance de classificação adequada, de aproximadamente 70%, indicando um desempenho superior de acerto para algumas classes como, por exemplo, calma e raiva. Entretanto, para esta situação foi relatada certa inconsistência durante o teste realizado pelo profissional. Segundo o recrutador, houve a oscilação durante a detecção dos sinais de voz da emoção/tonalidade calma (retornando como resultado tristeza) e, portanto, essa foi uma das sugestões de aperfeiçoamento na detecção de voz entregues pelo voluntário.

Com esse feedback comentado no parágrafo anterior e as métricas gerais alcançadas até o momento, identifica-se a necessidade de aprimoramento do sistema de detecção para os dois modelos de classificação em trabalhos futuros. Basicamente, analisar novas arquiteturas, ampliar o conjunto de dados e a otimização de hiperparâmetros são algumas alternativas que podem ser consideradas na busca por um melhor desempenho da plataforma.

REFERÊNCIAS

BRASIL. Ministério da Saúde. *OMS classifica coronavírus como pandemia*. GOV.BR. 11 mar. 2020. Disponível em: <<https://www.gov.br/pt-br/noticias/saude-e-vigilancia-sanitaria/2020/03/oms-classifica-coronavirus-como-pandemia>>. Acesso em: 06 fev. 2025.

CLARO, INSTITUTO. Ondas sonoras: conceitos fundamentais. Portal de Educação – Instituto Claro, 17 out. 2017. Disponível em: <<https://www.institutoclaro.org.br/educacao/para-ensinar/planos-de-aula/ondas-sonoras/>>. Acesso em: 12 jun. 2025.

DIEMERLING, H.; STRESEMANN, L.; BRAUN, T.; OERTZEN, T. *Implementing machine learning techniques for continuous emotion prediction from uniformly segmented voice recordings*. *Frontiers in Psychology*, v. 15, 2024. DOI: 10.3389/fpsyg.2024.1300996.

EKMAN, P. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York: Times Books, 2003.

EXAME. Software promete detectar mentiras em entrevista de emprego. *Exame*, 26 abr. 2013. Disponível em: <<https://exame.com/carreira/software-promete-detectar-mentiras-em-entrevista-de-emprego/>>. Acesso em: 18 fev. 2025.

FURTADO, M. I. V. *Redes Neurais Artificiais: Uma Abordagem Para Sala de Aula*. Atena Editora, p. 9-27, 2019. DOI: 10.22533/at.ed.262191504.

GLOBO, O. *Terapia online cresce e amplia alcance de tratamentos psicológicos*. O Globo Saúde, 24 out. 2021. Disponível em: <<https://oglobo.globo.com/saude/terapia-online-cresce-amplia-alcance-de-tratamento-psicologicos-25247975>>. Acesso em: 06 fev. 2025.

GOODFELLOW, I, J. et al. *Challenges in Representation Learning: A report on three machine learning contests*. 2013. FER2013 (Facial Expression Recognition 2013 Dataset). DOI: 10.48550/arXiv.1307.0414.

GOOGLE. Colaboratory. *Perguntas frequentes*. Disponível em: <<https://research.google.com/colaboratory/intl/pt-BR/faq.html>>. Acesso em: 10 mai. 2025.

HAYKIN, S. *Neural Networks and Learning Machines*. 3. ed. Upper Saddle River: Pearson Prentice Hall, 2009.

IBM. *O que é uma rede neural?*. Disponível em: <<https://www.ibm.com/br-pt/think/topics/neural-networks>>. Acesso em: 25 mai. 2025a.

IBM. *Guia do IBM SPSS Modeler CRISP-DM*. Disponível em: <https://www.ibm.com/docs/pt-br/SS3RA7_18.5.0/nl/pt/BR/pdf/ModelerCRISPDM.pdf>. Acesso em: 17 abr. 2025b.

LIVINGSTONE, S. R.; RUSSO, F. A. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*. PLoS ONE 13(5): e0196391, 2018. DOI: 10.1371/journal.pone.0196391.

MARCIANO, C. E.; SOUZA, F. A.; SOUZA, R. B. *Arquitetura Par-a-Par (Peer to Peer Networks)*. GTA UFRJ. Disponível em: <<https://www.gta.ufrj.br/ensino/eel878/redes1-2018-1/trabalhos-v1/p2p/arquitetura.html>>. Acesso em: 15 mai. 2025.

MEDICINA S/A. *87% das pessoas que aderiram ao tratamento psicológico online perceberam melhorias*. Medicina SA, 11 mar. 2025. Disponível em: <<https://medicinasa.com.br/terapia-online/>>. Acesso em: 02 abr. 2025.

MEHRABIAN, A. *Silent Messages: Implicit Communication of Emotions and Attitudes*. Belmont: Wadsworth Publishing Company, 1971.

MOLLAHOSSEINI, A.; HASANI, B.; MAHOOR, M. H. *AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild*. IEEE Transactions on Affective Computing, v. 10, no. 1, pp. 18–31, 21 ago. 2019. DOI: 10.1109%2Ftaffc.2017.2740923.

MOZILLA. WebRTC API. Disponível em: <https://developer.mozilla.org/pt-BR/docs/Web/API/WebRTC_API>. Acesso em: 10 mar. 2025a.

MOZILLA. RTCPeerConnection. Disponível em: <<https://developer.mozilla.org/en-US/docs/Web/API/RTCPeerConnection>>. Acesso em: 10 mar. 2025b.

MÜHLER, V. *Face-api.js — JavaScript API for Face Recognition in the Browser with tensorflow.js*. Medium, 25 jun. 2018. Disponível em: <<https://itnext.io/face-api-js-javascript-api-for-face-recognition-in-the-browser-with-tensorflow-js-bcc2a6c4cf07>>. Acesso em: 07 jul. 2025.

MÜLLER, M. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. 1. ed. Basileia, Switzerland: Springer International Publishing, 2015.

PARIS, L. *Restrições da pandemia mudam entrevistas de emprego para o formato online, e candidatos precisam se adaptar: veja dicas*. G1 PR – Curitiba, 01 mai. 2021. Disponível em: <<https://g1.globo.com/pr/parana/concursos-e-emprego/noticia/2021/05/01/restricoes-da-pandemia-mudam-entrevistas-de-emprego-para-o-formato-online-e-candidatos-precisam-se-adaptar-veja-dicas.ghtml>> Acesso em: 06 fev. 2025.

PHAM, L.; VU, T. H.; TRAN, T. A. *Facial Expression Recognition Using Residual Masking Network*. 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4513–4519. DOI: 10.1109/ICPR48806.2021.9411919.

PRAMERDORFER, C.; KAMPEL, M. *Facial Expression Recognition using Convolutional Neural Networks: State of the Art*. Computer Vision Lab, TU Wien. 9 dez. 2016. DOI: 10.48550/arXiv.1612.02903.

SCIKIT-LEARN. `confusion_matrix`. 2025. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html> Acesso em: 01 jul. 2025.

SERENGIL, S. I.; OZPINAR, A. *HyperExtended LightFace: A Facial Attribute Analysis Framework*. International Conference on Engineering and Emerging Technologies (ICEET), Istanbul, Turkey, 2021, pp. 1-4, doi: 10.1109/ICEET53442.2021.9659697.

WIRTH, R.; HIPPEL, J. *CRISP-DM: Towards a Standard Process Model for Data Mining*. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, 11-13 Abr. 2000.

APÊNDICE A

Questionário - Análise Emocional: Abordagem de diferentes metodologias de processamento dos dados e técnicas de Deep Learning

Aplicação de diferentes técnicas e abordagens para identificar emoções através da combinação e análise de informações faciais e de voz de um indivíduo em situações variadas como, por exemplo, Atendimento Psicológico Virtual ou Entrevistas de Emprego Online. Com isso, um sistema será desenvolvido para auxiliar profissionais na interpretação emocional de um paciente ou candidato, durante um atendimento ou entrevista online.

Todas as suas respostas não serão associadas ao seu nome ou a qualquer outra informação pessoal. A análise dos resultados será feita de forma anônima.

1. Área de Atuação:

- ☐ Recursos Humanos
- ☐ Psicologia

2. Qual sua faixa etária?

- ☐ Entre 18 e 20 anos
- ☐ Entre 21 e 25 anos
- ☐ Entre 26 e 30 anos
- ☐ Entre 31 e 35 anos
- ☐ Entre 36 e 40 anos
- ☐ Entre 41 e 45 anos
- ☐ Entre 46 e 50 anos
- ☐ Acima de 50 anos

3. Em uma escala de 1 a 5, quão importante você considera a **análise das expressões faciais** durante um atendimento psicológico ou entrevista de emprego?

	1	2	3	4	5
Pouco Importante()	()	()	()	()	Muito Importante

4. Em uma escala de 1 a 5, quão importante você considera a **tonalidade de voz** na análise emocional em um atendimento psicológico ou entrevista de emprego?

	1	2	3	4	5
Pouco Importante()	()	()	()	()	Muito Importante

5. Quais dificuldades você encontra durante a interpretação das expressões faciais e da voz (interpretação emocional) de um paciente/candidato?

6. Você já utilizou alguma ferramenta/tecnologia para monitoramento emocional?

() Sim
() Não

7. Em sua opinião, quais benefícios um sistema de análise das expressões faciais e do tom da voz pode gerar durante um atendimento ou entrevista de emprego? (*Escolha até 2 opções*)

() Identificação de emoções não verbalizadas pelo candidato/paciente
() Redução de interpretações equivocadas na entrevista/atendimento
() Auxiliar na tomada de decisões durante a entrevista/atendimento
() Otimização de feedbacks
() Personalização de tratamentos ao longo das sessões (psicólogos)
() Outro: _____

8. Qual seria sua preocupação ao utilizar um sistema de análise/interpretação emocional?

() Acurácia na detecção

() Custo

() Privacidade dos dados

() Dificuldade no manuseio

() Tempo de processamento rápido/lento

() Outro: _____

APÊNDICE B

Questionário Sessão Experimental do Sistema de Análise Emocional

Aplicação de diferentes técnicas e abordagens para identificar emoções através da combinação e análise de informações faciais e de voz de um indivíduo em situações variadas como, por exemplo, Atendimento Psicológico Virtual ou Entrevistas de Emprego Online. O questionário será aplicado após a realização da sessão experimental com os voluntários.

Todas as respostas obtidas não serão associadas ao nome ou a qualquer outra informação pessoal do participante. A análise dos resultados será feita de forma anônima.

1. Área de Atuação:

- ☐ Recursos Humanos
- ☐ Psicologia

2. Qual sua faixa etária?

- ☐ Entre 18 e 20 anos
- ☐ Entre 21 e 25 anos
- ☐ Entre 26 e 30 anos
- ☐ Entre 31 e 35 anos
- ☐ Entre 36 e 40 anos
- ☐ Entre 41 e 45 anos
- ☐ Entre 46 e 50 anos
- ☐ Acima de 50 anos

3. Em uma escala de 1 a 5: O sistema é de fácil manuseio/uso.

	1	2	3	4	5	
Discordo Totalmente	()	()	()	()	()	Concordo Totalmente

4. Em uma escala de 1 a 5: Me senti confiante ao usar o sistema.

	1	2	3	4	5	
Discordo Totalmente	()	()	()	()	()	Concordo Totalmente

5. Em uma escala de 1 a 5: O sistema funcionou sem erros.

	1	2	3	4	5	
Discordo Totalmente	()	()	()	()	()	Concordo Totalmente

6. Em uma escala de 1 a 5: Todas as funcionalidades da ferramenta estavam ativas/disponíveis para o uso.

	1	2	3	4	5	
Discordo Totalmente	()	()	()	()	()	Concordo Totalmente

7. Em uma escala de 1 a 5: A performance do sistema atendeu às minhas expectativas.

Discordo Totalmente 1 2 3 4 5 Concordo Totalmente

8. Em uma escala de 1 a 5: A interface é de fácil aprendizado.

Discordo Totalmente 1 2 3 4 5 Concordo Totalmente
 () () () () ()

9. Em uma escala de 1 a 5: Avalie o tempo de processamento do sistema durante a troca de informações.

	1	2	3	4	5	
Lento	()	()	()	()	()	Rápido

10. Em uma escala de 1 a 5: O áudio e o vídeo apresentaram excelente qualidade durante a interação entre os participantes na sessão experimental.

	1	2	3	4	5	
Discordo Totalmente	()	()	()	()	()	Concordo Totalmente

11. Você está satisfeito(a) com a experiência de uso?

() Sim
() Não

12. Esta foi a sua primeira experiência com alguma ferramenta/tecnologia para monitoramento emocional?

() Sim
() Não

13. Recomendaria o sistema para outras pessoas?

() Sim
() Não

14. O sistema de análise das expressões faciais e dos sinais de voz foi uma ferramenta que te auxiliou em algum momento, no processo de identificação de emoções não verbalizadas pelo candidato/paciente?

() Sim
() Não

15.O sistema de análise das expressões faciais e dos sinais de voz foi uma ferramenta que te auxiliou em algum momento, no processo de tomada de decisões durante o atendimento/entrevista de emprego?

() Sim

() Não

16.O sistema de análise das expressões faciais e dos sinais de voz foi uma ferramenta que te auxiliou em algum momento, na personalização de tratamentos ao longo das sessões?

() Sim

() Não

17.O que você achou do sistema de detecção emocional? Descreva a sua experiência e impressões (positivas ou negativas)

18.Quais foram as suas impressões gerais sobre a interface Web utilizada? Descreva a sua experiência e impressões (positivas ou negativas).

19.Você considera que a ferramenta facilitou ou dificultou o processo de recrutamento/atendimento?

20.A ferramenta de detecção de emoções apresentou resultados próximos ou semelhantes as suas próprias percepções/impressões durante a conversa?

APÊNDICE C

No decorrer do trabalho de graduação, foram utilizados rodapés com a finalidade de esclarecimento sobre termos técnicos, tecnologias e ferramentas complementares que não condizem com a lista principal de referências do projeto.

Buscando a facilitação do leitor durante a consulta, o APÊNDICE C visa reunir os rodapés utilizados.

Rodapé 1 – p. 3 – Capítulo 1

- [1](#): Link complementar de apresentação da plataforma online Zenklub.
- [2](#): Link complementar de apresentação da plataforma online Psitto.
- [3](#): Link complementar de acesso ao aplicativo de comunicação *WhatsApp*.
- [4](#): Link complementar de apresentação da plataforma online PSI Psicólogo e Terapia.

Rodapé 2 – p. 4 – Capítulo 1

- [5](#): Link de acesso para a plataforma de videoconferência *Google Meet*.
- [6](#): Link de acesso para a plataforma de videoconferência *Zoom*.
- [7](#): Link de acesso para a plataforma de videoconferência *ConferênciaWeb*.
- [8](#): Link de acesso para a plataforma de videoconferência *Microsoft Teams*.

Rodapé 3 – p. 18 – Capítulo 3

- [9](#): Contextualização sobre aplicação da função *RTCPeerConnection*.
- [10](#): Breve contextualização sobre o conceito de codecs e utilização multimídia.
- [11](#): Link complementar para o material e documentação *socket.io*.

Rodapé 4 – p. 20 – Capítulo 3

- [12](#): Link de acesso para a plataforma de hospedagem de código fonte *GitHub*.

Rodapé 5 – p. 21 – Capítulo 3

- [13](#): Link complementar de acesso a ferramenta Microsoft Power BI.

Rodapé 6 – p. 23 – Capítulo 3

[14](#): Link complementar de acesso ao editor de código fonte *Visual Studio Code*.

Rodapé 7 – p. 25 – Capítulo 3

[15](#): Link complementar de acesso a documentação da biblioteca *React*.

[16](#): Link complementar de acesso a documentação da biblioteca *JavaScript Redux*.

[17](#): Link complementar para o material e documentação *socket.io*.

[18](#): Link complementar de acesso ao editor de código fonte *Visual Studio Code*.

Rodapé 8 – p. 29 – Capítulo 4

[19](#): Link de acesso ao editor gráfico de prototipagem de projetos de design *Figma*.

Rodapé 9 – p. 35 – Capítulo 4

[20](#): Link de acesso ao ambiente de *Jupyter Notebook* hospedado na nuvem: *Google Colaboratory*.

[21](#): Definição da abreviação GPUs.

[22](#): Definição da abreviação TPUs.

[23](#): Link complementar de acesso a plataforma de compartilhamento de arquivos *Google Drive*.

Rodapé 10 – p. 43 – Capítulo 4

[24](#): Link de acesso a API do *TensorFlow* de criação e treinamento de modelos *Deep Learning*.

Rodapé 11 – p. 57 – Capítulo 4

[25](#): Link de acesso ao dicionário Cambridge.

Rodapé 12 – p. 60 – Capítulo 4

[26](#): Link de acesso a documentação da função *React-Calendar*.

Rodapé 13 – p. 61 – Capítulo 4

[27](#): Link de acesso a documentação da função *Papa Parse*.

Rodapé 14 – p. 67 – Capítulo 4

[28](#): Link de acesso a documentação *Keras*.

[29](#): Link de acesso a ferramenta *TensorFlow*.

[30](#): Link de acesso a documentação da biblioteca *HDF5*.

Rodapé 15 – p. 68 – Capítulo 4

[31](#): Link de acesso a documentação *ONNX*.

Rodapé 16 – p. 70 – Capítulo 4

[32](#): Link de acesso a documentação da função *Recharts*.

Rodapé 17 – p. 71 – Capítulo 4

[33](#): Link de acesso a ferramenta *Google Lighthouse*, que analisa a qualidade das páginas da Web.