UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

WILLIAN GUERREIRO COLARES

CLASSIFICAÇÃO DE EMOÇÕES HUMANAS UTILIZANDO PONTOS DE REFERÊNCIA DA FACE E REDES NEURAIS PROFUNDAS

MANAUS
2024

UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

WILLIAN GUERREIRO COLARES

CLASSIFICAÇÃO DE EMOÇÕES HUMANAS UTILIZANDO PONTOS DE REFERÊNCIA DA FACE E REDES NEURAIS PROFUNDAS

Dissertação apresentada ao Curso de Mestrado em Engenharia Elétrica, área de concentração Controle e Automação de Sistemas e linha de pesquisa Reconhecimento de Padrões e Otimização do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Amazonas.

Orientador: Prof. Dr. Cícero Ferreira Fernandes Costa Filho
Coorientadora: Profª. Drª. Marly Guimarães Fernandes Costa

MANAUS
2024

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Colares, Willian Guerreiro
C683c Classificação de emoções humanas utilizando pontos de referência da face e redes neurais profundas / Willian Guerreiro Colares . 2024
76 f.: il. color; 31 cm.

Orientador: Cícero Ferreira Fernandes Costa Filho
Coorientadora: Marly Guimarães Fernandes Costa
Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Amazonas.

1. Expressões faciais. 2. Fusão. 3. Rede neural convolucional. 4. Affectnet. I. Costa Filho, Cícero Ferreira Fernandes. II. Universidade Federal do Amazonas III. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Engenharia Elétrica

FOLHA DE APROVAÇÃO

Poder Executivo Ministério da Educação
Universidade Federal do Amazonas
Faculdade de Tecnologia
Programa de Pós-graduação em Engenharia Elétrica

Pós-Graduação em Engenharia Elétrica. Av. General Rodrigo Octávio Jordão Ramos, nº 3.000 - Campus Universitário, Setor Norte - Coroado, Pavilhão do CETELI. Fone/Fax (92) 99271-8954 Ramal:2607. E-mail: ppgee@ufam.edu.br

WILLIAN GUERREIRO COLARES

CLASSIFICAÇÃO DE EMOÇÕES HUMANAS UTILIZANDO PONTOS DE REFERÊNCIA DA FACE E REDES NEURAIS PROFUNDAS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Amazonas, como
requisito parcial para obtenção do título de Mestre
em Engenharia Elétrica na área de concentração
Controle e Automação de Sistemas.

Aprovada em 11 de junho de 2024.

BANCA EXAMINADORA
Prof. Dr. Cícero Ferreira Fernandes Costa Filho- Presidente
Prof. Dr. José Raimundo Gomes Pereira - Membro Titular 2 - Externo
Prof. Dr. Frederico da Silva Pinagé - Membro Titular 1 - Externo

Manaus, 28 de maio de 2024.

Documento assinado eletronicamente por **Frederico da Silva Pinagé**, **Usuário Externo**, em 18/06/2024, às 15:51, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Cícero Ferreira Fernandes Costa Filho**, **Professor do Magistério Superior**, em 18/06/2024, às 15:54, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **José Raimundo Gomes Pereira**, **Professor do Magistério Superior**, em 20/06/2024, às 16:17, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2071340** e o código CRC **4A511F10**.

Av. Octávio Hamilton Botelho Mourão - Bairro Coroado 1 Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181
CEP 69080-900 Manaus/AM - mestrado_engeletrica@ufam.edu.br

Referência: Processo nº 23105.021268/2024-98 SEI nº 2071340

Criado por 31183646291, versão 4 por 31183646291 em 28/05/2024 14:10:31.

AGRADECIMENTOS

À Deus, por ter me concedido saúde e perseverança mediante os obstáculos enfrentados.

À minha família por ter sempre me apoiado em minhas decisões e sobretudo por ter me conduzido no caminho até aqui.

À minha noiva, pela paciência, apoio e compreensão ao longo desta caminhada.

Aos meus orientadores por terem se disponibilizado a partilhar seu conhecimento em prol do desenvolvimento desta pesquisa. Agradeço também por toda a paciência durante o processo de ideação, concepção e escrita da pesquisa.

Ao SIDI, que na figura de um Instituto de Pesquisa, proporcionou um ecossistema propício para o desenvolvimento desta pesquisa, seja por meio das atividades relacionadas à Visão Computacional, seja por meio da interação com os demais colegas da área.

Ao Centro de Pesquisa e Desenvolvimento de Tecnologia Eletrônica e da Informação - CETELI da Universidade Federal do Amazonas - UFAM por propiciar a infraestrutura necessária à realização desta dissertação.

RESUMO

As expressões faciais humanas desempenham um papel fundamental na comunicação não-verbal e na transmissão de emoções. Conceitualmente, as expressões faciais podem ser deduzidas a partir da disposição dos músculos faciais. Sendo uma avaliação subjetiva, a construção de uma base de dados para o reconhecimento de expressões faciais torna-se um desafio devido ao elevado risco de enviesamento decorrente de dados desequilibrados ou imprecisos. Por outro lado, os avanços nas técnicas de processamento de imagem e de aprendizagem profunda têm aumentado a precisão e a eficácia dos algoritmos de reconhecimento de expressões faciais. Neste trabalho, com o objetivo de melhorar o reconhecimento automático de expressões faciais, apresentamos a fusão de duas arquiteturas de redes neurais. A primeira compreende uma rede neural convolucional unidimensional (1D), com entrada caracterizada por pontos de referência da face, e uma segunda, uma rede neural convolucional baseada no backbone DenseNet, com a própria imagem do rosto como entrada. O otimizador ADAM foi utilizado durante o treino desta rede. Foi utilizada a base de dados AffectNet. O melhor resultado obtido foi uma precisão de 60,40% no subconjunto de teste, para a modalidade de 7 classes. Este resultado é comparável aos melhores resultados obtidos no conjunto de dados AffectNet.

Palavras-chave: expressões faciais, fusão, rede neural convolucional, affectnet

ABSTRACT

Human facial expressions play a fundamental role in nonverbal communication and the conveyance of emotions. Conceptually, facial expressions can be deduced from the arrangement of facial muscles. As a subjective assessment, constructing a database for facial expression recognition becomes a challenge due to the high risk of bias arising from unbalanced or inaccurate data. On the other hand, advances in image processing techniques and deep learning have boosted the accuracy and effectiveness of algorithms for facial expression recognition. In this work, aiming to improve the automatic facial expression recognition, we present the fusion of two neural network architectures. The first one comprises a one-dimensional convolutional neural network (1D), with input characterized by facial landmarks, and a second one, a convolutional neural network based on the DenseNet backbone, with the face image itself as the input. The ADAM optimizer was used during the training of this network. The AffectNet database was employed. The best result obtained was an accuracy of 60.40% in the test subset, for the 7 classes modality. This result is comparable to the best results obtained on the AffectNet dataset.

Keywords: facial expressions, fusion, convolutional neural network

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

AM	Aprendizado de Máquina
AP	Aprendizado Profundo
CNN	*Convolutional Neural Network*
CPU	*Central Processing Unit*
GPU	*Graphics Processing Unit*
IA	Inteligência Artificial
RAM	*Random Access Memory*
RGB	*Red-Blue-Green*

SUMÁRIO

1 INTRODUÇÃO

As expressões faciais humanas desempenham um papel fundamental na comunicação não verbal e na transmissão de emoções. Como afirmado por Ekman e Friesen (1971), as expressões faciais são um meio eficaz de comunicação entre os indivíduos. Através de sorrisos, franzimentos de sobrancelhas, olhares surpresos ou lágrimas de tristeza, é possível expressar uma ampla gama de emoções, como alegria, raiva, medo, entre outras. As expressões faciais constituem uma linguagem universal que facilita a conexão entre as pessoas, permitindo a compreensão e o compartilhamento de experiências emocionais de forma profunda e significativa.

O reconhecimento de expressões faciais humanas tem aplicações diversas em áreas como psicologia, medicina, segurança e interação homem-máquina (LI et al., 2017). Na psicologia, é utilizado para entender e analisar emoções, auxiliando na avaliação de estados mentais e no diagnóstico de distúrbios psicológicos. Na medicina, contribui no monitoramento de pacientes e na identificação de indicadores de dor, fadiga e outras condições de saúde. Na segurança, é aplicado em sistemas de vigilância e identificação de suspeitos. Além disso, em interfaces homem-máquina, permite aprimorar a interação e a comunicação, tornando possível que computadores e dispositivos compreendam as expressões faciais dos usuários. Essas aplicações destacam o potencial do reconhecimento de expressões faciais como uma ferramenta valiosa em várias áreas, impulsionando avanços em diagnóstico, segurança e interação tecnológica.

A classificação de expressões faciais humanas por meio de algoritmos de aprendizado de máquina continua sendo uma área de pesquisa em rápido desenvolvimento. O avanço das técnicas de processamento de imagem e aprendizado profundo tem impulsionado a precisão e a eficácia desses algoritmos. Estudos recentes, como o trabalho de Liu et al. (2021), propuseram novas abordagens baseadas em redes neurais convolucionais e redes neurais de longa memória para melhorar a detecção e classificação de expressões faciais. Além disso, a disponibilidade de grandes conjuntos de dados rotulados, como o *Facial Expression Recognition and Analysis* (FERA) e o AffectNet, tem permitido treinar modelos mais robustos e capazes de reconhecer uma ampla gama de expressões emocionais (DHALL et al., 2019; MOLLAHOSSEINI et al., 2019).

Na área de reconhecimento de expressões faciais, diversos estudos e trabalhos têm contribuído para o avanço da pesquisa e das aplicações. Pesquisadores têm explorado diferentes técnicas e abordagens para melhorar a precisão e a robustez dos modelos de

reconhecimento. Por exemplo, o trabalho de Mollahosseini et al. (2019) propôs o uso de uma rede neural profunda para a extração de características faciais e alcançou resultados promissores em termos de classificação de expressões emocionais.

Além disso, abordagens baseadas em técnicas de aprendizado profundo, como redes neurais convolucionais, têm sido amplamente investigadas, como demonstrado pelo estudo de Zhang et al. (2018), que propôs uma rede neural convolucional com uma arquitetura especializada para o reconhecimento de expressões faciais em imagens 3D.

Outro trabalho relevante é o de Dhall et al. (2015), que introduziu um conjunto de dados abrangente chamado *Static Facial Expressions in the Wild* (SFEW), que contém imagens com expressões faciais em condições naturais. Esse conjunto de dados tem sido amplamente utilizado para avaliar e comparar o desempenho de diferentes algoritmos de reconhecimento de expressões.

De maneira complementar, estudos recentes, como o trabalho de Li et al. (2020), têm explorado a utilização de técnicas de transferência de aprendizado e redes neurais generativas adversariais para melhorar o reconhecimento de expressões faciais em cenários de baixa iluminação.

Treinar um modelo para reconhecimento de expressões faciais na base de dados AffectNet apresenta desafios significativos devido à natureza peculiar desse conjunto de dados. AffectNet é uma base de dados em larga escala com um número diversificado de imagens rotuladas com diferentes expressões emocionais. No entanto, a distribuição das classes nessa base de dados é desbalanceada, com algumas classes sendo sub-representadas em relação às classes majoritárias. Esse desequilíbrio pode afetar negativamente o desempenho do modelo e levar a um viés em direção às classes majoritárias. De acordo com Mollahosseini et al. (2019), o desbalanceamento das classes em AffectNet pode dificultar o treinamento de modelos de reconhecimento de expressões faciais.

O uso de pontos de referência da face é essencial no reconhecimento de expressões faciais (GHOSH et al., 2020). Esses pontos, também conhecidos como marcos faciais ou *landmarks*, são posições específicas na face, como os cantos dos olhos, o nariz e a boca, que podem ser detectados automaticamente em imagens ou vídeos. A detecção e o rastreamento desses pontos permitem a extração de informações relevantes para descrever a configuração e os movimentos faciais associados a diferentes expressões emocionais. Essas informações são utilizadas para treinar modelos de aprendizado de máquina capazes de classificar e reconhecer expressões faciais (VALSTAR et al., 2016). O uso de pontos de referência da face

tem se mostrado uma abordagem eficaz e promissora nessa área de pesquisa (GHOSH et al., 2020).

Essa dissertação tem o objetivo de contribuir com o avanço do estado da arte do tema classificação de emoções humanas utilizando tanto a imagem da face, quanto pontos de referência da face, como entradas para redes neurais profundas. Para isso, utiliza-se a base de dados AffectNet, referida anteriormente, e avalia-se o desempenho de uma arquitetura de rede composta de dados multimodais. De forma a reduzir o custo computacional de treinar um modelo com base em um conjunto de dados tão grande quanto o da base AffectNet, são realizados pré-processamentos a fim de se obterem os pontos de referência da face, constituindo assim uma variação mais leve da base, com dados unidimensionais. Para endereçar o problema de desbalanceamento dos dados, é adotada a estratégia do treinamento ponderado, isto é, são aplicados pesos a cada classe proporcionais à sua frequência de ocorrência na base de dados. A aplicação de pesos força que o modelo concentre igual atenção às classes majoritárias e minoritárias, durante o treinamento. Por fim, esse trabalho propõe uma nova arquitetura de rede, baseada na fusão de uma rede convolutiva 1D e de uma rede convolutiva 2D para a classificação de emoções humanas.

1.1 OBJETIVO GERAL

Propor uma arquitetura e um algoritmo de treinamento de redes multimodais para classificação de emoções humanas, que permitam alcançar desempenhos no estado da arte.

1.2 OBJETIVOS ESPECÍFICOS

1. Abordar o problema de classificação de emoções utilizando um banco de dados de imagens desbalanceado.
2. Avaliar o desempenho de diversas arquiteturas de redes profundas 1D, utilizando pontos de referência da face, para classificação de emoções humanas.
3. Avaliar o desempenho de diversas arquiteturas de redes profundas 2D, utilizando imagens 2D da face, para classificação de emoções humanas.
4. Propor e avaliar o desempenho de uma arquitetura de rede neural multimodal e de um novo algoritmo de treinamento, com base na fusão de redes convolucionais 1D e 2D, para a classificação de emoções humanas.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado conforme a divisão descrita a seguir:

- Capítulo 1: Introdução;
- Capítulo 2: Revisão da Literatura;
- Capítulo 3: Referencial Teórico;
- Capítulo 4: Materiais e Métodos;
- Capítulo 5: Resultados e discussão;

O Capítulo 1 contextualiza as aplicações de sistemas de reconhecimento de expressões faciais, bem como descreve as dificuldades associadas às características dos dados, tais como variação étnica, condições de iluminação, subjetividade no processo de anotação e desproporcionalidade entre as categorias de emoções.

O Capítulo 2 apresenta trabalhos publicados na literatura na área de classificação de expressões faciais. Neste capítulo, são apresentados quadros comparativos elencando as diversas estratégias de solução do problema, destacando a base de dados utilizada, o método empregado e os resultados obtidos.

No Capítulo 3, são apresentados os fundamentos teóricos necessários para o desenvolvimento deste trabalho. As expressões faciais são definidas de forma objetiva, bem como suas categorias. Os conceitos de aprendizagem profunda e redes neurais convolutivas unidimensionais são descritos. As definições de treinamento e otimização dessas redes também são apresentadas.

No Capítulo 4 constam os materiais e os métodos utilizados neste trabalho. São apresentadas as características dos dados e as estratégias de aumento de dados utilizadas. Também são apresentados o ambiente de simulação, os experimentos realizados e descrição da métrica para análise de desempenho.

O Capítulo 5 apresenta os resultados decorrentes do treinamento, validação e teste das arquiteturas desenvolvidas. Os resultados são discutidos e comparados com outros trabalhos relacionados.

2 REVISÃO DA LITERATURA

Com o objetivo de obter o estado da arte do tema "reconhecimento de expressões faciais humanas" foi realizada uma revisão da literatura, a qual incluiu artigos, teses e dissertações. As pesquisas foram realizadas nas plataformas *Engineering Village,* IEEE *Explore* e *Web of Science*, acessadas a partir do Portal de Periódicos - CAPES. Foi adotada uma estratégia de busca avançada em todas as plataformas, e como forma de delimitação, foram utilizadas as palavras-chave *"Facial Expression Recognition", "Neural Network", "Deep Learning"* e *"Facial Landmarks"*.

Além disso, os trabalhos encontrados durante a pesquisa preliminar passaram por um processo de avaliação cuidadosa, com o objetivo de identificar artigos mais recentes e diretamente relevantes ao tema em questão. Durante esse estágio, foi dada preferência aos artigos publicados em revistas, considerando sua excelência, confiabilidade e relevância acadêmica.

A análise dos trabalhos selecionados possibilitou a identificação dos principais métodos aplicados, os quais consistem em redes neurais convolucionais, mecanismos de atenção, redes neurais baseadas em grafos e técnicas clássicas de processamento digital de imagens. Algumas abordagens foram baseadas na composição de uma ou mais técnicas, constituindo assim modelos híbridos. De maneira complementar, foram analisados os artigos relacionados ao uso da base de dados *AffectNet*, referência no problema de reconhecimento de expressões faciais e um dos bancos de imagem mais extensos.

O Quadro 1 apresenta trabalhos relacionados à classificação de emoções humanas através de imagem e redes convolutivas, além de trabalhos que utilizaram pontos de referência da face, conhecidos na literatura como *landmarks*.

2.1 ANÁLISE DOS TRABALHOS - CLASSIFICAÇÃO DE EMOÇÕES UTILIZANDO REDES NEURAIS CONVOLUCIONAIS E/OU PONTOS DE REFERÊNCIA DA FACE

Kim e colaboradores (2017) contribuíram com o desenvolvimento de um classificador de expressões faciais baseado em *deep learning* hierárquico, isto é, as características visuais extraídas são fundidas com as características geométricas de forma hierárquica. A primeira rede recebe como entrada uma imagem após aplicação de LBP, e desta forma aprende características visuais. A segunda rede, extrai as características geométricas relacionadas aos pontos de referência da face, e de maneira complementar, contribui com a classificação final. O pré-processamento realizado consiste na detecção de rostos em uma imagem, seguida de recorte e aplicação de filtro de borramento, a fim de mitigar ruídos. Para a validação final foram utilizadas as bases CK+ e JAFFE, e a acurácia foi utilizada como métrica de avaliação principal, obtendo os valores 96,46% e 91,27% para cada base, respectivamente.

Li et al. (2019) propuseram uma rede neural convolucional com mecanismo de atenção (ACNN) que identifica regiões ocluídas e foca nas áreas desobstruídas mais relevantes. A ACNN, uma estrutura de aprendizado de ponta a ponta, pesa representações de regiões faciais de interesse (ROIs) usando uma unidade de gate adaptativa. Duas versões são apresentadas: pACNN, que foca em patches locais, e gACNN, que integra representações locais e globais. Avaliações em oclusões reais e sintéticas mostraram que as ACNNs melhoram a precisão do reconhecimento, redirecionando a atenção de patches ocluídos para não obstruídos, superando métodos de última geração em vários conjuntos de dados de expressões faciais.

Mollahosseini et al. (2019) apresentam como contribuição a base de dados AffectNet, a maior base de expressões faciais da atualidade, contendo mais de 1 milhão de instâncias. Dentre os resultados apresentados neste trabalho, estão os valores de referência para demais pesquisadores que queiram explorar esse conjunto de dados. Os resultados foram apresentados considerando um modelo de rede convolutiva chamado AlexNet. Além disso, foram consideradas diversas estratégias de particionamento dos dados devido à característica não uniforme do conjunto de dados, sendo elas: *down-sampling, up-sampling* e *weighted-Loss*. Para a composição do Quadro 1, foi considerada a referência de valores do método *Down-Sampling*.

Verma e colaboradores (2019) desenvolveram uma arquitetura de rede composta por dois ramos, um visual e outro cujo domínio de dados são os pontos de referência da face. Além disso, os dados de entrada foram tratados como uma sequência de imagens,

significando a gradação entre um estado neutro até um estado de expressão facial mais intenso. Todas as imagens foram normalizadas, bem como os pontos de referência da face. Cada ramo da rede neural foi pré treinado de forma individual, para que em uma etapa posterior de ajustes, as últimas camadas fossem unificadas e o modelo como um todo treinado. A fusão dos dados de domínios distintos impulsionou a performance do classificador, tendo obtido 97,60% de acurácia na base CK+.

O trabalho desenvolvido por Chen e colaboradores (2022) propôs a utilização de uma *Lightweight Neural Network*, caracterizada pela esparsidade de sua matriz de pesos, o que se traduz em uma quantidade menor de conexões entre os neurônios. Neste trabalho são consideradas convoluções separadas em profundidade, blocos residuais invertidos e camadas de *pooling* globais para reduzir os parâmetros do modelo. Durante o processamento das bases de dados, foi utilizado um algoritmo de detecção facial, e em casos de não haver detecção, a imagem foi excluída. Os rostos detectados foram alinhados e centralizados. Durante a fase de treinamento do modelo foi empregado aumento de dados com as operações de rotação e translação com respeito ao eixo horizontal. A principal métrica utilizada para avaliar o desempenho do modelo treinado foi a acurácia, apresentando um valor de 98,38% para a base de dados CK+.

Soylemez e Ergen (2022) realizaram diversos experimentos variando a área de interesse das faces detectadas, isto é, buscaram investigar os efeitos de realizar o reconhecimento de expressões faciais sob diferentes escalas de aproximação. Para esta tarefa foi realizada a normalização das imagens, o que proporciona uma convergência mais rápida no treinamento de redes neurais. Foram analisadas 3 escalas de aproximação, e comprovou-se que informações que não sejam da face geram impacto no tempo de treinamento e no desempenho do modelo de classificação. Foi utilizada a arquitetura Resnet50 e foram consideradas 6 classes do conjunto de dados CK+. Ao final do treinamento, foi possível obter uma acurácia média de 98% no conjunto de testes.

Huang e colaboradores (2023) utilizaram uma rede neural profunda (DNN) para reconhecimento de emoções faciais (FER). Foi utilizada uma rede neural convolucional (CNN) que combina redes *squeeze-and-excitation* e residual para a tarefa de FER, utilizando as bases de dados AffectNet e Real-World Affective Faces Database (RAF-DB). A análise dos mapas de características mostrou que as regiões ao redor do nariz e da boca são fundamentais para as redes neurais. Validações cruzadas revelaram que o modelo treinado no AffectNet alcançou 77,37% de precisão no RAF-DB, enquanto o modelo pré-treinado no AffectNet e ajustado no RAF-DB obteve 83,37% de precisão.

Por fim, Wadhawan e Gandhi (2023) propuseram o método de transferência de aprendizado da tarefa de detecção dos pontos de referência da face para a tarefa de reconhecimento de expressões faciais. O modelo de detecção mapeia 68 pontos de referência, os quais são subdivididos em 5 grupos, correspondendo ao contorno da sobrancelha, olhos, nariz, boca e mandíbula. Para cada grupo uma rede neural é treinada, tendo como saída um extrator de características baseado em CNN e um localizador de pontos da face. Após o treinamento, a saída do detector de pontos de referência é substituída por uma saída do tipo classificador, e então um ajuste fino é realizado, seguido da técnica de *transfer learning*. Na etapa de pré-processamento os autores empregaram melhorias no contraste das imagens, bem como equalização de histograma. Durante o treinamento ocorreu o aumento de dados, caracterizado por rotação, translação e cisalhamento. Além do mais, as imagens foram normalizadas. Após o treinamento, o modelo foi validado nas bases CK+ e JAFFE, obtendo 97,31% e 97,14% de acurácia, respectivamente.

Quadro 1 - Resumo dos trabalhos relacionados a classificação de emoções utilizando redes neurais convolucionais

Citação	**Base(s) de dados utilizada(s)**	**Dados de entrada**	**Pré-processamento**	**Método**	**Acurácia**	**Outras métricas**
(H. KIM; B. KIM; P. ROY; D. JEONG, 2017)	CK+, JAFFE	Imagens	Detecção e recorte da área do rosto, aplicação de filtro de *blurring* bilateral	Estrutura de rede neural profunda hierárquica	**CK+:** 96,46% **JAFFE:** 91,27%	-
(VERMA; NAKASHIMA; TAKEMURA; NAGAHARA, 2019)	CK+, OULU-CASIA	Imagens + Pontos de referência	Normalização e alinhamento das imagens. Normalização dos pontos de referência da face	Fusão de dados entre arquitetura CNN para imagens e extrator de pontos de referência da face	**CK+:** 97,60% **OULU-CASIA:** 84,17%	-
(MOLLAHOSSEINI; MAHOOR, 2019)	AFFECTNET	Imagens	Cálculo da importância das características.	Rede neural convolucional	**AFFECTNET**: 58%	**F1-Score**: 57% **Kappa**: 51% **Apha**: 51% **AUC**: 85%
(LI et al., 2019)	AFFECTNET	Imagens + mapas de características	Detecção da face, alinhamento e corte da área do rosto	Rede neural convolucional com mecanismo de atenção	**AFFECTNET**: 58,78%	-
(CHEN; JING; ZHANG; MU, 2022)	FER2013, JAFFE, CK+	Imagens	Normalização, aumento de dados	Rede neural convolucional	**FER2013:** 69,51% **CK+:** 98,38% **JAFFE:** 99,17%	-
(SOYLEMEZ; ERGEN., 2022)	CK+	Imagens	Normalização, Corte da área de interesse	Rede neural convolucional	**CK+:** 98%	-

(HUANG et al., 2023)	AFFECTNET	Imagens	Aumento de dados e corte central da área do rosto	Rede neural convolucional	**AFFECTNET**: 56,54%	-
(WADHAWAN;GANDHI, 2023)	JAFFE, CK+	Imagens + Pontos de referência	Ajuste de contraste, detecção e recorte da área do rosto, anotação de pontos de referência da face, aumento de dados	Transfer learning do problema de detecção de pontos de referência da face para o problema de reconhecimento de expressões faciais	**CK+:** 97,31% **JAFFE:** 97,14%	-

2.2 DISCUSSÃO DOS TRABALHOS

Os trabalhos destacam a importância do uso de técnicas como detecção facial, aumento de dados, normalização e pré-processamento para melhorar o desempenho dos modelos de reconhecimento de expressões faciais. Além disso, a utilização de redes neurais profundas e a combinação de diferentes tipos de características (visuais, geométricas) têm se mostrado eficazes para alcançar altas taxas de acurácia nas bases de dados analisadas. A disponibilidade de grandes conjuntos de dados, como a base AffectNet, tem impulsionado o avanço nessa área de pesquisa.

Ao relacionar os trabalhos supracitados, é possível observar algumas tendências e abordagens comuns no reconhecimento de expressões faciais:

- Construção de grandes conjuntos de dados: Mollahosseini et al. (2019) desenvolveram a base de dados AffectNet, que se tornou uma referência para pesquisas na área. A disponibilidade de grandes conjuntos de dados é essencial para o avanço do reconhecimento de expressões faciais, permitindo treinar modelos mais robustos e generalizáveis.
- Fusão de características visuais e geométricas: Verma et al. (2019) propuseram uma arquitetura que utiliza ramos separados para capturar características visuais e geométricas da face. A fusão dessas informações provenientes de diferentes domínios melhorou a performance do classificador.
- Qualidade da imagem avaliada: Li et al. (2019) propuseram uma arquitetura resiliente a oclusões da face. Esta preocupação reflete as condições desafiadoras no que tange à qualidade das imagens do conjunto de dados, o que inclui também a visibilidade da própria face.
- Pré-processamento e aumento de dados: Vários estudos mencionaram o pré-processamento das imagens, incluindo a normalização e melhorias no contraste, bem como o aumento de dados por meio de operações de rotação, translação e cisalhamento. Essas técnicas ajudam a melhorar a generalização do modelo e a lidar com variações nas expressões faciais.
- Importância da escala de aproximação: O estudo de Soylemez e Ergen (2022) ressalta a importância da escala de aproximação no reconhecimento de expressões faciais. Eles demonstraram que a seleção adequada da área de interesse dos rostos pode impactar o desempenho e o tempo de treinamento dos modelos.

- Uso de redes neurais leves: Chen et al. (2022) propuseram o uso de uma *Lightweight Neural Network*, enquanto outros estudos também se preocuparam em reduzir a complexidade dos modelos, como forma de otimizar o desempenho e eficiência computacional.
- Identificação de características relevantes para o reconhecimento de expressões faciais: Huang et al. (2023) estudaram os mapas de características gerados após as convoluções e detectaram que as regiões do nariz e boca são fundamentais para o processo de reconhecimento de expressões faciais. Este dado corrobora o fato de que os pontos de referência associados a estas regiões carregam informação útil para o processo de classificação de expressões faciais.
- Transferência de aprendizado: Wadhawan e Gandhi (2023) exploraram a transferência de aprendizado ao utilizar um modelo pré-treinado para a detecção dos pontos de referência da face e adaptá-lo para o reconhecimento de expressões faciais. Essa abordagem permite aproveitar o conhecimento prévio adquirido em tarefas relacionadas, acelerando o processo de treinamento e melhorando os resultados.

Essas percepções indicam a importância de abordagens eficientes de modelagem, uso de conhecimento prévio, pré-processamento adequado dos dados, consideração da escala de aproximação, fusão de características e disponibilidade de conjuntos de dados abrangentes para avançar na área de reconhecimento de expressões faciais.

3 REFERENCIAL TEÓRICO

Neste capítulo serão apresentados os conceitos relacionados ao desenvolvimento deste trabalho. A Seção 3.1 discorre sobre expressões faciais e suas categorias. A Seção 3.2 discorre sobre os pontos de referência da face, ou *landmarks*. A Seção 3.3 aborda as definições de aprendizagem profunda. Por fim, a Seção 3.4 apresenta os conceitos de redes neurais convolutivas 1D.

3.1 EXPRESSÕES FACIAIS HUMANAS

As expressões faciais universais são padrões de expressões emocionais que são reconhecidos e compreendidos por pessoas de diferentes culturas e origens. Essas expressões são consideradas universais porque são exibidas e reconhecidas em todo o mundo, independentemente das diferenças culturais ou linguísticas. A pesquisa sobre expressões faciais universais tem sido amplamente estudada e discutida na psicologia e nas ciências cognitivas.

Um dos principais pesquisadores a estudar as expressões faciais universais foi Paul Ekman (EKMAN; FRIESEN, 1971). Em seus estudos pioneiros, Ekman (1971) identificou seis expressões faciais básicas que são consideradas universais: felicidade, tristeza, raiva, medo, surpresa/espanto e aversão (nojo), assim como ilustra a Figura 1. Ele argumentou que essas expressões são inatas e biologicamente determinadas. Além de que as pessoas as exibem e reconhecem independentemente da cultura.

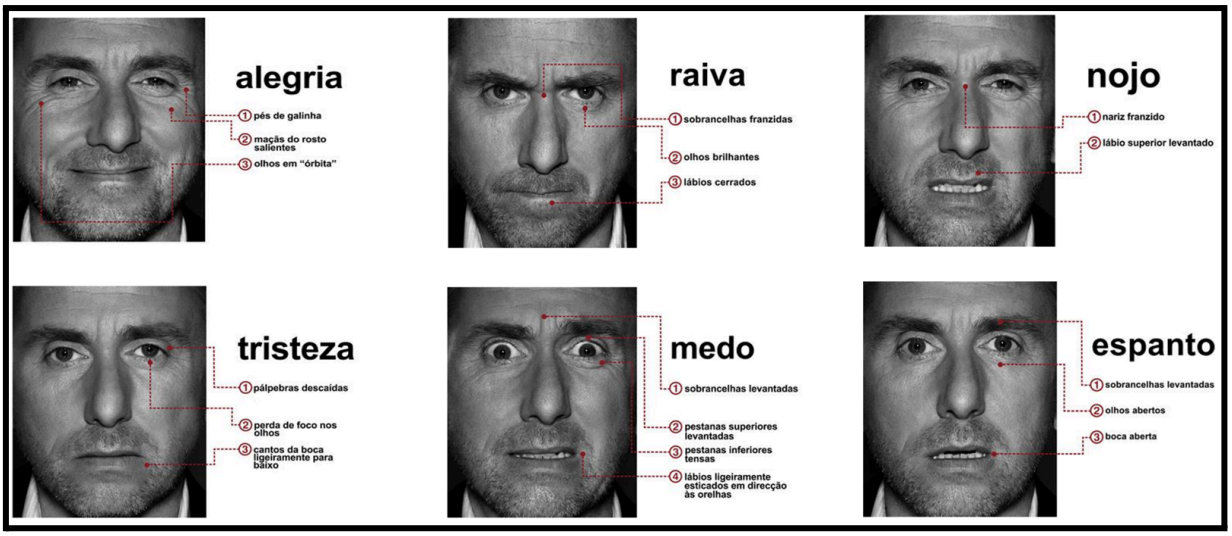


Figura 1 - As 6 expressões faciais de acordo com Paul Ekman. Fonte: Adaptado de Ekman, 2011

Essas descobertas foram apoiadas por estudos transculturais que mostraram a universalidade das expressões faciais. Um estudo conduzido por Matsumoto e Hwang (2019) examinou as expressões faciais em diferentes culturas e descobriu que as seis emoções básicas identificadas por Ekman (1971) eram reconhecidas com alta precisão em todas as culturas estudadas. Isso sugere que a capacidade de reconhecer e interpretar expressões faciais é inata e compartilhada pela humanidade.

Além das pesquisas sobre as emoções básicas, os pesquisadores também investigaram as variações culturais nas expressões faciais. Estudos como o de Elfenbein e Ambady (2002) mostraram que, embora as expressões faciais básicas sejam universais, as culturas podem ter regras específicas sobre como e quando exibir certas emoções. Essas regras culturais podem influenciar a forma como as pessoas interpretam e respondem às expressões faciais em diferentes contextos sociais.

Outro estudo importante realizado por Jack e Schyns (2014) mostrou que as expressões faciais universais são processadas de forma rápida e automática pelo cérebro humano, independentemente da atenção consciente. Isso sugere que a capacidade de reconhecer e interpretar expressões faciais universais é um processo inato e fundamental.

Apesar da literatura de Paul Ekman (1971) especificar 6 expressões faciais, em trabalhos de reconhecimento de emoções humanas, considera-se também o estado neutro, ou seja, sem emoções aparentes, como uma categoria adicional.

3.2 PONTOS DE REFERÊNCIA DA FACE - LANDMARKS

Os pontos de referência faciais, conhecidos como *facial landmarks* em inglês, são pontos anatômicos específicos na face humana que desempenham um papel importante na identificação e localização de características faciais distintas. Esses pontos são amplamente utilizados em diversas aplicações, como reconhecimento facial automatizado, análise de expressões faciais, realidade aumentada, entre outras. Vários estudos têm sido realizados para explorar as características e a relevância desses pontos de referência faciais.

Um estudo significativo na detecção e rastreamento de pontos de referência faciais foi conduzido por Saragih, Lucey e Cohn (2010). Nessa pesquisa, foi proposto um método baseado em descritores de características e modelos de regressão para localizar com precisão os pontos de referência faciais em tempo real. Os resultados obtidos demonstraram a eficácia e a utilidade desses pontos de referência em tarefas como análise de expressões faciais e rastreamento de movimentos faciais.

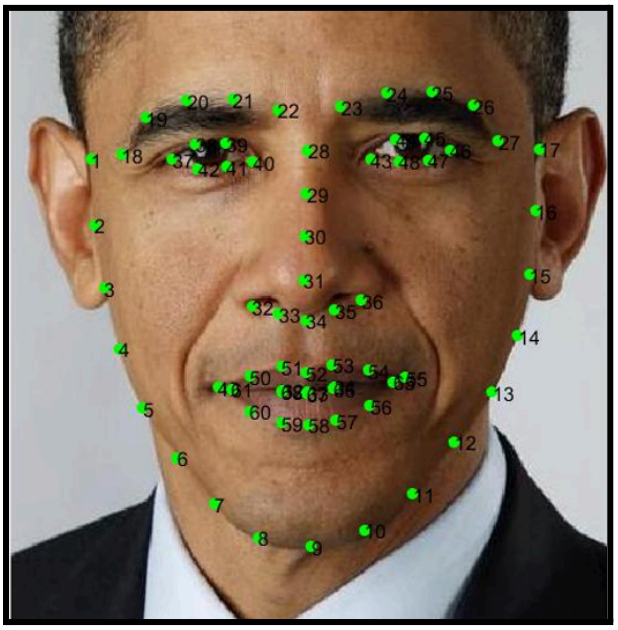


Figura 2 - Pontos de referência da face - *Facial Landmarks.* Fonte: Adaptado de Lai et al., 2016

Conforme elucidado na Figura 2, os pontos de referência podem ser referenciados em 2 coordenadas (x e y) em uma dada imagem. Atualmente, diversas bibliotecas na área de visão computacional proporcionam modelos ou ferramentas para a detecção e extração de *landmarks* em um rosto humano, sendo comum o fornecimento de no mínimo 68 pontos.

3.3 APRENDIZAGEM PROFUNDA

A inteligência artificial (IA) é um campo de estudo que busca desenvolver sistemas e algoritmos capazes de realizar tarefas que normalmente exigiriam inteligência humana. No contexto da IA, existem duas abordagens principais de aprendizado: o aprendizado clássico e o aprendizado profundo.

O aprendizado clássico, também conhecido como aprendizado de máquina tradicional, envolve a construção de modelos e algoritmos que extraem características relevantes dos dados e utilizam essas características para fazer previsões ou tomar decisões. Essa abordagem geralmente requer que os especialistas definam manualmente as características a serem extraídas, o que pode ser uma tarefa complexa e demorada.

Por outro lado, o aprendizado profundo é uma abordagem mais recente e poderosa dentro da IA, que se baseia em redes neurais artificiais profundas. As redes neurais profundas são capazes de aprender automaticamente representações de alto nível a partir dos dados, permitindo a extração de características complexas e sutis. Segundo Goodfellow et al. (2016), "o aprendizado profundo é uma abordagem para IA que busca aprender representações

múltiplas e em camadas de dados, com cada camada de representação se tornando gradualmente mais abstrata".

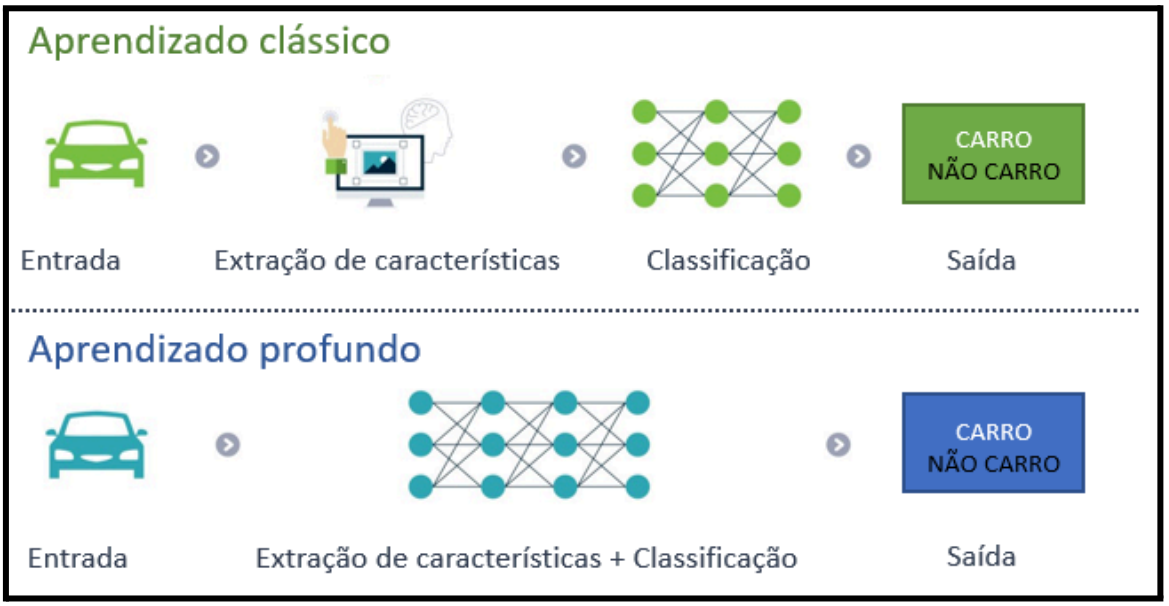


Figura 3 - Aprendizado profundo e aprendizado clássico.

Conforme a Figura 3 ilustra, o aprendizado profundo muitas das vezes pode automatizar o processo de seleção de características, criando inclusive relações entre os dados que poderiam não ser notadas via seleção manual.

O aprendizado profundo tem demonstrado um desempenho impressionante em várias tarefas de IA, como reconhecimento de imagens, processamento de linguagem natural, detecção de anomalias e muito mais. Através de suas camadas de neurônios, as redes neurais profundas são capazes de aprender representações abstratas e complexas dos dados, o que as torna particularmente eficazes em problemas de alta dimensionalidade. LeCun et al. (2015) destacam que "o aprendizado profundo tem o potencial de revolucionar muitas áreas da ciência e da tecnologia, especialmente aquelas relacionadas ao processamento e à compreensão de dados sensoriais, como visão computacional e reconhecimento de voz".

No contexto da relação entre IA, aprendizado profundo e aprendizado clássico, é importante destacar que o aprendizado profundo tem sido considerado uma extensão do aprendizado clássico. Embora o aprendizado clássico ainda seja amplamente utilizado e tenha suas aplicações, o aprendizado profundo tem ganhado destaque devido à sua capacidade de aprender automaticamente representações de alto nível, eliminando a necessidade de características pré-definidas. Segundo Bishop (2006), "o aprendizado profundo pode ser visto

como uma maneira de aprender representações automáticas de dados que são cada vez mais complexas e abstratas".

3.4 REDES NEURAIS CONVOLUTIVAS

As redes neurais convolutivas (CNNs) são uma classe de arquiteturas de aprendizado profundo que foram desenvolvidas para processar dados de natureza espacial, como imagens. Elas se destacaram em tarefas de visão computacional devido à sua capacidade de aprender características hierárquicas e complexas dos dados de entrada. As CNNs aplicam operações de convolução, *pooling* e camadas densas para processar as informações.

3.4.1 Camada convolutiva

A camada de convolução em redes neurais convolucionais (CNNs) é uma operação fundamental que desempenha um papel crucial na extração de características visuais. Matematicamente, a convolução é definida como uma operação linear entre uma matriz de entrada e um filtro, seguida por uma função de ativação não-linear. Essa operação é realizada em cada posição da entrada para gerar um mapa de características.

A fórmula matemática para a convolução em uma posição específica é dada por:

$$(I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (1)$$

Na Equação (1), I representa a matriz de entrada, K o filtro convolucional e (I*K) o resultado da convolução.

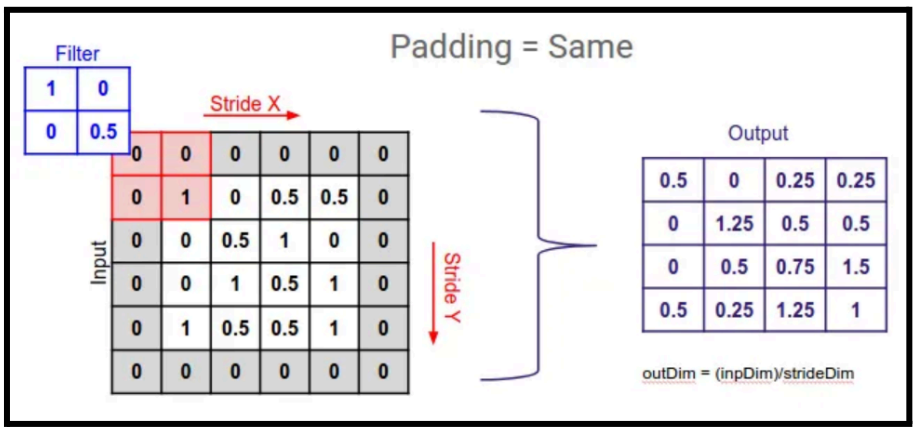


Figura 4 - Operação de convolução. Fonte: (MEDIUM, 2019)

Na Figura 4, um aspecto intrínseco à operação de convolução está sendo ilustrado, isto é, a utilização de *paddings* e o parâmetro de *stride*. Em termos práticos, os *paddings* referem-se à adição de valores extras ao redor da matriz de entrada antes de aplicar o filtro convolucional. O objetivo do padding é controlar o tamanho do mapa de características resultante e preservar as informações nas bordas da matriz de entrada. Já o *stride* indica o avanço, ou o passo com quem o filtro percorre cada linha da matriz alvo.

Durante o processo de treinamento, os filtros convolucionais são ajustados iterativamente usando algoritmos de otimização, como o gradiente descendente, para minimizar uma função de perda. Isso permite que a rede aprenda automaticamente os padrões e características relevantes nos dados de entrada.

A camada de convolução em CNNs desempenha um papel fundamental na extração de características locais, como bordas, texturas e padrões. Essas características são essenciais para a identificação de objetos e reconhecimento de padrões em imagens.

3.4.2 Camada de *pooling*

As camadas de *pooling* são responsáveis por reduzir a dimensionalidade dos dados, preservando as características mais relevantes. A técnica de *pooling* mais comumente utilizada é o *max pooling*, onde apenas o valor máximo de uma região é preservado.

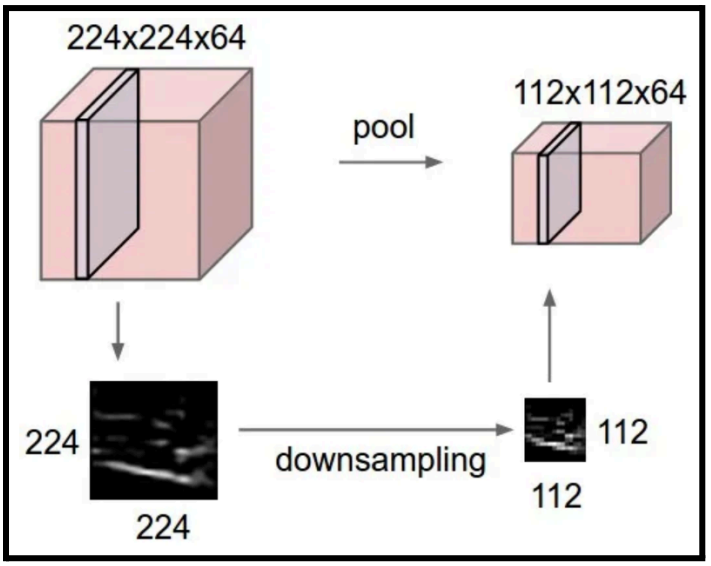


Figura 5 - Operação de *pooling*. Fonte: (MEDIUM, 2019)

3.4.3 Camadas densas

As camadas densas (ou completamente conectadas) são uma parte fundamental das redes neurais, onde cada neurônio em uma camada está conectado a todos os neurônios na camada anterior. Essas camadas são responsáveis por aprender representações mais abstratas e complexas dos dados de entrada.

Matematicamente, a camada densa é representada por uma transformação linear seguida de uma função de ativação. Supondo que existam *N* neurônios na camada anterior e *M* neurônios na camada densa, cada neurônio *j* na camada densa calcula uma combinação linear das saídas dos neurônios da camada anterior ponderada pelos pesos Wij , seguido pela adição de um termo de bias bij :

$$Zj = \sum_{i=1}^{N} W_{ij}.x_i + bj \quad (2)$$

Na Equação (2), x_i é a saída do neurônio i na camada anterior. Em seguida, a função de ativação não linear, geralmente aplicada elemento a elemento, é aplicada no resultado z_j para obter a saída y_j do neurônio j, onde f é a função de ativação:

$$y_j = f(z_i) \quad (3)$$

A matriz de pesos W e o vetor de bias b são os parâmetros que a rede neural aprende durante o processo de treinamento, ajustando seus valores para otimizar a tarefa em questão.

3.4.4 Redes convolutivas de 1 dimensão

Redes convolucionais de 1 dimensão (1D) são amplamente utilizadas para processar dados sequenciais unidimensionais, como séries temporais e dados de linguagem natural. Essas redes aplicam filtros convolucionais ao longo da dimensão da sequência para capturar informações locais e extrair características relevantes.

A operação de convolução em redes convolucionais 1D é matematicamente definida como a multiplicação entre um filtro (ou kernel) e uma janela deslizante da entrada. Isso é seguido por uma função de ativação não linear para introduzir a capacidade de aprendizado não linear na rede.

Uma referência importante no desenvolvimento das redes convolucionais 1D é o trabalho de Zhang et al. (2018), que propõe o uso de redes convolucionais de caracteres para

classificação de texto. O estudo demonstra a eficácia das redes convolucionais 1D na extração de informações discriminativas de sequências de caracteres.

Outro estudo relevante é o de Sainath et al. (2015), que descreve o uso de redes convolucionais 1D para aprendizado de recursos a partir de formas de onda brutas de fala. O trabalho destaca a capacidade das redes convolucionais 1D em processar dados sequenciais e extrair características relevantes para a tarefa de reconhecimento de fala.

De maneira complementar, Santana e colaboradores (2021) trabalharam com séries temporais para a classificação de arritmias cardíacas, e no âmbito desta pesquisa, foram elencadas duas arquiteturas de redes convolutivas, sendo uma delas unidimensional.

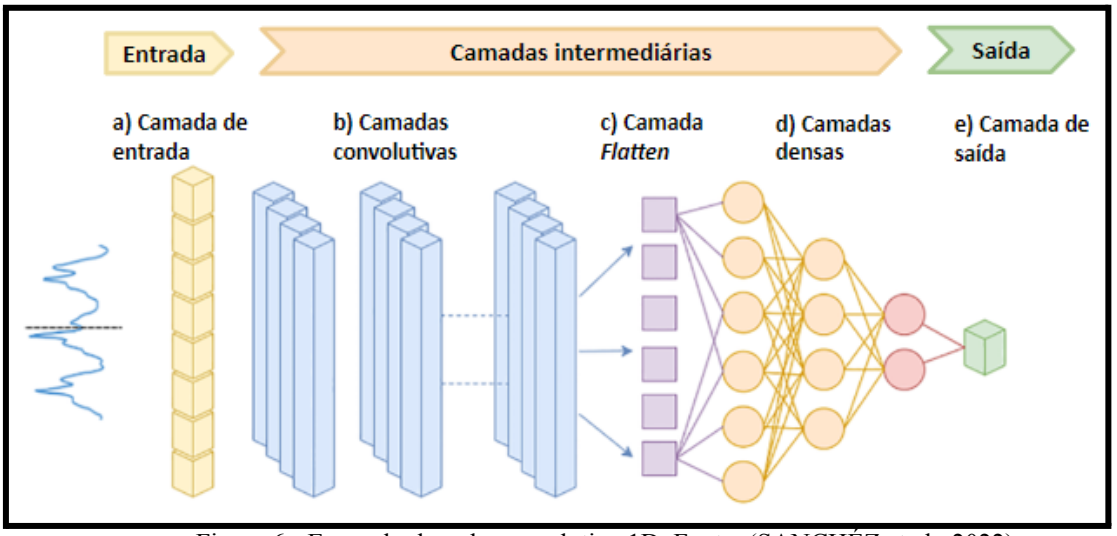


Figura 6 - Exemplo de rede convolutiva 1D. Fonte: (SANCHÉZ et al., 2022)

A Figura 6 ilustra um exemplo de rede convolutiva 1D. Como pode ser observado, a topologia da rede muito se assemelha aos padrões de redes convolutivas 2D. A principal diferença está na dimensão dos kernels de convolução logo nas camadas iniciais, que deverão estar de acordo com a característica dos dados de entrada.

4 MATERIAIS E MÉTODOS

Para avaliar se redes neurais que realizam a fusão de características provenientes dos pontos de referência da face e da própria imagem da face são capazes de melhorar o reconhecimento de expressões faciais, é necessário treinar modelos com a fusão dessas técnicas de classificação. Neste capítulo, são descritos os materiais necessários para executar o treinamento deste modelo, bem como os métodos empregados para contornar problemas de desbalanceamento da base de dados, melhorando a generalização do mesmo.

4.1 MATERIAIS

4.1.1 Conjunto de dados

A base de dados utilizada neste trabalho foi a *AffecNet,* disponibilizada por Mollahosseini et al. (2019). A base de dados é distribuída oficialmente com os conjuntos de treinamento e validação, contendo 287401 e 3500 amostras, respectivamente (considerando a versão com 7 classes).

As amostras da base foram agrupadas em pastas, sendo elas nomeadas de acordo com as expressões faciais alvo (*neutral, happy, sad, surprise, fear, disgust, anger)*, as quais correspondem às classes no procedimento de classificação. Todas as imagens foram dispostas com as dimensões 224 x 224 pixels, no padrão RGB. A Tabela 1 detalha as expressões faciais existentes no conjunto de dados, bem como a sua distribuição quantitativa.

Tabela 1 - Distribuição dos dados.

Índice	Classe	Quantidade
0	*Neutral* - Neutro	75374
1	*Happy* - Feliz	134915
2	*Sad* - Triste	25959
3	*Surprise* - Surpresa	14590
4	*Fear* - Medo	6878
5	*Disgust* - Nojo	4304
6	*Anger* - Raiva	25382

Fonte: (MOLLAHOSSEINI et al., 2019).

Com o objetivo de melhor visualizar a distribuição dos dados, foi elaborado um histograma a partir do conjunto de treinamento, o qual pode ser visto na Figura 7.

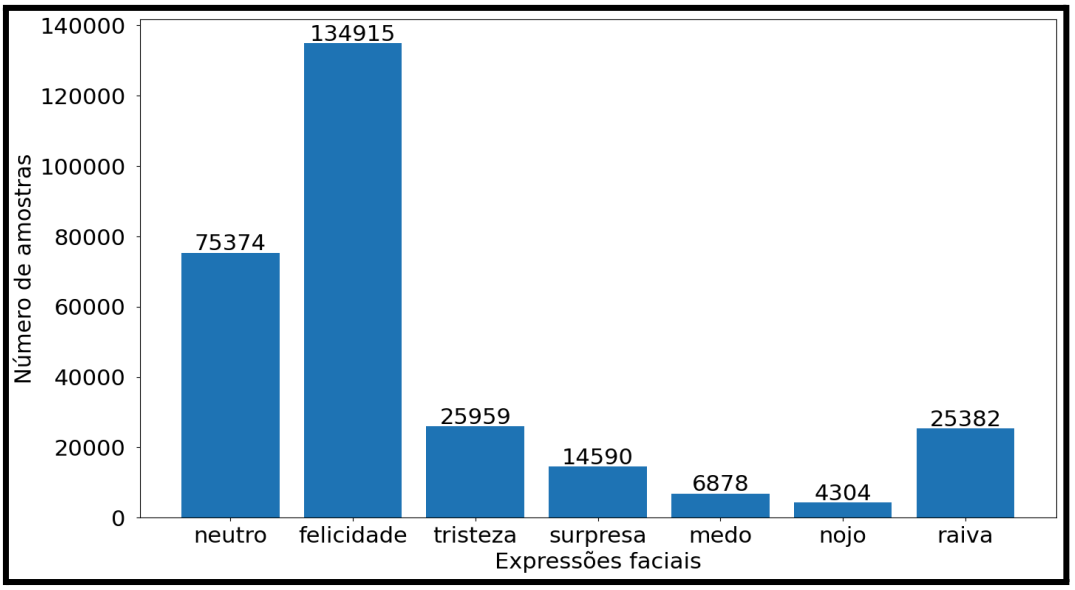


Figura 7 - Distribuição de classes

Conforme pode ser observado na Figura 7, há um claro desbalanceamento entre as classes. Mollahosseini et al. (2019), ao observar esta distribuição, sugeriu técnicas de balanceamento, a exemplo de *downsampling* - remoção de amostras, *upsampling* - adição de amostras ou aplicação de pesos durante o treinamento (*class-weights),* tendo maior sucesso nesta última.

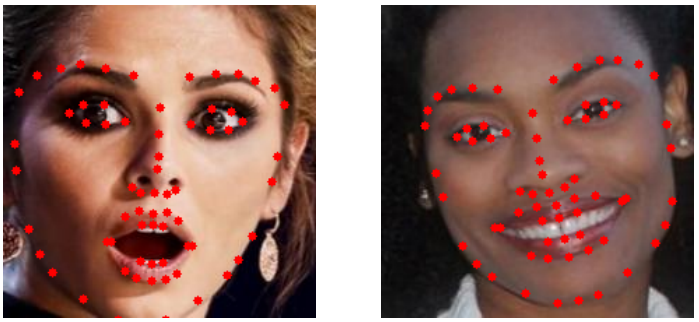
As imagens presentes na base foram coletadas, em sua maioria, a partir de mecanismos de busca na *internet*. Desta maneira, os rostos não estão dispostos em condições controladas de laboratório. A Figura 8(a) exemplifica uma amostra da base de dados categorizada como a expressão surpresa e a Figura 8(b) exemplifica uma amostra da classe feliz.

(a) (b)

Figura 8 - Amostras da base de dados utilizada neste trabalho: (a) Surpresa; (b) Feliz

Com o objetivo de avaliar a relação dos pontos de referência da face com as expressões faciais, foram realizados testes preliminares considerando o pacote *mediapipe* da linguagem Python para localizar e representar estes pontos. A função *face_mesh.process d*o pacote *mediapipe* tem como parâmetro uma imagem no padrão RGB e retorna um objeto

contendo as coordenadas dos pontos de referência da face, as quais foram impressas em vermelho na imagem original, a partir da biblioteca openCV, conforme pode ser visto nas Figuras 9(a) e 9(b).



(a) (b)

Figura 9 - Pontos de referência da face para as expressões: (a) Surpresa e (b) Feliz

Ainda de maneira preliminar, o estudo da base de dados permitiu associar algumas características dos pontos de referência da face à expressão facial. Isto é, na Figura 9(a) é possível perceber que os pontos no contorno da boca se organizam de forma circular para a expressão de surpresa, enquanto que, para a expressão de felicidade, se organizam de forma horizontal. Desta forma, evidenciou-se a relação entre as expressões faciais e os pontos de referência da face.

4.1.2 Ambiente utilizado para a implementação

Este trabalho foi implementado integralmente através da linguagem de programação Python 3.8. Para a etapa de preparação da base de dados foi utilizado um computador Intel I7-11800H com 2.30GHz, 32GB de memória RAM e uma GPU GeForce RTX3050 com 4GB e 2560 núcleos CUDA.

O ambiente colaborativo Google Colab Pro foi utilizado para a realização dos experimentos e análise de desempenho. Para este ambiente foi alocada uma GPU modelo Tesla T4 com 16GB e 2560 núcleos CUDA.

Foram importados os pacotes Tensorflow versão 2.9.1, Keras versão 2.9.0, Pandas versão 1.3.5, Numpy versão 1.21.6, Matplotlib versão 3.2.2 e OpenCV versão 4.7.0.

A integração do pacote Keras com a linguagem de programação Python viabiliza a utilização das mais diversas camadas presentes em arquiteturas de redes neurais, a exemplo das camadas densas, normalização de lote, agrupamento máximo, *dropout* e camadas

convolutivas. Além do mais, a importação de redes pré-treinadas também está disponível nesta ferramenta.

A Figura 10 apresenta uma arquitetura de rede neural convolucional 1D criada por intermédio do pacote Keras. Nesta rede as camadas convolutivas e normalização de lotes são intercaladas 7 vezes, para que por fim as características extraídas sejam propagadas até as últimas camadas inteiramente conectadas.

```
model = Sequential()
model.add(Conv1D(filters=96, kernel_size=2, activation='relu',
                 input_shape=(136, 1)))
model.add(BatchNormalization())
model.add(Conv1D(filters=96, kernel_size=2, activation='relu'))
model.add(BatchNormalization())
model.add(Conv1D(filters=96, kernel_size=2, activation='relu'))
model.add(BatchNormalization())
model.add(Conv1D(filters=96, kernel_size=2, activation='relu'))
model.add(BatchNormalization())
model.add(Conv1D(filters=96, kernel_size=2, activation='relu'))
model.add(BatchNormalization())
model.add(Conv1D(filters=96, kernel_size=2, activation='relu'))
model.add(BatchNormalization())
model.add(Conv1D(filters=96, kernel_size=2, activation='relu'))
model.add(BatchNormalization())
model.add(Flatten())
model.add(Dense(200, activation='relu'))
model.add(Dropout(0.1))
model.add(Dense(7, activation='softmax'))
```

Figura 10 - Ilustração de uma arquitetura de rede neural implementada na linguagem Python

Além da definição da arquitetura da rede, o pacote Keras também fornece mecanismos para controle de parada antecipada do treinamento do modelo, conhecido como *EarlyStopping*. Como critério de parada antecipada adotou-se uma paciência de 3 considerando a métrica *val_loss*, isto é, caso a perda no conjunto de validação não diminua por 3 épocas consecutivas, o treinamento é interrompido. Além disso, também há mecanismos para checagem e armazenamento das melhores versões de modelos ao longo de cada etapa de validação durante o treinamento.

Os demais parâmetros de treinamento adotados constituíram-se do otimizador ADAM, da função de perda *categorical_crossentropy*, do número de épocas escolhidas para o treinamento definida em 60, tamanho do lote de dados, conjunto de treinamento, conjunto de validação, e por fim, o parâmetro *class_weight*, futuramente empregado para aplicar pesos proporcionais ao volume de dados de cada classe durante o processo de treinamento.

A função de perda *categorical_crossentropy*, ou entropia cruzada categórica, tem larga utilização em problemas de classificação multiclasse em aprendizado de máquina (GOODFELLOW; BENGIO; COURVILLE, 2016). Esta função mensura a discrepância entre a distribuição verdadeira dos rótulos das classes e a distribuição prevista pelo modelo. Para uma amostra, a entropia cruzada pode ser definida tal como na Equação (4).

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \tag{4}$$

Na Equação 4, C representa o número de classes, y é o vetor que representa a classe verdadeira da amostra e $\widehat{y}_i$ é a probabilidade prevista pelo modelo para cada classe.

De maneira complementar, a Equação (5) ilustra que para um conjunto de dados com N amostras, a entropia cruzada categórica média é calculada como a média das perdas individuais de todas as amostras:

$$L_{média} = \frac{1}{N}\sum_{j=1}^{N} L(y^{(j)}, \hat{y}^{(j)}) \tag{5}$$

Estes parâmetros foram configurados tal como se ilustra na Figura 11.

```
model.compile(optimizer='adam',
              loss='categorical_crossentropy',
              metrics=['accuracy'])


history = model.fit(train_data_generator,
                    epochs=60,
                    validation_data=val_data_generator,
                    batch_size=8,
                    callbacks = [earlystopping, model_checkpoint],
                    class_weight=class_weights
                    )
```

Figura 11 - Ilustração dos parâmetros de treinamento do modelo

4.2 MÉTODOS

O presente trabalho propõe um método para classificação de expressões faciais com base na fusão das características dos pontos de referência da face e da imagem da própria face, empregando redes neurais profundas. Para que os objetivos geral e específicos deste trabalho fossem alcançados, foram seguidas as etapas apresentadas no diagrama em blocos da Figura 12.

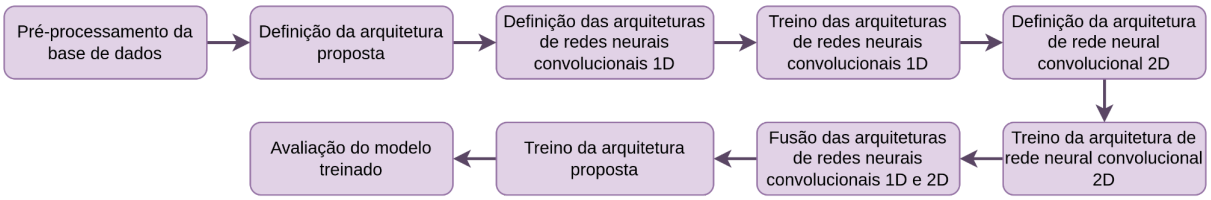


Figura 12 - Etapas de desenvolvimento do sistema proposto

4.2.1 Pré-processamento da base de dados

A base de dados AffectNet é distribuída apenas com os conjuntos de treinamento e validação. Não há oficialmente um conjunto de testes estabelecido. Por este motivo, Mollahosseini et al. (2019) recomendam que os demais pesquisadores realizem as comparações e análises de desempenho considerando o conjunto de validação.

Para que a capacidade de generalização do modelo seja alcançada e para que a comparação com os demais trabalhos seja válida, optou-se por considerar a partição de validação original como a partição de teste para este problema, e por conseguinte, a partição de treinamento original foi segregada em treinamento e validação, em uma proporção de 90% e 10%, respectivamente. A Figura 13 ilustra este processo de organização da base de dados.

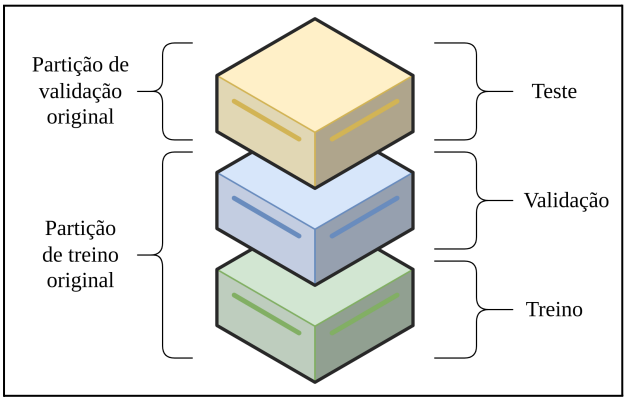


Figura 13 - Particionamento da base de dados

Considerando a distribuição de dados presente na Tabela 1 e o particionamento estabelecido na etapa de organização da base de dados, é esperado que os subconjuntos de treinamento, validação e teste possuam as quantidades previstas na Tabela 2.

Tabela 2 - Organização da base de dados

Classe	**Total de amostras**	**Conjunto de treino original**		**Conjunto de validação original**
		Treino (90%)	**Validação (10%)**	**Teste**
Neutro	75374	67387	7487	500
Feliz	134915	120974	13441	500
Triste	25959	22914	2545	500
Surpresa	14590	12681	1409	500
Medo	6878	5741	637	500
Nojo	4304	3424	380	500
Raiva	25382	22394	2488	500

As amostras da base de dados são compostas em sua totalidade por imagens de rostos de pessoas. Dada a proposta deste trabalho, foi necessária a manipulação de dois tipos de dados: a imagem da face e seus pontos de referência. A imagem da face caracteriza-se por um tipo de dados bidimensional, a ser processado por uma rede neural convolucional tradicional, a exemplo da *GoogleNet, AlexNet, VGG16, ResNet* e outras. Por outro lado, os pontos de referência da face são dados unidimensionais, necessitando de uma rede neural compatível com esta modalidade. Conforme ilustrado pela Figura 14, os passos de pré-processamento são distintos para cada tipo de dado. A normalização prevista para os dados de imagem é tida como padrão em problemas de classificação, enquanto que a detecção da face em uma etapa anterior à extração dos pontos de referência opera como um passo mandatório.

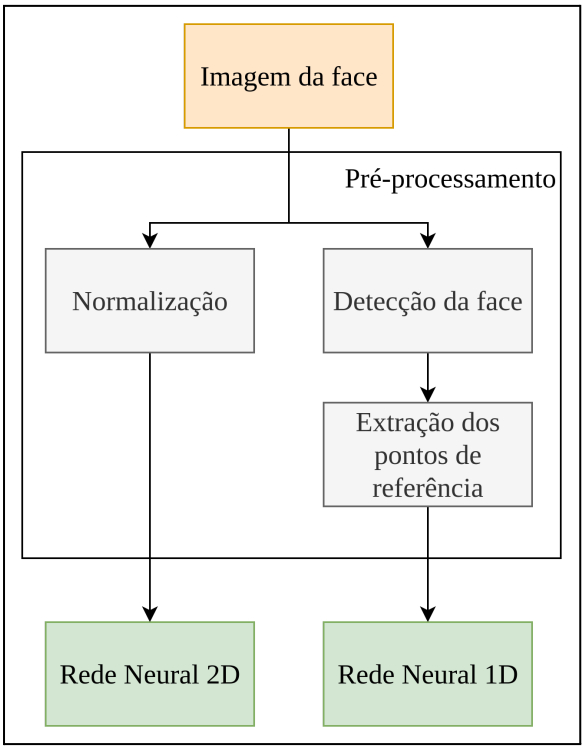


Figura 14 - Pré-processamento dos dados

A partir da organização dos conjuntos de treinamento, validação e teste, foi realizada uma etapa posterior de extração de dados. Esta extração leva em consideração os pontos de referência da face de cada uma das imagens.

Dentre as principais motivações para a utilização de pontos de referência da face para o treinamento de modelos de reconhecimento de expressões faciais, estão:

- Compressão de dados: uma única imagem colorida, nas dimensões 224x224x3 pode chegar a ocupar 7,7KB em memória. Ao utilizar esse dado como entrada de uma rede neural, o treinamento tende a ter um maior tempo de execução, considerando o tamanho da base de dados. Enquanto que 68 pontos de referência da face chegam a ocupar apenas 544B em memória, considerando um vetor de *float* com 136 posições.
- Segurança e privacidade: um modelo treinado apenas com pontos de referência da face dispensa qualquer tipo de viés relacionado à aparência física, além de que os dados de imagem não são aproveitados pelo modelo.

De posse dessas informações, buscou-se projetar um modelo de rede neural cuja entrada fosse composta pelos 68 pontos de referência da face. Como cada ponto é descrito por 2 coordenadas, o tamanho total do vetor totalizou 136 coordenadas.

x_1	y_1	x_2	y_2	...	...	x_{67}	y_{67}	x_{68}	y_{68}

Figura 15 - Disposição dos dados na entrada da rede 1D

O vetor de características é composto das coordenadas dispostas de maneira sequencial. Os valores de x são sucedidos pelos valores de y para todos os pontos, assim como ilustra a Figura 15.

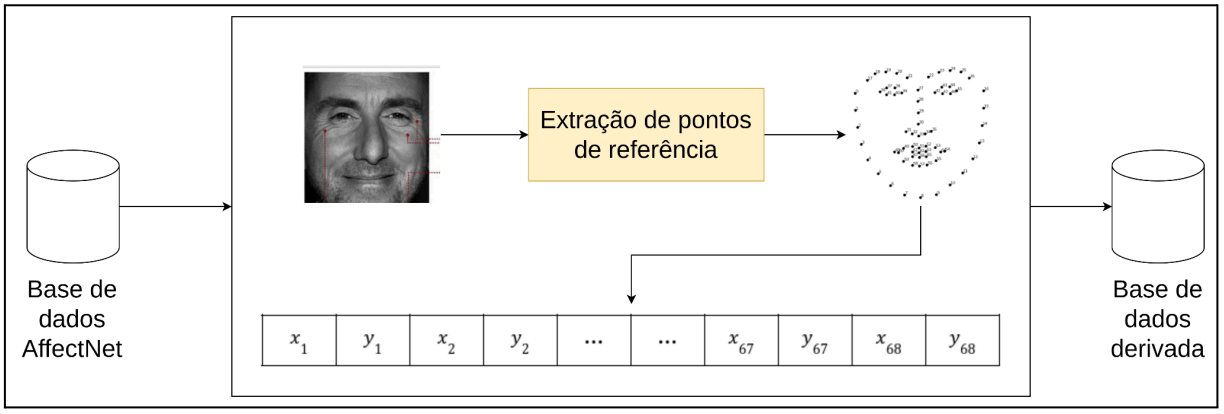


Figura 16 - Aumento de dados seguido de extração de características

A Figura 16 ilustra o processo de obtenção de uma variante da base de dados *AffectNet*. Conforme ilustrado, para cada imagem da base de dados ocorre o processo de extração dos pontos de referência da face através da biblioteca *mediapipe* da linguagem de programação Python. Desta maneira, uma segunda base de dados é gerada contendo apenas os pontos de referência da face extraídos de cada imagem.

O processo de extração de pontos de referência de face foi aplicado em toda a base de dados disponível. Este processo foi conduzido de maneira interativa, isto é, as primeiras imagens foram inspecionadas com os pontos de referência sobrepostos, tal como ilustra a Figura 9. Desta forma, assegurou-se que os pontos foram extraídos corretamente.

Por outro lado, foram observadas amostras de imagens não conformes para o problema em questão, seja por oclusão, seja por incongruência entre a classe alvo e o que está representado na imagem. Exemplos destas imagens podem ser encontrados na Figura 17.

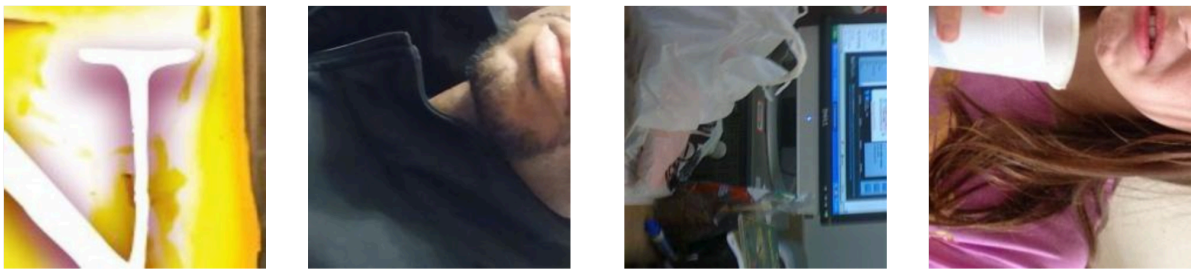


Imagem anotada com a classe *Neutral* | Imagem anotada com a classe *Neutral* | Imagem anotada com a classe *Neutral* | Imagem anotada com a classe *Disgust*

Figura 17 - Exemplos de imagens que foram descartadas

Por não ser possível a detecção dos pontos de referência para estas imagens, este processo atuou também como um filtro, eliminando ruídos provenientes da anotação da base de dados e contribuindo, portanto, para uma melhor qualidade dos dados de treinamento do modelo. A consequência direta deste filtro foi a eliminação de amostras da partição de treinamento, a qual teve uma diminuição, conforme mostrado na Tabela 3.

Tabela 3 - Distribuição dos dados resultantes do processo de extração de pontos de referência

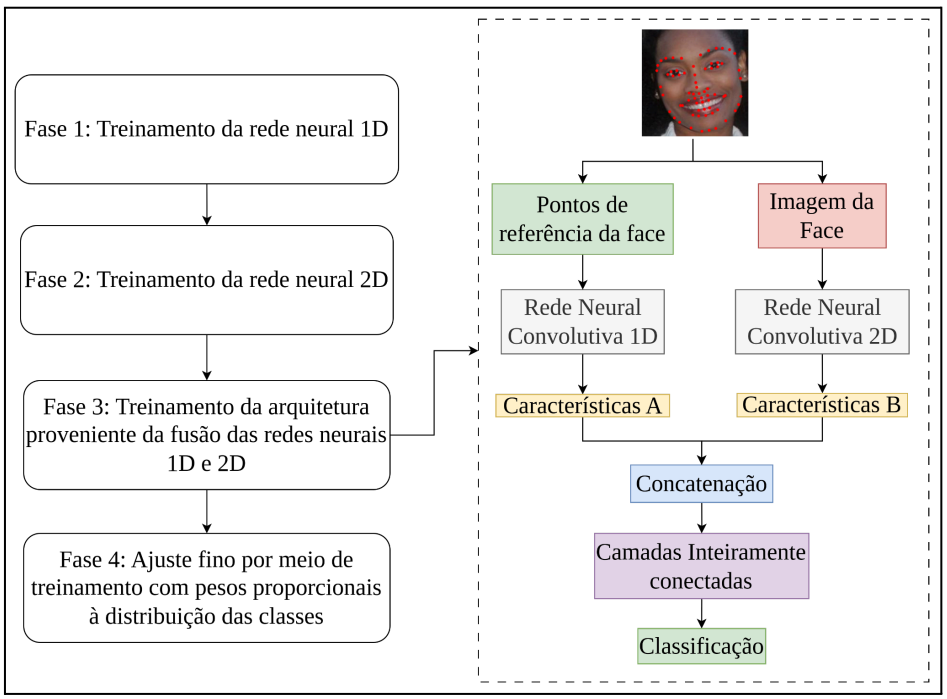
Índice de classe	**Classe**	**Total de amostras antes da extração**	**Total de amostras após a extração**	**Imagens descartadas**	**Percentual de perda (%)**
0	Neutro	75374	74743	631	0,84
1	Feliz	134915	134181	734	0,54
2	Triste	25959	25395	564	2,17
3	Surpresa	14590	14065	525	3,60
4	Medo	6878	6363	515	7,49
5	Nojo	4304	3797	507	11,78
6	Raiva	25382	24789	593	2,34

De acordo com o indicado na Tabela 3, todas as classes sofreram uma redução na quantidade de imagens. Esta redução atingiu um percentual máximo de 11,78% para a classe 5, o que se justifica pela relação entre a quantidade original de exemplos desta classe na base de dados e sua relação com a quantidade de imagens descartadas.

4.2.2 Arquitetura da rede neural proposta

Para o reconhecimento de expressões faciais, foi proposta uma arquitetura de rede neural que comporta a fusão de dados heterogêneos. Os dados referentes aos pontos de referência da face e à imagem da face são processados por uma rede neural convolucional 1D e por uma rede neural convolucional 2D, respectivamente. A arquitetura e o mecanismo de treinamento propostos podem ser visualizados na Figura 18.

Conforme ilustrado na Figura 18(a), nas fases 1 e 2 do treinamento, realiza-se o treinamento separado das redes 1D e 2D. Os dados 1D, contendo pontos de referência da face são utilizados para o treinamento da rede neural 1D, enquanto que, os dados 2D, contendo imagens da face são utilizados para o treinamento da rede neural 2D. Após o treinamento, os pesos dessas redes são congelados e as mesmas são unidas em uma única arquitetura de rede, conforme mostrado na Figura 18(b). Da última camada de classificação de cada uma das redes, então, são extraídas características, características A e B, que são concatenadas em um único vetor. Na fase 3 do treinamento, são ajustados os parâmetros das últimas camadas inteiramente conectadas, que realizam a classificação. Por fim, na fase 4, realiza-se um ajuste fino dos valores desses parâmetros, através da realização de um treinamento com ponderação dos pesos.



(a) (b)

Figura 18 – Arquiteturas e mecanismo de treinamento propostos;
(a) Mecanismo de treinamento; (b) Arquitetura proposta

O mecanismo de treinamento do modelo de rede neural envolveu, então, 4 etapas, listadas a seguir:

- Fase 1: Treinamento da rede neural 1D;
- Fase 2: Treinamento da rede neural 2D;
- Fase 3: Treinamento da arquitetura proveniente da fusão das redes neurais 1D e 2D;
- Fase 4: Ajuste fino por meio de treinamento com pesos proporcionais à distribuição das classes.

A fase 1 teve como objetivo selecionar a arquitetura de rede neural convolucional unidimensional mais aderente ao problema em questão. Para fins de validação foram selecionadas as arquiteturas apresentadas nos trabalhos de Machine Learning (2020), Azizjon et al. (2020), e Santana et al. (2021). Como critério de seleção adotou-se o melhor desempenho em termos de acurácia obtido no treinamento destas redes considerando o problema de reconhecimento de expressões faciais.

A fase 2 teve como objetivo selecionar uma arquitetura de rede neural convolucional aplicável ao problema de classificação de imagens. As redes neurais convolucionais 2D utilizadas para a extração de características da imagem do rosto foram a DenseNet121 e a DenseNet169. Essas CNN's são modelos de rede profunda pré-treinadas disponíveis no pacote *TensorFlow* da linguagem Python. Ambas as arquiteturas DenseNet são conhecidas pela sua conectividade densa, em que cada camada recebe dados não só da sua antecessora, mas também de todas as camadas anteriores. A DenseNet121 e a DenseNet169 consistem em 121 camadas e 169 camadas, respetivamente. Ambas se destacam em tarefas de visão computacional, nomeadamente na classificação de imagens. As ligações densas promovem uma reutilização eficiente das características, permitindo que a rede capte padrões complexos de maneira eficaz (HUANG et al., 2017).

Na fase 3 ocorre a fusão das arquiteturas 1D e 2D selecionadas. Esta fase considera que os modelos de rede neural 1D e 2D já foram previamente treinados e, portanto, será realizado o treinamento apenas da últimas camadas inteiramente conectadas da arquitetura proposta, que realizam a tarefa de classificação.

Na fase 4 ocorre um treinamento adicional da arquitetura proposta com a finalidade de realizar um ajuste fino dos pesos da rede neural. Para este treinamento são aplicados pesos proporcionais à distribuição das classes no conjunto de treino, tal como o aplicado por Mollahosseini et al. (2019), uma vez que a base de dados não é uniforme.

4.2.3 Definição das arquiteturas de redes neurais convolucionais 1D

Para a fase 1, três arquiteturas de redes neurais profundas são propostas para a classificação de expressões faciais humanas. Essas arquiteturas são baseadas em redes neurais convolucionais de 1 dimensão e variam entre si em termos de profundidade. Apesar destas arquiteturas serem originalmente utilizadas em outros problemas envolvendo dados unidimensionais, suas características construtivas podem ser reutilizadas dada a possibilidade de treinar uma rede neural alterando o conjunto de dados alvo.

4.2.3.1 Arquitetura 1 (CNN 1D - Pequeno porte)

A arquitetura 1 (MACHINE LEARNING, 2020) foi originalmente utilizada no problema de reconhecimento de atividades humanas. Para este problema a rede neural é alimentada com os dados das coordenadas x, y e z fornecidas por um sensor acelerômetro, muito utilizado em análise de movimentos e gestos. Para o contexto do presente trabalho, os dados de entrada da rede possuem dimensão 1× 136 × 1. O dado é composto exclusivamente das coordenadas dos pontos de referência da face.

Na primeira arquitetura, uma rede neural convolutiva com 3 camadas convolucionais e função de ativação ReLU foi utilizada para extração de características. As três camadas convolucionais possuem 64 filtros cada e kernel de tamanho 3 × 1. São conectadas sequencialmente. Em seguida, há uma camada *max pooling* com *kernel* de tamanho 2 × 2 e passo de 2. Posteriormente, há uma camada do tipo *Flatten* e uma camada densa com 100 neurônios. A última camada é constituída por sete neurônios com função de ativação do tipo *softmax*. A Figura 19 apresenta o diagrama em blocos dessa arquitetura.

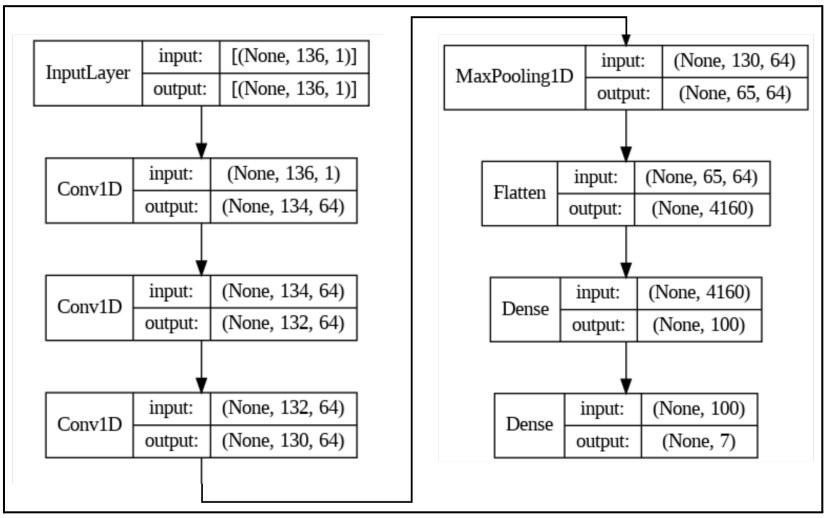


Figura 19 - Diagrama da arquitetura proposta 1.Fonte: (MACHINE LEARNING, 2020)

4.2.3.2 Arquitetura 2 (CNN 1D - Médio porte)

A arquitetura 2, proposta por Azizjon et al. (2020), teve sua aplicação na detecção de intrusão em redes de computadores. Neste problema, as características do dado de entrada estão diretamente relacionadas aos parâmetros do protocolo TCP/IP. Em se tratando de dados numéricos e tabulares, o autor aplicou uma arquitetura de rede neural unidimensional.

Os dados de entrada para treinamento da arquitetura 2 possuem dimensão $1 \times 136 \times 1$, assim como na arquitetura 1. A segunda arquitetura consiste em uma rede neural convolutiva. Nesta arquitetura um bloco de extração de características é formado por 4 camadas convolucionais e função de ativação ReLU, sendo a primeira camada convolutiva composta por 32 kernels e as demais por 16. Após uma camada de *MaxPooling* seguida de uma *Flatten*, são adicionadas 4 camadas inteiramente conectadas, todas com função de ativação ReLu. A última camada é constituída por sete neurônios com função de ativação do tipo *softmax*. A Figura 20 apresenta o diagrama em blocos dessa arquitetura.

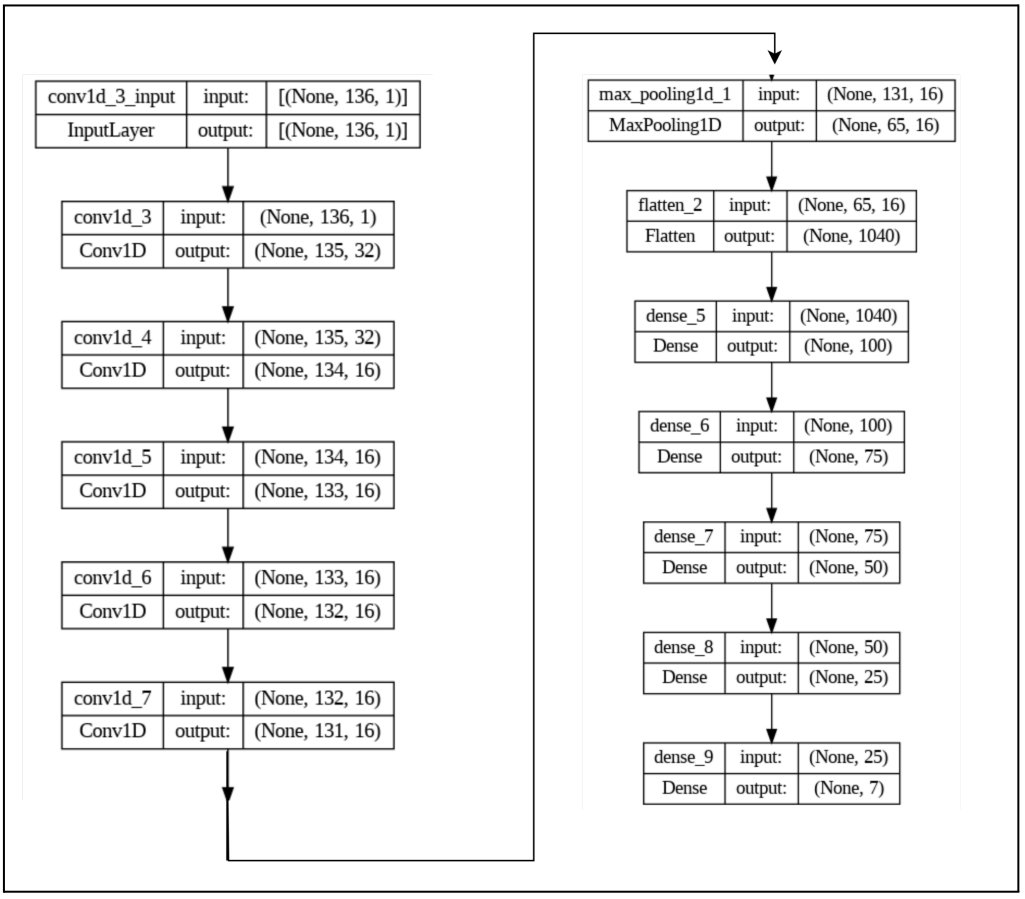


Figura 20 - Diagrama da arquitetura proposta 2. Fonte: Adaptado de Azizjon, 2020

4.2.3.3 Arquitetura 3 (CNN 1D - Grande porte)

A arquitetura 3, proposta por Santana et al. (2021), teve sua aplicação no processo de classificação de arritmias cardíacas baseadas no sinal de ECG. O trabalho manipula o sinal de ECG de maneira unidimensional e bidimensional para a classificação de arritmias, e para isso, propõe e avalia a aplicação de uma rede convolutiva 1D e 2D.

Assim como na arquitetura 1 e 2, os dados de entrada desta arquitetura foram adaptados para a dimensão 136 × 1. A terceira arquitetura consiste em uma rede neural convolutiva 1D. Nesta arquitetura, um bloco de extração de características é formado por uma camada convolutiva com 96 kernels com tamanho de 2×1, seguido de uma camada de *BatchNormalization*. Essa estrutura se repete por 6 vezes. Posteriormente, há uma camada do tipo *Flatten* e uma camada densa com 200 neurônios. A última camada é constituída por sete neurônios com função de ativação do tipo *softmax*. A Figura 21 apresenta o diagrama em blocos dessa arquitetura.

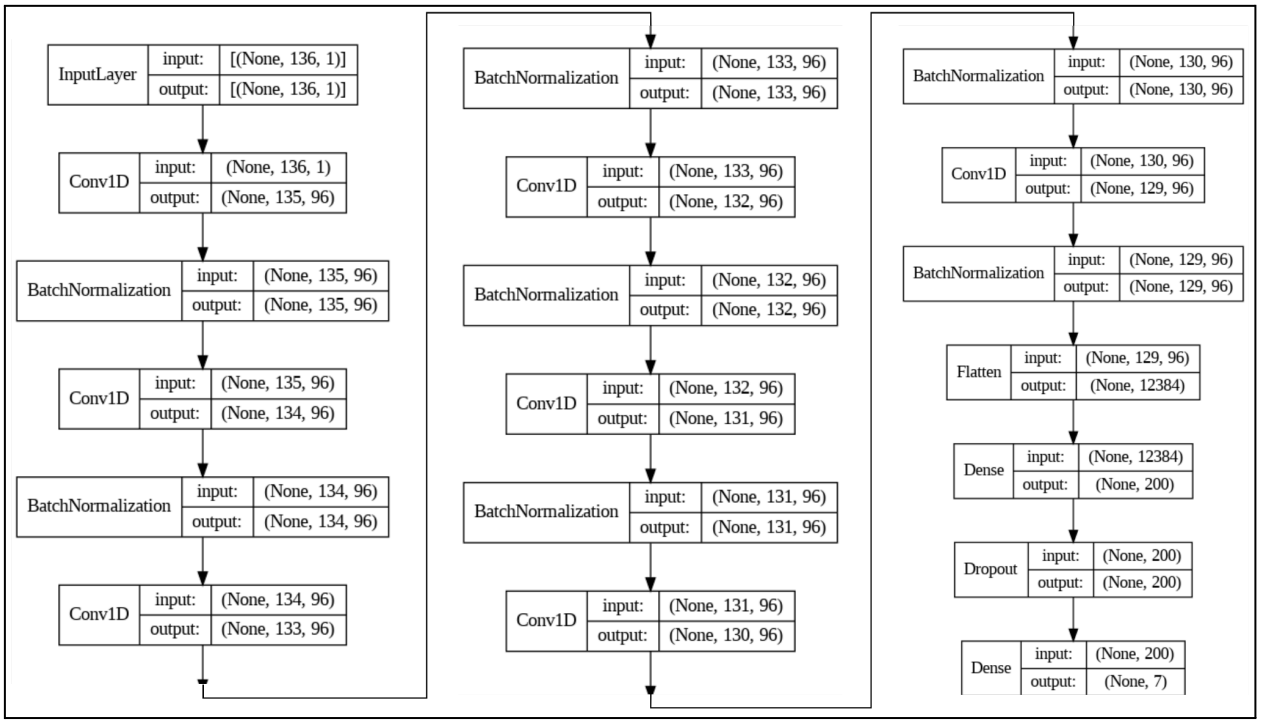


Figura 21 - Diagrama da arquitetura proposta 3. Fonte: (SANTANA et al. , 2021)

4.2.4 Treino das arquiteturas de redes neurais convolucionais 1D

Na etapa de treinamento das redes neurais convolucionais 1D propostas, foram estabelecidos os hiperparâmetros taxa de aprendizagem, tamanho do mini-lote e decaimento da taxa de aprendizagem. Os valores destes hiperparâmetros estão listados na Tabela 4. A escolha dos valores mais adequados foi baseada na análise da curva de aprendizado. O critério de parada adotado foi a taxa de decaimento do erro no conjunto de validação. Para este cenário, foi utilizado o critério de parada antecipada com uma paciência de 5, significando que caso não haja melhoria no erro de convergência da rede em 5 épocas consecutivas, o treinamento é encerrado.

Tabela 4 - Hiperparâmetros de treinamento das redes convolutivas 1D

Hiperparâmetro	**Faixa de valores**
Taxa de aprendizagem	0,0001 - 0,001
Tamanho do mini-lote	32/64
Decaimento da taxa de aprendizagem	0,2
Número de épocas	100 - 300

Para o processo de definição da arquitetura de rede neural 1D a ser utilizada, foram realizados diversos experimentos considerando cada uma das arquiteturas de redes neurais citadas na seção 4.2.3. Durante a fase de treinamento, os hiperparâmetros foram ajustados a partir do comportamento das curvas de aprendizado.

Como critério de avaliação, buscou-se selecionar a arquitetura que obtivesse uma curva de aprendizado estável do ponto de vista da variação da acurácia ao longo do treinamento, e que ao mesmo tempo alcançasse o máximo valor possível desta métrica dentre as demais arquiteturas.

As Figuras 22, 23 e 24 ilustram as curvas de treinamento obtidas quando as arquiteturas foram submetidas às combinações de hiperparâmetros que propiciaram os melhores desempenhos.

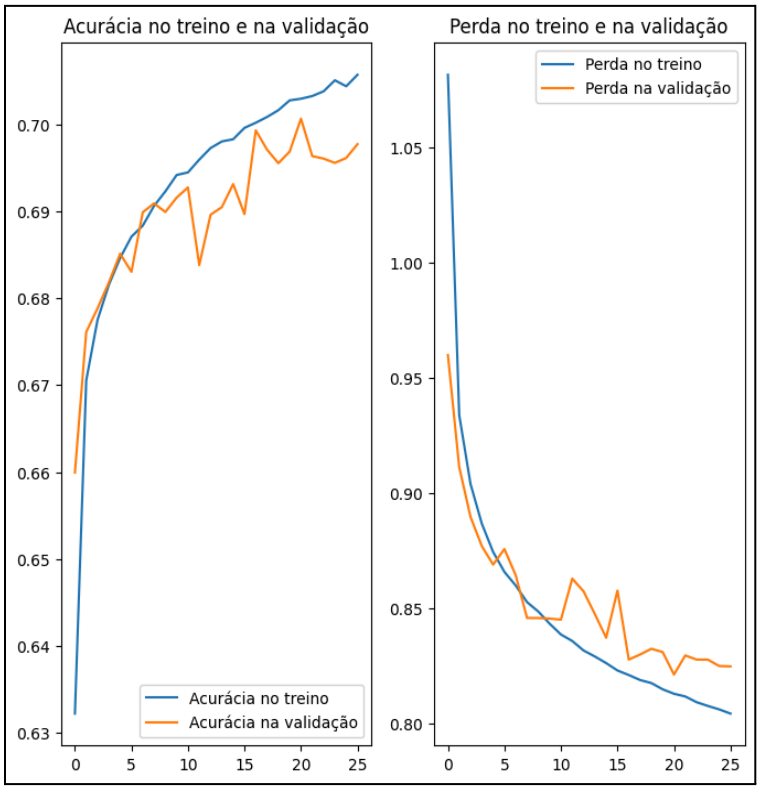


Figura 22 - Curva de treinamento referente à arquitetura de rede neural 1

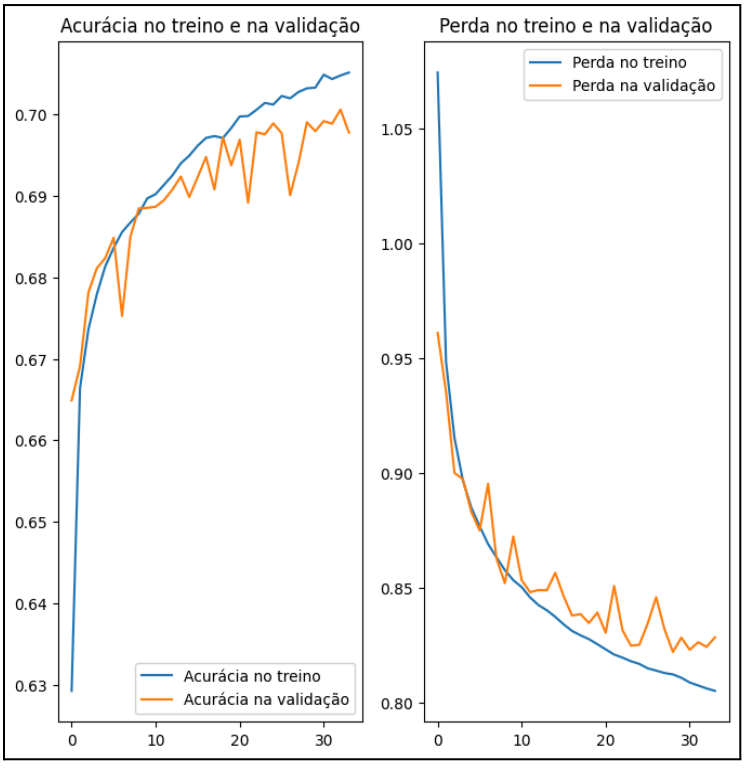


Figura 23 - Curva de treinamento referente à arquitetura de rede neural 2

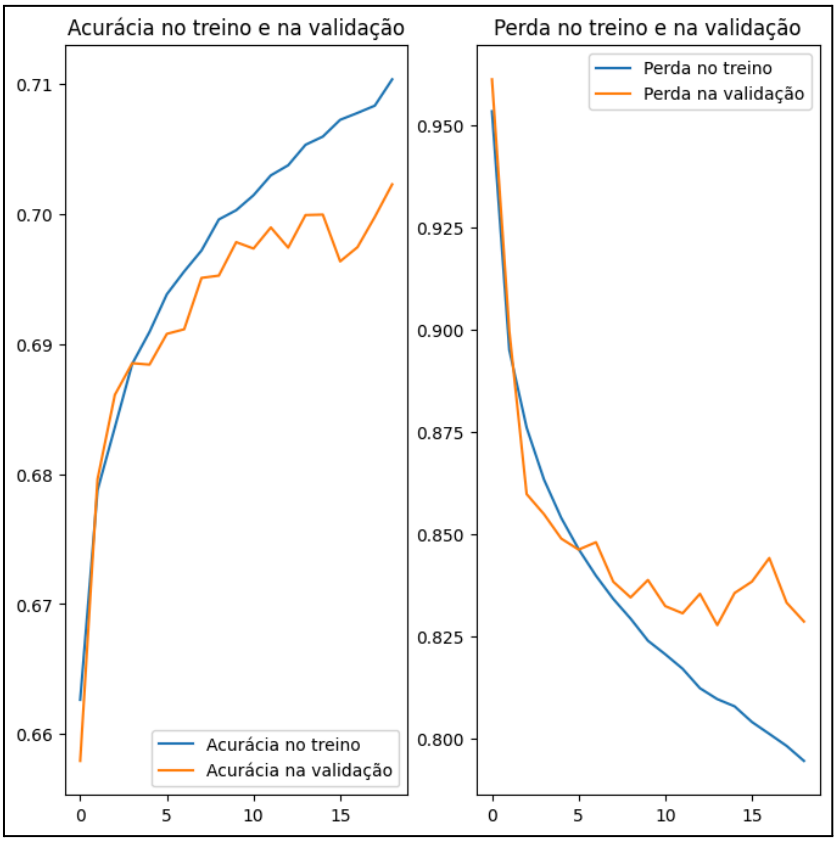


Figura 24 - Curva de treinamento referente à arquitetura de rede neural 3

Além da análise das curvas de aprendizado, os modelos foram submetidos ao conjunto de testes para fins de aferição de desempenho. A Tabela 5 condensa os melhores resultados obtidos com os experimentos realizados para as três arquiteturas. Estes experimentos foram separados em dois estágios. No primeiro estágio, não foram utilizadas estratégias para mitigar os problemas de desbalanceamento da base de dados. No segundo estágio, foram utilizados pesos proporcionais à ocorrência das classes, tal como sugerido por Mollahosseini et al. (2019).

Tabela 5 - Experimentos realizados para seleção de arquitetura

Estratégia de compensação da base desbalanceada	Arquitetura	Acurácia	Épocas	Taxa de Aprendizado
*	#1	39	80	0,0001
*	#2	40,71	30	0,0001
*	**#3**	**48,82**	**20**	**0,0001**
Aplicação de pesos	#1	42,83	30	0,0001
Aplicação de pesos	#2	40,36	6	0,0001
Aplicação de pesos	**#3**	**51,53**	**30**	**0,0001**

A Tabela 5, proveniente de diversas experimentações, indicou uma tendência para a seleção da arquitetura 3, devido ao valor obtido para a acurácia, que se manteve maior entre as demais arquiteturas. A partir destes experimentos também foi possível consolidar uma taxa de aprendizado de 0,0001, ocorrendo uma convergência em até no máximo 30 épocas.

4.2.5 Definição da arquitetura de rede neural convolucional 2D

Nesse trabalho, optou-se pela utilização de arquiteturas DenseNet para a rede convolucional 2D. De acordo com os estudos realizados por Huang et al (2017) a respeito da arquitetura DenseNet, suas conexões densas possibilitam a obtenção de altos valores de acurácia em conjuntos de dados de referência como *ImageNet (*DENG et al., 2009*), CIFAR-10* e *CIFAR-100* (KRIZHEVSKY;HINTON, 2009) quando comparados com a arquitetura ResNet (HE et al., 2016), referência em problemas de classificação de imagens.

Ainda segundo Huang et al (2017), dado que cada camada recebe mapas de características provenientes das camadas anteriores, a rede é capaz de possuir uma quantidade menor de parâmetros treináveis, tornando-se mais compacta que suas concorrentes.

A Tabela 6 detalha as variações da arquitetura DenseNet. As estruturas DenseNet121, DenseNet169, DenseNet201 e DenseNet264 diferem-se entre si quanto ao número de camadas com pesos treináveis.

Tabela 6 - Variações da arquitetura DenseNet

Camadas	Dimensões de saída	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolução	112 x 112	Convolução 7 x 7, *Stride 2*			
Pooling	56 x 56	*MaxPooling 3* x 3, *Stride 2*			
Bloco Denso	56 x 56	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 6$
Camada de transição	56 x 56	Convolução 1 x 1			
	28 x 28	Average Pooling 2 x 2, *Stride 2*			
Bloco Denso	28 x 28	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 12$
Camada de transição	28 x 28	Convolução 1 x 1			
	14 x 14	Average Pooling 2 x 2, *Stride 2*			
Bloco Denso	14 x 14	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 48$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 64$
Camada de transição	14 x 14	Convolução 1 x 1			
	7 x 7	Average Pooling 2 x 2, *Stride 2*			
Bloco Denso	7 x 7	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \ conv \\ 3 \times 3 \ conv \end{bmatrix} \times 48$
Classificação	1 x 1	Average Pooling 7 x 7			
		Camada inteiramente conectada, *Softmax*			

4.2.6 Treino da arquitetura de rede neural convolucional 2D

Na etapa de treinamento das redes neurais convolucionais 2D, foram estabelecidos os hiperparâmetros taxa de aprendizagem, tamanho do mini-lote e decaimento da taxa de aprendizagem. Os valores destes hiperparâmetros estão listados na Tabela 7. A escolha dos valores mais adequados foi baseada na análise da curva de aprendizado. O critério de parada adotado foi a taxa de decaimento do erro no conjunto de validação. Para este cenário, foi utilizado o critério de parada antecipada com uma paciência de 3, significando que caso não haja melhoria no erro de convergência da rede em 3 épocas consecutivas, o treinamento é encerrado.

Tabela 7 - Hiperparâmetros de treinamento das redes convolutivas 1D

Hiperparâmetro	**Faixa de valores**
Taxa de aprendizagem	0,00001 - 0,001
Tamanho do mini-lote	32/64
Decaimento da taxa de aprendizagem	0,2
Número de épocas	100 - 300

Para o processo de definição da arquitetura de rede neural 2D a ser utilizada, foram realizados diversos experimentos considerando as arquiteturas DenseNet121 e DenseNet169 citadas na seção 4.2.5. Durante a fase de treinamento, os hiperparâmetros foram ajustados a partir do comportamento das curvas de aprendizado.

Como critério de avaliação, buscou-se selecionar a arquitetura que obtivesse uma curva de aprendizado estável do ponto de vista da variação da acurácia ao longo do treinamento, e que ao mesmo tempo alcançasse o máximo valor possível desta métrica dentre as demais arquiteturas.

As Figuras 25 e 26 ilustram as curvas de treinamento obtidas quando as arquiteturas foram submetidas às combinações de hiperparâmetros que propiciaram os melhores desempenhos.

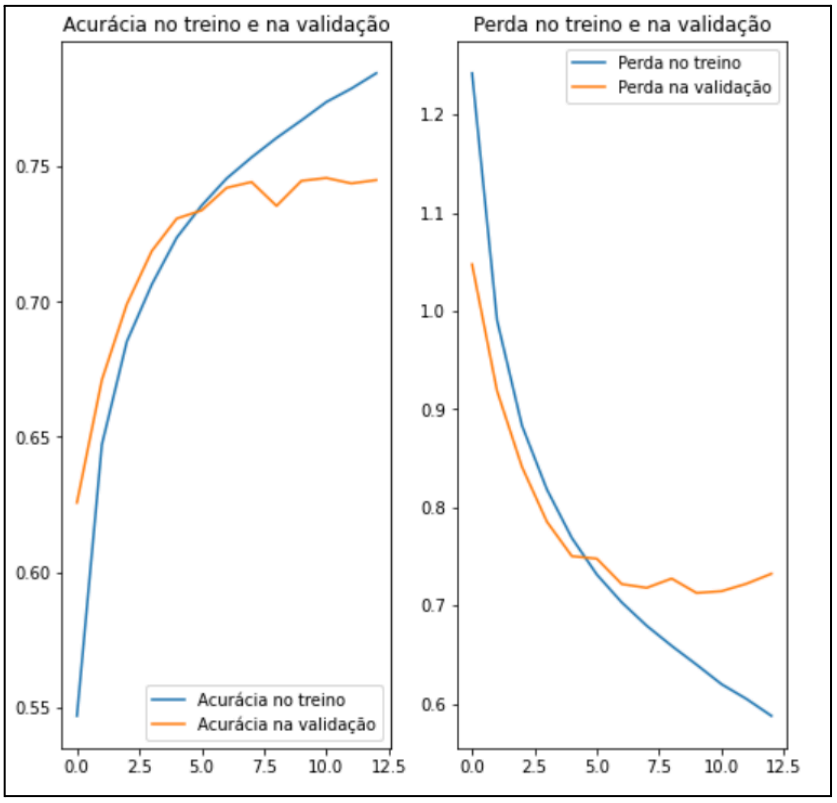


Figura 25 - Curva de treinamento referente à arquitetura de rede neural DenseNet121

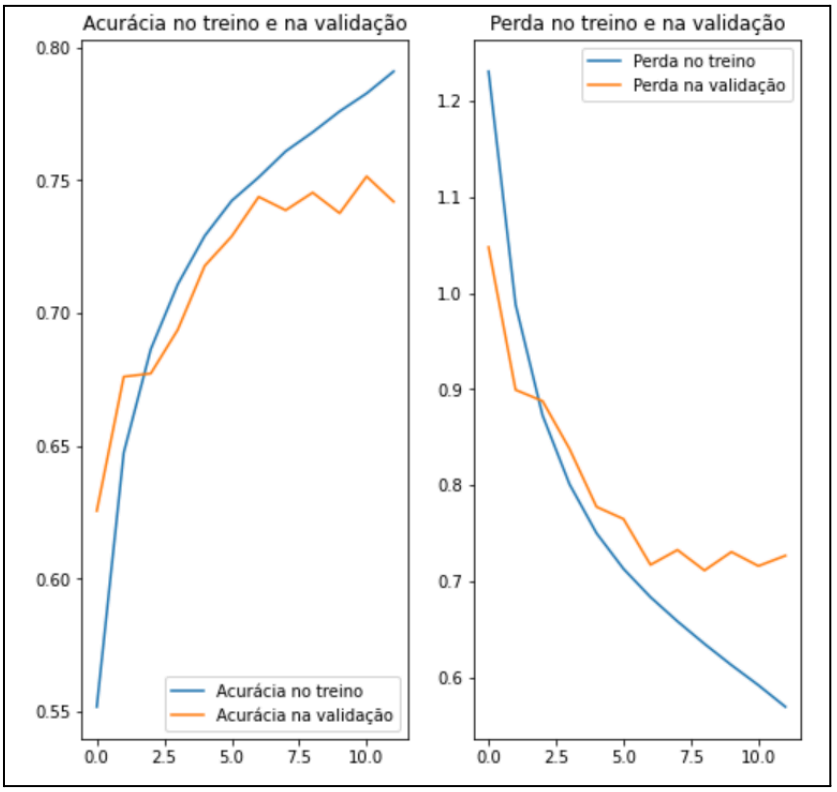


Figura 26 - Curva de treinamento referente à arquitetura de rede neural DenseNet169

Além da análise das curvas de aprendizado, os modelos foram submetidos ao conjunto de testes para fins de aferição de desempenho. A Tabela 8 condensa os experimentos realizados para as duas arquiteturas.

Tabela 8 - Experimentos realizados para seleção de arquitetura de rede 2D

Arquitetura	Acurácia	Épocas	LR
DenseNet121	**50,83**	**13**	**0,00001**
DenseNet169	50,14	11	0,00001

A Tabela 8 indica uma tendência para a seleção da arquitetura DenseNet121, pelo fato da acurácia obtida ter sido levemente superior ao valor da acurácia da rede DenseNet169.

4.2.7 Arquitetura de rede neural proposta para a classificação de expressões faciais

A arquitetura de rede neural proposta neste trabalho foi obtida a partir da fusão entre as melhores arquiteturas de rede neural 1D e 2D definidas na fase de experimentação e testes. Com base nas Tabelas 5 e 8, foram selecionadas a arquitetura 3 para o caso da rede convolutiva 1D e a arquitetura DenseNet121 para o caso da rede 2D.

De acordo com Huang et al. (2020), o uso de modelos de aprendizado profundo multimodais traz grandes vantagens por possibilitar o uso de dados contextuais que não sejam apenas a informação dos pixels de uma imagem. Desta maneira, o paradigma de fusão atua em tarefas complexas que não podem ser facilmente resolvidas a partir de uma única modalidade de dados.

Ainda de acordo com Huang et al. (2020), existem 3 principais estratégias de fusão, a citar:

- Fusão Inicial: Combina todas as modalidades de entrada em uma única representação antes do processamento.
- Fusão Conjunta: Processa as modalidades de entrada separadamente, mas simultaneamente, compartilhando informações durante o processo.
- Fusão Tardia: Processa as modalidades de entrada separadamente e combina suas saídas em um estágio posterior do processo.

Neste trabalho a modalidade de fusão tardia será utilizada pelo fato de proporcionar um processamento separado para cada modalidade de entrada, isto é, a imagem da face e os pontos de referência da face. A Figura 27 ilustra as 3 diferentes estratégias de fusão.

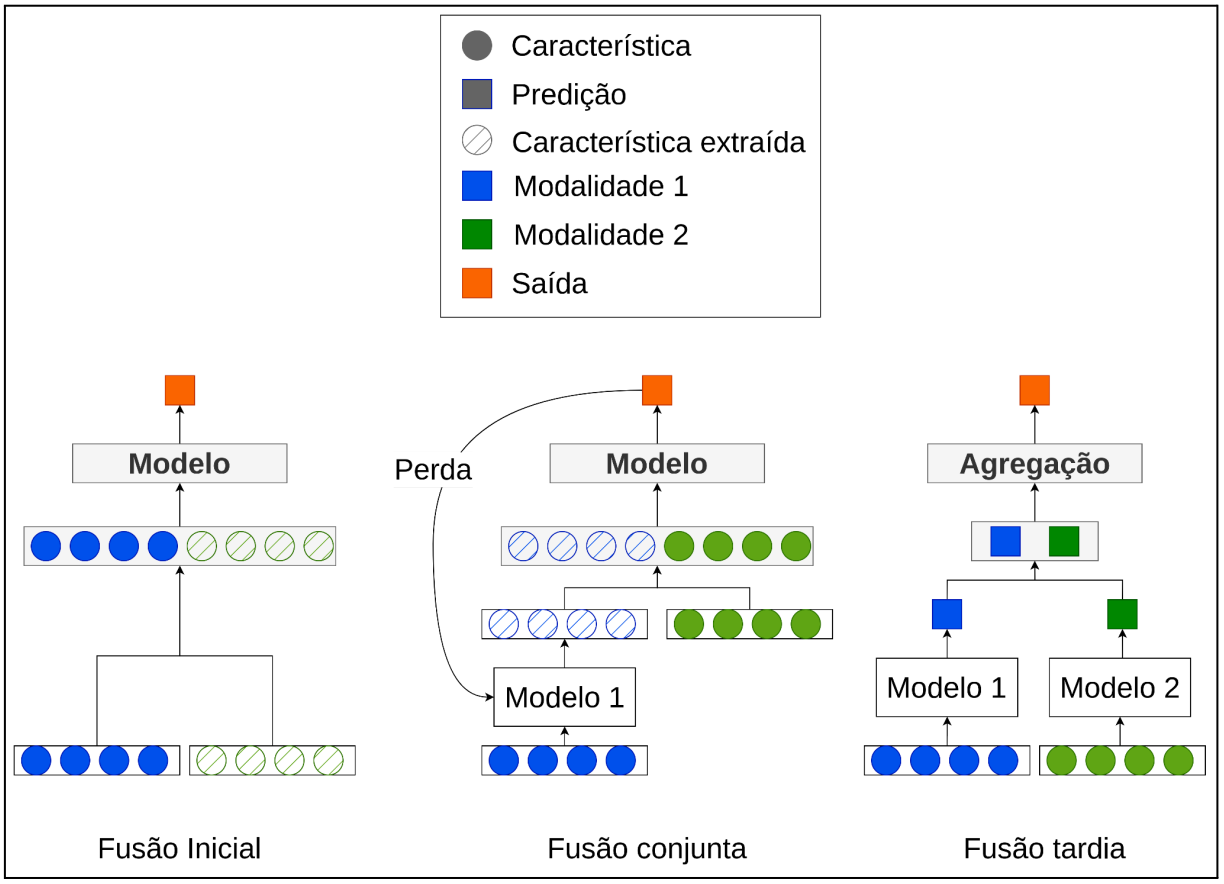


Figura 27 - Estratégias de fusão. Fonte: Adaptado de Huang et al., 2020

Considerando a estratégia de fusão tardia, as penúltimas camadas inteiramente conectadas das arquiteturas 1D e 2D foram concatenadas. Desta maneira, a tomada de decisão da arquitetura proposta levará em consideração as características extraídas por ambos os ramos da rede.

4.2.8 Treinamento da arquitetura de rede neural proposta

Na etapa de treinamento da rede neural proposta, foram estabelecidos os hiperparâmetros taxa de aprendizagem, tamanho do mini-lote e decaimento da taxa de aprendizagem. Os valores destes hiperparâmetros estão listados na Tabela 9. A escolha dos valores mais adequados foi baseada na análise da curva de aprendizado e nos experimentos prévios realizados com as arquiteturas 1D e 2D isoladamente. O critério de parada adotado foi

a taxa de decaimento do erro no conjunto de validação. Para este cenário, foi utilizado o critério de parada antecipada com uma paciência de 2, significando que caso não haja convergência por 2 épocas consecutivas, o treinamento é encerrado.

Tabela 9 - Hiperparâmetros de treinamento da rede proposta

Hiperparâmetro	Valores utilizados
Taxa de aprendizagem	0,001
Tamanho do mini-lote	32
Decaimento da taxa de aprendizagem	0,2
Número de épocas	30

O treinamento da arquitetura proposta foi dividido em duas etapas. A primeira etapa consistiu no treinamento da rede logo após o processo de fusão. Nesta primeira etapa buscou-se verificar a convergência da nova arquitetura gerada. Na segunda etapa, o treinamento foi realizado com a aplicação de pesos proporcionais ao número de amostras de cada classe no conjunto de treino. Estes pesos foram calculados de acordo com a estratégia sugerida pela biblioteca Tensorflow (ABADI et al., 2015) e pelo estudo de caso proposto por Toward Data Science (2021). Com base nestas implementações, os pesos podem ser calculados de acordo com a Equação (6). Nesta definição, o peso W de uma classe i (W_i), é obtido através da relação entre o número total de amostras da base de treinamento N, o número de classes n_c e o número de elementos n_i da classe i.

$$W_i = \frac{N}{n_C \times n_i} \tag{6}$$

A partir da aplicação da Equação (6), foram obtidos os pesos proporcionais à cada classe. Os mesmos tornaram-se hiperparâmetros do treinamento. Os valores calculados encontram-se na Tabela 10.

Tabela 10 - Pesos proporcionais ao número de exemplos de cada classe

Classe	Peso Calculado
Neutro	0,54
Feliz	0,3
Triste	1,59
Surpresa	2,87
Medo	6,35
Nojo	10,66
Raiva	1,63

4.3 AVALIAÇÃO DO MODELO TREINADO

4.3.1 Métricas de desempenho

Para a tarefa de classificação de expressões faciais humanas foram utilizadas as seguintes métricas: acurácia (ACC), sensibilidade (SEN), especificidade (SPE), precisão (PRE), F1-Score e curva ROC.

Este processo de avaliação de desempenho consiste na maneira com que o modelo ou arquitetura de rede classifica corretamente uma imagem. Para duas classes, esta classificação pode ser representada por uma matriz de confusão tal como ilustrada na Figura 28. Essa matriz relaciona o resultado esperado e o resultado predito pelo modelo. É esperado que estes dois valores sejam o mais próximo possível (PATTERSON;GIBSON, 2017).

Matriz de confusão		**Classe predita**	
		Positiva	**Negativa**
Classe original	**Positiva**	VP	FN
	Negativa	FP	VN

Figura 28 - Matriz de confusão. Fonte: Adaptado de (FERRARI E SILVA, 2017)

Por definição, tem-se que:

- VP (Verdadeiro Positivo): O modelo classifica uma imagem como pertencente a uma classe e a decisão está correta.
- VN (Verdadeiro Negativo): O modelo classifica uma imagem como não pertencente a uma classe e a decisão está correta.
- FN (Falso Negativo): O modelo classifica uma imagem como não pertencente a uma classe e a decisão está incorreta.
- FP (Falso Positivo): O modelo classifica uma imagem como pertencente a uma classe e a decisão está incorreta

Em decorrência destas definições, é possível estabelecer métricas de desempenho que levam em conta relações entre estes índices de erros ou acertos. De maneira genérica, a acurácia avalia o percentual de acertos na classificação. A sensibilidade avalia a capacidade de classificar com sucesso resultados positivos. A especificidade, por outro lado, avalia a capacidade do modelo classificar com sucesso resultados negativos. A precisão revela quantas das classes previstas como verdadeiras estão rotuladas corretamente. O score F1 é a média

harmônica entre a sensibilidade e a precisão. Estas métricas estão matematicamente definidas no Quadro 2.

Quadro 2 - Métricas de desempenho

Acurácia	$ACC = \frac{VP + VN}{VP + FP + VN + FN}$	(7)
Sensibilidade	$SEN = \frac{VP}{VP + FN}$	(8)
Especificidade	$SPE = \frac{VN}{VN + FP}$	(9)
Precisão	$PRE = \frac{VP}{VP + FP}$	(10)
Score F1	$F1 = 2 \cdot \frac{SEN \cdot PRE}{SEN + PRE}$	(11)

Estas métricas são úteis para a comparação com demais trabalhos, sendo a acurácia a métrica de partida para a maioria das referências.

A curva ROC (*Receiver Operating Characteristic*), mostrada na Figura 29, é uma ferramenta gráfica amplamente utilizada na avaliação de modelos de classificação, especialmente em campos como aprendizado de máquina e estatística (FAWCETT, 2006). Ela representa a relação entre a fração de verdadeiros positivos (True Positive Fraction →TPF, sensibilidade) e a fração de falsos positivos (False Positive Fraction → FPF = 1-especificidade) para diferentes pontos de corte em um classificador binário.

Uma curva ROC ideal se aproxima do canto superior esquerdo do gráfico, representando um modelo que maximiza tanto a sensibilidade quanto a especificidade. A área sob a curva (AUC) é frequentemente usada como uma medida resumida do desempenho do modelo, onde um valor próximo de 1 indica um desempenho excelente.

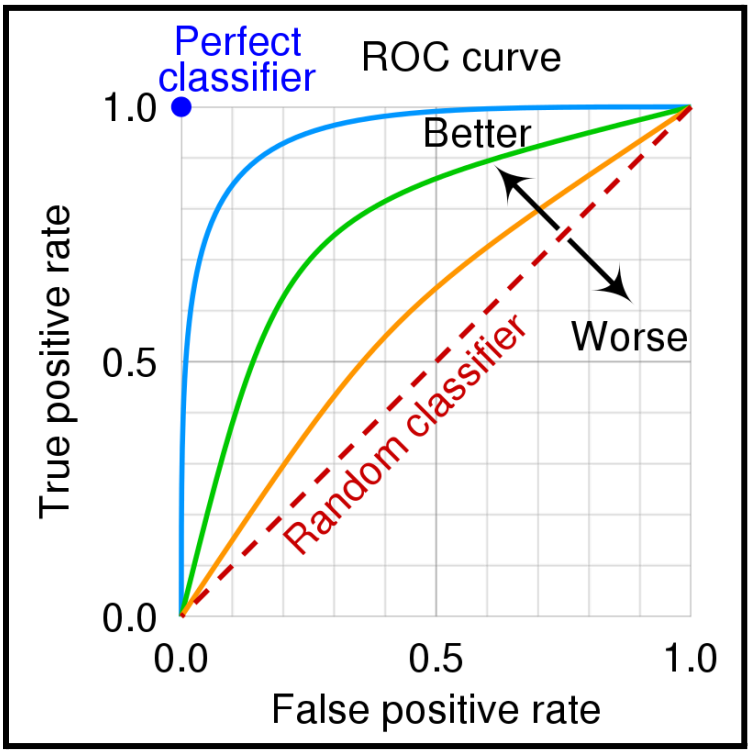


Figura 29 - Curva ROC. Fonte: (THE ROC CURVE, 2023)

Na Figura 29 a linha tracejada representa um classificador que realiza classificações de maneira aleatória, com AUC de 0,5. Abaixo dessa linha um classificador é considerado ruim. Acima da linha e o quanto mais próximo do canto superior direito, um classificador é considerado de boa qualidade.

5 RESULTADOS E DISCUSSÕES

Para avaliar se a abordagem de fusão de arquiteturas de rede 1D e 2D melhora a classificação das expressões faciais humanas, foi necessário comparar os modelos gerados a partir deste método. Neste capítulo, são apresentados os desempenhos das redes 1D e 2D isoladamente, e em seguida, o desempenho do modelo resultante da fusão destas duas arquiteturas. Por fim, é realizada uma comparação com demais trabalhos que lidam com o problema de classificação de expressões faciais humanas.

5.1 RESULTADO DOS MODELOS

A Tabela 11 exibe as métricas obtidas na fase de experimentação com as arquiteturas de redes neurais. Os resultados obtidos foram colhidos a partir do conjunto de testes da base de dados *AffectNet*.

Tabela 11 - Métricas obtidas a partir do conjunto de testes

Modelo	**Métricas**			**Acurácia Global (%)**
	Precisão	**Revocação**	**Score F1**	
1D CNN	51,36	50,94	51,15	51,53
2D CNN – DenseNet121	59,9	50,82	54,98	50,83
2D CNN – DenseNet169	60,1	50,14	54,67	50,14
Fusão das arquiteturas (1D CNN + DenseNet121)	61	60	60	60,4

O modelo de rede neural 1D e as redes neurais convolutivas atingiram valores de acurácia similares. A precisão das redes convolutivas 2D se mostrou expressiva, e isso ressalta o potencial dessas redes em classificar por meio de aspectos globais das imagens.

Por outro lado, a rede neural 1D selecionada, mesmo possuindo uma quantidade inferior de parâmetros que as demais redes, atingiu resultados de acurácia similares às das redes 2D apenas tendo como dados de entrada os pontos de referência da face. Esse aspecto reforça a capacidade da rede 1D atuar de maneira complementar durante o processo de classificação.

A rede DenseNet121 obteve performance discretamente superior à rede DenseNet169. Esta diferença sutil pode ser devida ao tamanho superior da rede DenseNet169 em termos de quantidade de camadas, e isto pode provocar o efeito de *overfitting* durante o treinamento do modelo.

A arquitetura de rede proposta, resultante da fusão entre as redes que obtiveram melhores resultados na etapa de experimentação, foi submetida ao mesmo conjunto de testes.

A acurácia de 60,40% supera alguns resultados de pesquisas que endereçaram o problema de reconhecimento de expressões faciais humanas considerando 7 classes e

utilizando a mesma base de dados utilizada nesse trabalho . A Tabela 12 elucida as bases de comparação.

Tabela 12 - Análise comparativa com demais trabalhos

Modelo	**Acurácia (%)**
Li et al. (2019)	58,78
Huang et al. (2023)	56,54
Fusão das arquiteturas (1D CNN + DenseNet121)	**60,4**

5.2 RESULTADO DOS EXPERIMENTOS

O gráfico da Figura 30 dispõe os melhores resultados obtidos na fase de experimentação. A análise do gráfico permite notar que o uso de pesos proporcionais às ocorrências das classes durante o treinamento do modelo contribui para um ganho na métrica acurácia.

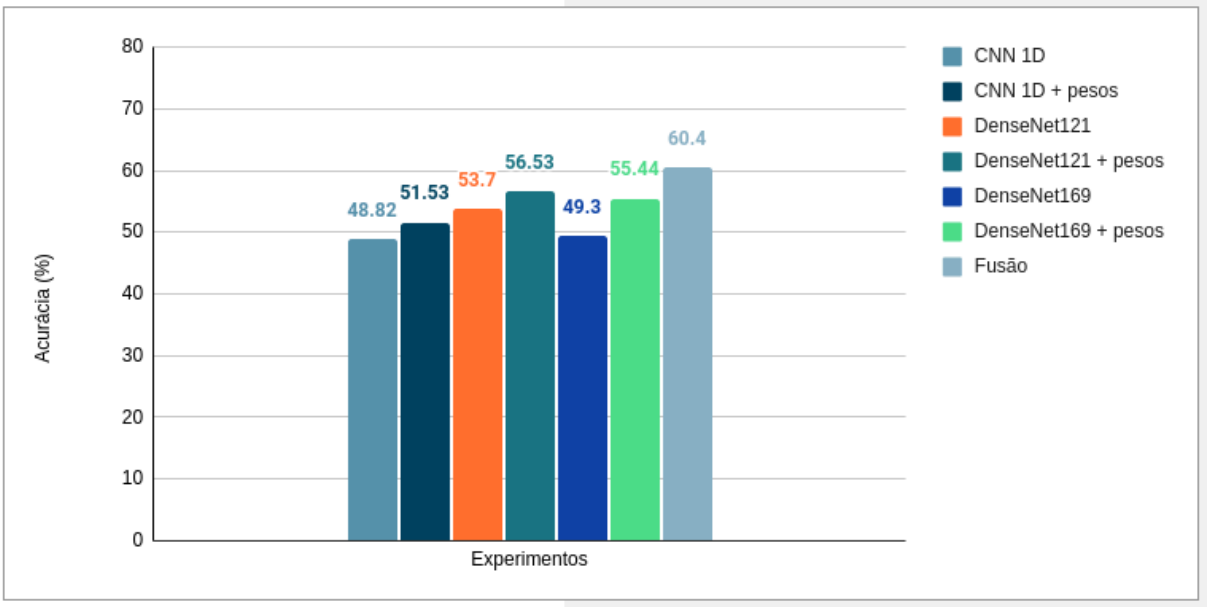


Figura 30 - Melhores resultados obtidos nos experimentos realizados

A técnica de fusão não apenas proporcionou um ganho no desempenho global, como também evidenciou o potencial do uso de modelos multimodais, isto é, com entradas de dados de domínios distintos.

5.3 MÉTRICAS DE DESEMPENHO

Após a avaliação do modelo de classificação, foram obtidas várias métricas de desempenho para analisar sua eficácia e generalização. Entre essas métricas, destacam-se a curva ROC (*Receiver Operating Characteristic*), que oferece uma visão abrangente da relação entre TPF e FPF em diferentes pontos de corte do classificador, e a área sob a curva (AUC), que resume a capacidade discriminativa do modelo.

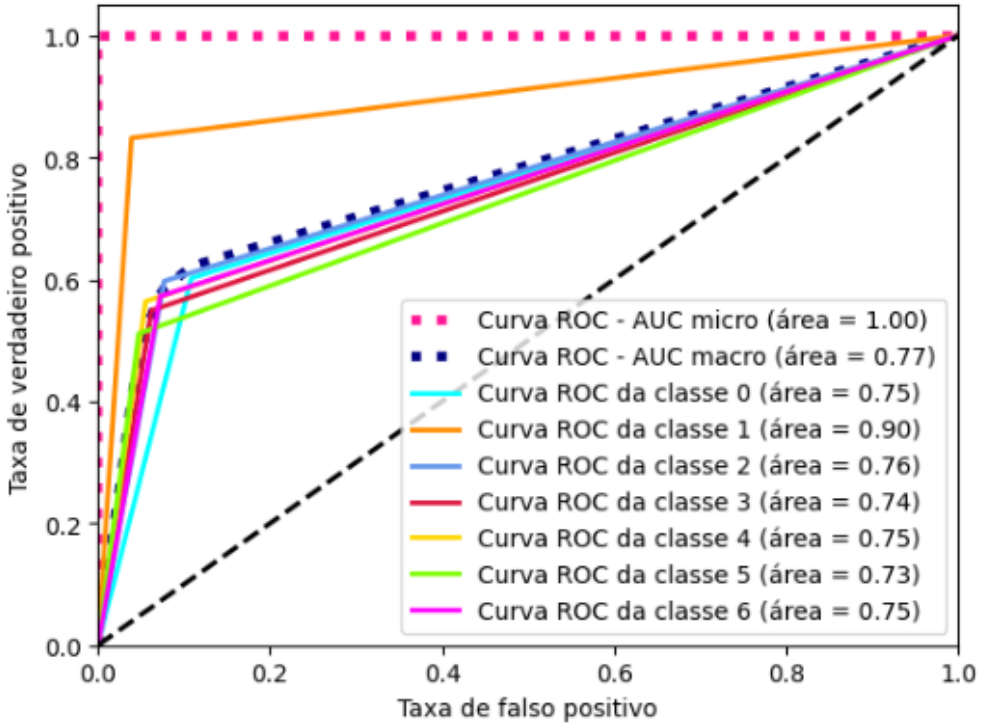


Figura 31 - Curva ROC

Globalmente, esses valores de AUC sugerem que o modelo tem um desempenho variado, com algumas classes sendo mais bem classificadas do que outras. A classe com o maior AUC (0,90) indica um modelo particularmente forte na distinção entre instâncias positivas e negativas dessa classe, enquanto os valores em torno de 0,75 indicam um desempenho consistente e razoável em várias outras classes.

Para fins de avaliação por classe, foi elaborado um relatório considerando as métricas individuais, as quais podem ser visualizadas na Tabela 13.

Tabela 13 - Métricas de desempenho globais e locais

Classe	Precisão	Sensibilidade	Score F1
Neutro	0,48	0,60	0,53
Feliz	0,78	0,83	0,81
Triste	0,56	0,60	0,58
Surpresa	0,60	0,55	0,57
Medo	0,63	0,56	0,59
Nojo	0,64	0,51	0,57
Raiva	0,57	0,57	0,57
Acurácia			0,60
Média	0,61	0,60	0,60

Além disso, a matriz de confusão, ilustrada na Figura 32, forneceu uma representação tabular das classificações corretas e incorretas do modelo, permitindo uma análise detalhada das taxas de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. Essas métricas são fundamentais para avaliar a precisão, robustez e utilidade do modelo em diferentes cenários e contextos de aplicação.

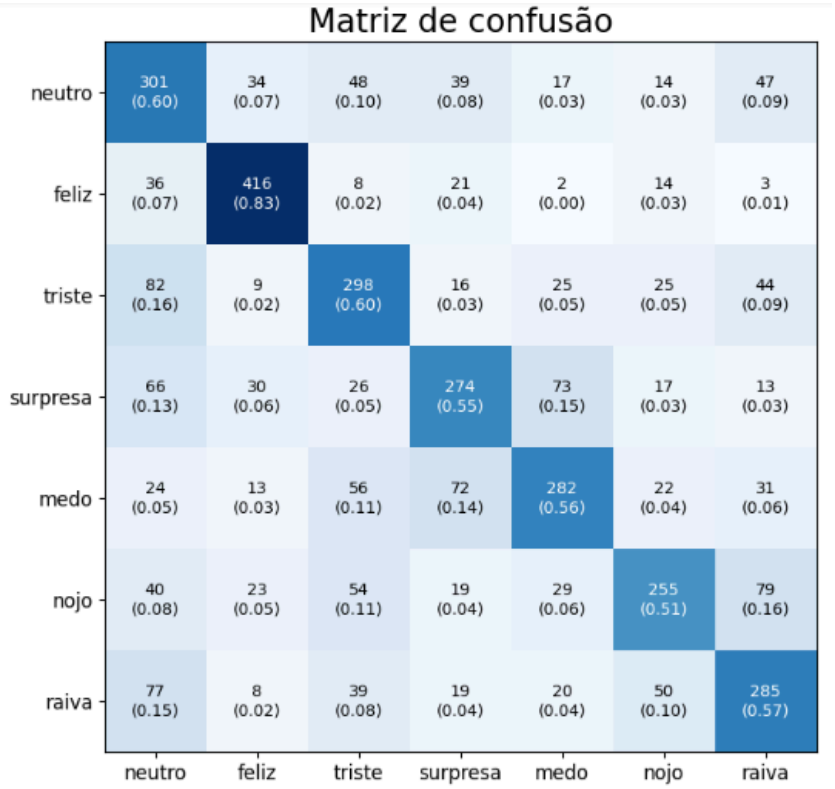


Figura 32 - Matriz de confusão

6 CONCLUSÕES

A pesquisa bibliográfica permitiu visualizar o estado da arte no que se refere à classificação de expressões faciais humanas por meio de imagens da face. Partindo destas informações, o presente trabalho buscou propor uma arquitetura de rede neural capaz de proporcionar bons resultados, mantendo a premissa de ser computacionalmente viável e apta a ser executada em computadores de borda.

Para que estes objetivos fossem atingidos, a metodologia empregada estabeleceu a fusão de uma rede neural convolutiva 1D, tendo como entrada os pontos de referência da face e uma rede neural convolutiva 2D. Após algumas experimentações, a rede convolutiva 2D escolhida foi a rede DenseNet121, uma das mais simplificadas da família de redes DenseNet.

A base de dados AffectNet foi utilizada tanto para fins de treinamento do modelo proposto como para comparação dos resultados obtidos entre si e com resultados da literatura. Apesar da base de dados AffectNet ser uma das mais volumosas em termos de dados, também possui um grande desbalanceamento, dispondo de mais amostras de determinada classe em detrimento de outras. Este problema foi parcialmente contornado a partir do treinamento ponderado, isto é, foi possível atribuir pesos proporcionais à frequência de ocorrência de cada classe na base de dados. Desta maneira, algumas classes são mais fortemente atingidas pela correção de erro no algoritmo de *backpropagation.*

Os experimentos realizados permitiram concluir que os pontos de referência da face por si só não asseguram uma boa classificação, entretanto, a conjunção com outros mecanismos de tomada de decisão, como as redes convolutivas, propiciam melhor desempenho.

Os resultados obtidos em termos de acurácia 60,40% , precisão de 60% e AUC média de 0,77 permitiram estabelecer a arquitetura proposta como um ponto de partida para futuras melhorias, dado que se tornou competitiva frente aos trabalhos de Li et al. (2019) e Huang et al. (2023).

REFERÊNCIAS

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; GOODFELLOW, I.; HARP, A.; IRVING, G.; ISARD, M.; JIA, Y.; JOZEFOWICZ, R.; KAISER, Ł.; KUDLUR, M., ... ZHENG, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Retrieved from https://www.tensorflow.org/

AZIZJON, M.; JUMABEK, A.; KIM, W. (2020). "1D CNN based network intrusion detection with normalization on imbalanced data," 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 218-224, doi: 10.1109/ICAIIC48513.2020.9064976.

CHEN, Q.; JING, X.; ZHANG, F.;MU, J.(2022). "Facial Expression Recognition Based on A Lightweight CNN Model," 2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Bilbao, Spain, 2022, pp. 1-5, doi: 10.1109/BMSB55706.2022.9828739.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., & FEI-FEI, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).

DHALL, A.; GOECKE, R.; JOSHI, J.; SIKKA, K.; GEDEON, T. (2015). Static facial expression in the wild: A novel emotion recognition dataset. IEEE Transactions on Affective Computing, 6(2), 144-158.

DHALL, A.; GOECKE, R.; JOSHI, J.; GEDEON, T. (2019). Emotion recognition in the wild challenge 2019: Baseline, data and protocol. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 0-0).

EKMAN, P.; FRIESEN, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 1971, 17 (2), 124-129.

EKMAN, P. (2011). A linguagem das emoções, Lua de Papel.

ELFENBEIN, H. A.; AMBADY, N.; MANDAL, M. K.; HARIZUKA, S. (2002) Cross-Cultural patterns in emotion recognition: Highlighting design and analytical techniques. Emotion, 2(1), 75–84. DOI: 10.1037//1528-3542.2.1.75.

FAWCETT, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. Robin, X., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77.

FERRARI, D. G.; DE CASTRO SILVA, L. N. Introdução à mineração de dados. [s.l.] Saraiva Educação S.A., 2017.

GHOSH, S.; LAHA, A.; MURSHED, N. A. (2020). Facial landmarks: A comprehensive survey, applications, performance evaluation, and open issues. IEEE Transactions on Multimedia, 22(8), 2066-2094.

GOODFELLOW, I.; BENGIO,Y.; COURVILLE,A. DeepLearning.[S.l.]:MITPress, 2016.Disponível em: <http://www.deeplearningbook.org>

HE, K.; ZHANG, X.; REN, S.; SUN, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

HUANG, G.; LIU, Z.; VAN DER MAATEN, L.; WEINBERGER, K. Q. (2017). Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

HUANG, SC., PAREEK, A., SEYYEDI, S. et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digit. Med. 3, 136 (2020). https://doi.org/10.1038/s41746-020-00341-z

HUANG, Z.; CHIANG, C.; CHEN, J.; CHEN, Y.; CHUNG, H.; CAI, Y.; HSU, H. (2023). A study on computer vision for facial emotion recognition. Sci Rep. 2023 May 24;13(1):8425. doi: 10.1038/s41598-023-35446-4. PMID: 37225755; PMCID: PMC10209161.

JACK, R. E.; GARROD, O. G. B.; SCHYNS, P. G. (2014). Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time. Current Biology, 24(2), 187-192.

KRIZHEVSKY, A.;HINTON, G. (2009). Learning Multiple Layers of Features from Tiny Images (Technical Report). University of Toronto. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

LI, S.; DENG, W.; DU, Y. (2017). Deep Facial Expression Recognition: A Survey. arXiv preprint arXiv:1703.08396.

LI, S.; WANG, W.; LI, L.; ZHANG, H. (2020). LFA-Net: Low-light facial expression recognition via attention-based adversarial learning. IEEE Transactions on Information Forensics and Security, 15, 2856-2867.

LI, Y.; ZENG, J. S.; SHAN, S.; CHEN, X. (2019). Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism. IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439-2450, May 2019, doi: 10.1109/TIP.2018.2886767.

LIU, W.; LIU, X.; YANG, J.; ZHANG, Y.; WANG, L.; WANG, Y. (2021). A novel convolutional neural network for facial expression recognition. Information Sciences, 569, 413-427.

LUCEY, P.; COHN, J.F.; KANADE, T.; SARAGIH, J.; AMBADAR, Z.; MATTHEWS, I. (2010). The extended cohn-kanade dataset(ck+): A complete dataset for action unit and emotion-specified expression. In:IEEE.2010 ieee computer society conference on computer vision and pattern recognition-workshops. [S.l.], 2010.p.94–101

MACHINE LEARNING. 1D Convolutional Neural Network Models for Human Activity Recognition. Disponível em: https://machinelearningmastery.com/cnn-models-for-human-activity-recognition-time-series-classification/

MATSUMOTO, D.; HWANG, H. C. (2019). Assessing the accuracy of detecting truths and lies in real-time: A meta-analysis of highly controlled laboratory studies. Journal of Applied Research in Memory and Cognition, 8(4), 403-412.

MEDIUM. Uma introdução as redes neurais convolucionais utilizando o Keras. Disponível em: <https://medium.com/data-hackers/uma-introdu%C3%A7%C3%A3o-as-redes-neurais-convolucionais-utilizando-o-keras-41ee8dcc033e>.

MOLLAHOSSEINI, A.; HASANI, B.;MAHOOR, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

PATTERSON, J.; GIBSON, A. Deep Learning: A Practitioner's Approach. O'Reilly Media, 2017. ISBN 9781491914236. Disponível em: <https://www.safaribooksonline.com>.

RIBEIRO, W. (2020). Curva ROC e AUC em Machine Learning. Disponível em: https://cienciaenegocios.com/curva-roc-e-auc-em-machine-learning/

SANCHÉZ, R.; LÓPEZ, F.; LÓPEZ, M. T.; FERNANDÉZ, A. (2022). One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity. Elsevier

SANTANA, J. R. G.; COSTA, M. G. F.; COSTA FILHO C. F. F. (2021). A New Approach to Classify Cardiac Arrythmias Using 2D Convolutional Neural Networks. PubMed

SOYLEMEZ, O.F.; ERGEN, B. (2022). Facial Landmark Based Region of Interest Localization for Deep Facial Expression Recognition. Tehnicki vjesnik - Technical Gazette.

THE ROC CURVE. The ROC Curve: Application and Interpretation in the Health Context. Disponível em: <https://medium.com/@evertongomede/the-roc-curve-application-and-interpretation-in-the-health-context-cc06af05a6ec>. Acesso em: 10 mai. 2024.

TOWARD DATA SCIENCE. Dealing with Imbalanced Data in TensorFlow: Class Weights. Disponível em: <https://towardsdatascience.com/dealing-with-imbalanced-data-in-tensorflow-class-weights-60f876911f99> . Acesso em: 5 maio. 2024.

TRACY, J. L.; MATSUMOTO, D. (2008). The Spontaneous Expression of Pride and Shame: Evidence for Biologically Innate Nonverbal Displays. Proceedings of the National Academy of Sciences, 105(33), 11655-11660.

VALSTAR, M. F.; MARTINEZ, B.; BINEFA, X.; PANTIC, M. (2016). Facial expression analysis. In Handbook of Face Recognition (2nd ed., pp. 433-459). Springer.

VERMA, M.; KOBORI, H.; NAKASHIMA,Y.; TAKEMURA, N.; NAGAHARA, H. (2019). "Facial Expression Recognition with Skip-Connection to Leverage Low-Level Features," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 51-55, doi: 10.1109/ICIP.2019.8803396.

WADHAWAN, R.; GANDHI, T. K. "Landmark-Aware and Part-Based Ensemble Transfer Learning Network for Static Facial Expression Recognition from Images," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 2, pp. 349-361, April 2023, doi: 10.1109/TAI.2022.3172272.

ZHANG, Z.; SONG, Y.; QI, H.; LIU, Y.; ZHANG, J.; LIU, Z. (2018). Deep learning for 3D face recognition: A survey. Pattern Recognition, 83, 34-58.

APÊNDICE A – ARTIGO

Neste apêndice é apresentada a cópia do artigo originado deste trabalho. O artigo intitulado: *"Enhancing Emotion Recognition: A Dual-Input Model for Facial Expression Recognition Using Images and Facial Landmarks",* foi submetido e aceito para apresentação na 46ª Conferência Internacional Anual da Sociedade IEEE de Engenharia em Medicina e Biologia, a realizar em Orlando, Flórida, EUA.

Enhancing Emotion Recognition: A Dual-Input Model for Facial Expression Recognition Using Images and Facial Landmarks

Willian Guerreiro Colares
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
0000-0003-0960-2668

Marly. G. F. Costa
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
0000-0002-6839-1402

Cícero F. F . Costa Filho
Center for Research and Development in Electronics and Information Technology
Federal University of Amazonas
Manaus, Brazil
0000-0003-3325-5715

***Abstract*— Human facial expressions play a fundamental role in nonverbal communication and the conveyance of emotions. Conceptually, facial expressions can be deduced from the arrangement of facial muscles. As a subjective assessment, constructing a database for facial expression recognition becomes a challenge due to the high risk of bias arising from unbalanced or inaccurate data. On the other hand, advances in image processing techniques and deep learning have boosted the accuracy and effectiveness of algorithms for facial expression recognition. In this work, aiming to improve the automatic facial expression recognition, we present the fusion of two neural network architectures. The first one comprises a one-dimensional convolutional neural network (1D), with input characterized by facial landmarks, and a second one, a convolutional neural network based on the DenseNet backbone, with the face image itself as the input. The ADAM optimizer was used during the training of this network. The AffectNet database was employed. The best result obtained was an accuracy of 60.17% in the test subset, for the 7 classes modality. This result is comparable to the best results obtained on the AffectNet dataset.**

Keywords—facial expressions, fusion, convolutional neural network

I. Introduction

Facial expression recognition has diverse applications in fields such as psychology, medicine, and human-machine interactions. In psychology, it is used to understand and analyze emotions, assisting in the assessment of mental states and the diagnosis of psychological disorders. In medicine, it contributes to patient monitoring and the identification of indicators of pain and fatigue. Additionally, in human-machine interfaces, it enhances interaction and communication, enabling computers and devices to understand users' facial expressions. These applications highlight the potential of facial expression recognition as a valuable tool in various areas, driving advancements in diagnostics and technological interaction. The availability of large labeled datasets, such as AffectNet, has allowed the training of more robust models capable of recognizing a wide range of emotional expressions [1]. Training a facial expression model on the AffectNet dataset presents significant challenges due to unbalanced data. The under-representation of some classes can lead to bias towards the majority classes [1]. The use of facial landmarks is essential in facial expression recognition [2]. These points, also known as landmarks, are specific positions on the face, such as the corners of the eyes, nose, and mouth. Detecting and tracking these points allow the extraction of relevant information to describe the configuration and facial movements associated with different facial expressions [3].

With the aim of enhancing automated facial expression recognition, this work proposes a model based on the fusion of features extracted from two data types: face images and facial landmarks. The face images features were extracted with the DenseNet deep neural network, while the facial landmarks features were extracted with a 1D deep neural network. The performance of this proposed model was evaluated with facial expressions of the AffectNet dataset [1].

II. Literature Review

The classification of human facial expressions through machine learning algorithms remains a rapidly developing research area.

Based on the literature, it was possible to group the reviewed papers in two groups. One of them uses only images for facial expression recognition [1,4,5,6], while the other uses features from facial landmarks associated with face images [7,8].

Given the unbalanced nature of the AffectNet dataset, in [1] the authors applied 3 types of techniques to mitigate it. They performed down-sampling on the majority classes, up-sampling on the minority classes, or training with weights proportional to the distribution of the classes. Results from these three approaches were obtained and compared.

The most common procedure for recognizing facial expressions consists initially of detecting the face, followed by pre-processing and applying this face to a neural network. In approaches that use the face landmarks, they are extracted with a neural network and inserted into a one-dimensional vector for further processing. In [4] the authors proposed a facial expression classifier with hierarchical deep learning, combining visual and geometric features. Pre-processing includes face detection and Local Binary Pattern [14] application, followed by cropping and a blur filter. Validated on the CK+ [15] and JAFFE [16] datasets, the model achieved accuracies of 96.46% and 91.27%, respectively. In [7] the authors proposed a network architecture with two branches, face image and facial landmarks, treating the input data as a sequence of images. Each branch was pre-trained individually and then the last layers were unified for final adjustments. Normalization of images and landmarks was carried out, resulting in outstanding performance. The model achieved 97.60% accuracy on the CK+ dataset [15].

In [5], the authors proposed a lightweight neural network with a sparse weight matrix, using depth-separated convolutions and global pooling to reduce parameters. Pre-processing included face detection, exclusion of undetected images and face alignment. The model achieved an accuracy of 98.38% on the CK+ dataset [15].

In [6], the authors investigated the impact of variation in the area of interest of the detected faces on the recognition of facial expressions, using three approximation scales. Normalizing the images accelerated the training with the deep neural networks. The Resnet50 architecture was used to classify six classes in the CK+ dataset [15]. At the end of training, they achieved an average accuracy of 98% in the test set, showing the influence of non-facial information on the model's performance.

In [12], the authors proposed a novel convolutional neural network (CNN) architecture with an attention mechanism, to tackle challenges in facial expression recognition under uncontrolled conditions. The CNN effectively addresses issues as occluded faces, emphasizing discriminative non-occluded regions. Evaluated on diverse datasets, including AffectNet and RAF-DB [17], the proposed CNN outperforms existing methods. In the AffectNet dataset [1] this study achieved an accuracy of 58.78%. Visualization results demonstrate the ability of the proposed CNN to shift attention dynamically, emphasizing its efficacy in real-world scenarios.

In [13], the authors Implemented a deep neural network (DNN), specifically a CNN combining squeeze-and-excitation and residual networks, for facial emotion recognition (FER). The proposed architecture explores critical facial features by extracting feature maps from residual blocks, revealing the importance of nose and mouth regions. An accuracy of 56.54% was obtained in the AffectNet dataset [1].

In this work we intend to evaluate the performance of a neural network architecture made up of two branches. One of the branches, a 1D deep neural network, extracts features from facial landmarks, while the other branch, a 2D deep neural network, extracts features from face images. A similar strategy was adopted by [7]. However, in [7] the authors used a sequence of images of the same individual to perform the inference. In this work, a single image of the individual was used to perform the inference. As the target database comprises only one image per individual, it is unfeasible to assess performance through the comparison of multiple images of the same person. Such an analysis necessitates a dataset presented in video format.

III. Materials and Methods

A. Materials

In this work we used facial expressions from the AffectNet dataset [1]. This dataset is distributed in two versions, a larger one with around 1 million images, and a smaller one with around 291,000 instances. The experiments were carried out on the reduced version. The RGB images of AffectNet have 224x224 pixels and comprises 7 classes: angry, disgust, fear, happy, neutral, sad and surprise. The test set is composed of 500 imagens from each class.

Given the unbalanced nature of the dataset, the following strategy was adopted to balance it: down-sampling on the majority classes and up-sampling on the minority classes, using rotations, zooms and horizontal flips. Table I shows the number of images in each class before and after data balancing. The perfect balance cannot be achieved as it would necessitate generating a disproportionately large volume of images for certain classes, potentially leading to overfitting. The dataset has been split in the proportion 90/10 for the training and validation. As a perfect balanced dataset is not possible to obtain, a complementary procedure employed was training with weights proportional to the distribution of the classes [11].

TABLE I. NUMBER OF IMAGES IN EACH DATABASE, BEFORE AND AFTER DATABASE BALANCING

Class	Before balancing	After balancing		
		Train	Validation	Test
Neutral	74743	67279	7474	500
Happy	134181	120763	13418	500
Sad	25395	22856	2539	500
Surprise	14065	12659	1406	500
Fear	6363	5727	636	500
Disgust	3797	3417	380	500
Anger	24789	22310	2479	500

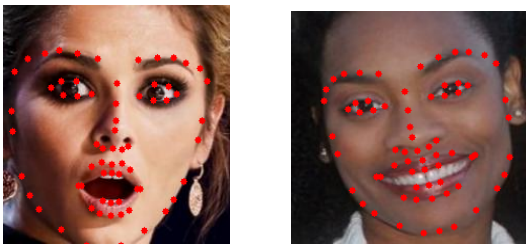
The validation set was employed for fine-tuning the model's hyperparameters. As a stopping criterion during training, early stopping with a patience of 5 was utilized, meaning that if the loss on the validation set did not improve for 5 consecutive epochs, the training process was halted.

B. Methodology

Since the proposal takes into account the use of facial reference points, an initial data extraction stage was necessary. The reference points are made up of 68 points in

relative coordinates. The reference points were extracted and stored sequentially in a one-dimensional vector. Fig. 1 illustrates examples of facial landmarks extraction from the AffectNet dataset. Fig. 2 illustrates the procedure used for extraction of reference points. During the training process the images were normalized. Fig. 3 shows a block diagram of the preprocessing steps performed on both data types.

As shown in Fig.4 the proposed methodology consists of four stages. The first stage involved training the 1D CNN. Before it, the face reference points from the entire database were extracted and stored. The second stage consists of training the 2D CNN. During this stage, online data augmentation operations were carried out, including rotations, zooms and horizontal flips. In the third stage, features from the last dense layer from both architectures were concatenated. In the fourth stage, a new training session involving the entire architecture was carried out. At this stage, the weights of the last dense layer shown in Fig. 4 are adjusted.



(a) (b)

Fig. 1. Example of face images of databases used in this work: (a) Example of surprise expression; (b) Example of happiness expression.

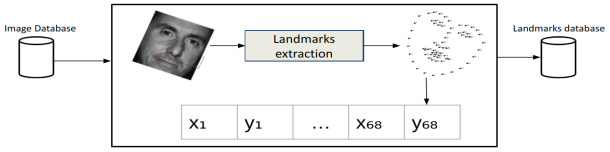


Fig. 2. Landmarks data extraction

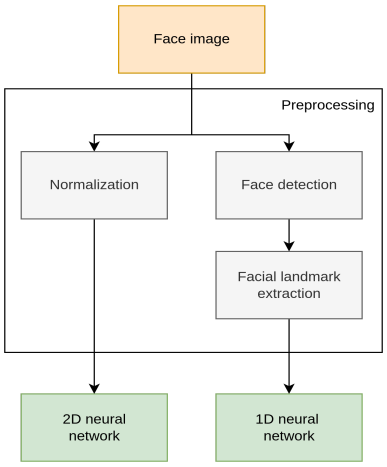


Fig. 3. Flowchart showing preprocessing steps applied in the input image.

C. *Methods*

Fig. 4 shows the adopted methodology.

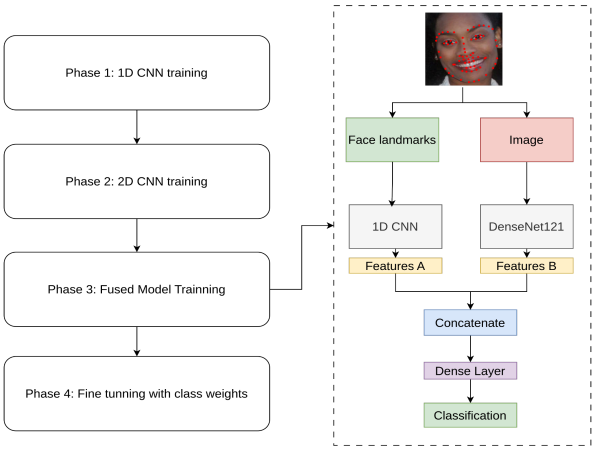


Fig. 4. Block diagram of the proposed methodology for recognizing face pain expressions.

Facial landmark extraction for face detection was conducted utilizing two libraries: Mediapipe and OpenCV2. Mediapipe was selected for its enhanced precision compared to Dlib, a widely used library offering 68 landmark points. In contrast, Mediapipe provides a comprehensive set of 478 points per detected face. Despite the extensive capabilities of Mediapipe, this study adhered to the convention of utilizing only 68 points, a practice commonly observed in scientific literature focusing on facial landmarks [6, 7, 8, 12]. Consequently, a mapping process was undertaken to align the selected landmarks, followed by the storage of coordinates in a numpy array file.

The 1D CNN used for feature extraction from the facial landmarks assumes that the data involves sequential correlated information. 1D CNN excel at capturing local patterns and dependencies within sequential data, making them well-suited for tasks like processing facial landmarks, where the order and relationships among landmarks are crucial. The 1D architecture chosen follows the CNN proposed in [9], for detecting and classifying different types of arrhythmias in ECG signals. In this CNN architecture there are seven feature extraction blocks, made up of the following layers: Conv→Batch→Relu. The number of filters of the convolutional layers is equal to 96. A stride of 1 is used. In the sequence, there are the following layers: fully connected layer (200 neurons) → dropout layer → fully connected layer (17 neurons) → softmax layer → classification layer.

The 2D CNNs used for feature extraction from the face image were the DenseNet121 and DenseNet169. These CNNs are pre-trained deep network models available in TensorFlow. They were pretrained on the ImageNet dataset. Both DenseNet architectures are known for their dense connectivity, where each layer receives input not only from its predecessor but also from all preceding layers. DenseNet121 and DenseNet169 consists of 121 layers and 169 layers, respectively. Both of them excels in computer vision tasks, particularly image classification. The dense connections foster efficient feature reuse, enabling the network to capture intricate patterns effectively [10].

During all training phases, a learning rate of 0.001 was used with the Adam optimizer. The stopping criterion was a patience of 3 epochs until the loss in the validation set stopped improving.

IV. Results and discussion

The training and tests were performed aiming at the 7 classes configuration of the AffectNet. Table II shows the best results obtained for each phase of the pipeline. As shown, the best performance was achieved by the fusion model trained with the DenseNet121. We believe that the larger size of the DenseNet169 network caused overfitting during training. The following metrics were obtained with the fusion model using DenseNet121: average precision of 61%, average recall of 60%, average F1-score of 60%, and overall accuracy of 60.17%. Fig. 5 shows the corresponding confusion matrix. By analyzing the confusion matrix, it can be seen that the happiness class performs better than the others, which is reflected in the clear distinction from other classes.

The best performance obtained in this study in the AffectNet dataset [1], an accuracy of 60.17%, was greater than that obtained in [13], 56.54%. In [13], the authors also verified the importance of facial landmarks in recognizing facial expressions with deep neural networks. The accuracy obtained in this work also outperformed [12] by 1.39%. In [12], the authors worked with attention mechanisms, mainly to treat cases where there is occlusion of the face.

V. Conclusion

Seeking to contribute to research in the area of facial expression recognition, this paper proposed the fusion of features extracted from two CNN architectures. The strategy combines local patterns of facial landmarks with global patterns of the face image. Even with an unbalanced dataset, and working with a partition of the original dataset, the method proved to be competitive with other results in the literature. An accuracy of 60.17% was obtained with 7 classes.

Acknowledgment

This research, carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 39 of Decree n°10.521/2020, was funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law n°8.387/1991 through agreement 001/2020, signed with UFAM and FAEPI, Brazil.

References

[1] A. Mollahosseini, B. Hassani, M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” *2019 IEEE Transactions on Affective Computing, 10(1), 18-31*.

[2] S. Ghosh, A. Laha, N. A. Murshed, “Facial landmarks: A comprehensive survey, applications, performance evaluation, and open issues,” 2020 *IEEE Transactions on Multimedia, 22(8), 2066-2094*.

[3] M. F. Valstar, B. Martinez, X. Binefa, M. Pantic, “Facial expression analysis,” 2016 *In Handbook of Face Recognition (2nd ed., pp. 433-459). Springer*.

[4] J. -H. Kim, B. -G. Kim, P. P. Roy and D. -M. Jeong, "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," in IEEE Access, vol. 7, pp. 41273-41285, 2019, doi: 10.1109/ACCESS.2019.2907327.

[5] Q. Chen, X. Jing, F. Zhang, J. Mu, “Facial Expression Recognition Based on A Lightweight CNN Model,” 2022 *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Bilbao, Spain, 2022, pp. 1-5, doi: 10.1109/BMSB55706.2022.9828739*.

[6] O. F. Soylemez, B. Ergen, “Facial Landmark Based Region of Interest Localization for Deep Facial Expression Recognition,” 2022 *Tehnicki vjesnik - Technical Gazette*.

[7] M. Verma, H. Kobori, Y. Nakashima, N. Takemura, H. Nagahara, “Facial Expression Recognition with Skip-Connection to Leverage Low-Level Features,” 2019 *IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 51-55, doi: 10.1109/ICIP.2019.8803396*.

[8] R. Wadhawan, T. K. Gandhi, “Landmark-Aware and Part-Based Ensemble Transfer Learning Network for Static Facial Expression Recognition from Images,” 2023 *IEEE Transactions on Artificial Intelligence, vol. 4, no. 2, pp. 349-361, April 2023, doi: 10.1109/TAI.2022.3172272*.

[9] J. R. Santana, M. G. F. Costa, C. F. F. Costa Filho, “A New Approach to Classify Cardiac Arrhythmias Using 2D Convolutional Neural Networks,” 2021 *In PubMed.*

[10] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[11] C. M. Bishop, “Pattern Recognition and Machine Learning,” Springer, Madison, 2006.

[12] Y. Li, J. Zeng, S. Shan and X. Chen, "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism," in IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439-2450, May 2019, doi: 10.1109/TIP.2018.2886767.

[13] Huang ZY, Chiang CC, Chen JH, Chen YC, Chung HL, Cai YP, Hsu HC. A study on computer vision for facial emotion recognition. Sci Rep. 2023 May 24;13(1):8425. doi: 10.1038/s41598-023-35446-4. PMID: 37225755; PMCID: PMC10209161.

[14] A. Hadid, "The Local Binary Pattern Approach and its Applications to Face Analysis," 2008 First Workshops on Image Processing Theory, Tools and Applications, Sousse, Tunisia, 2008, pp. 1-9, doi: 10.1109/IPTA.2008.4743795.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[16] K. Miyuki, L. Michael and G. Jiro. “The japanese female facial expression (jaffe) database, ” 1997 Available: http://www. kasrl. org/jaffe. html.

[17] S. Li and W. Deng. “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition” IEEE Transactions on Image Processing, 28(1):356–370, 2019. 5

TABLE II. Results obtained for face expressions recognition with the affectnet dataset

Model	Mean Values (%)			Global Accuracy (%)
	Precision	Recall	F1-Score	
1D CNN	51.36	50.94	50.78	51.53
2D CNN – DenseNet 121	59.90	50.82	48.82	50.83
2D CNN – DenseNet 169	60.10	50.14	47.67	50.14
Fusion Model (1D CNN + DenseNet 121)	61	60	60	60.17
Fusion Model (1D CNN + DenseNet 169)	58	58	58	58

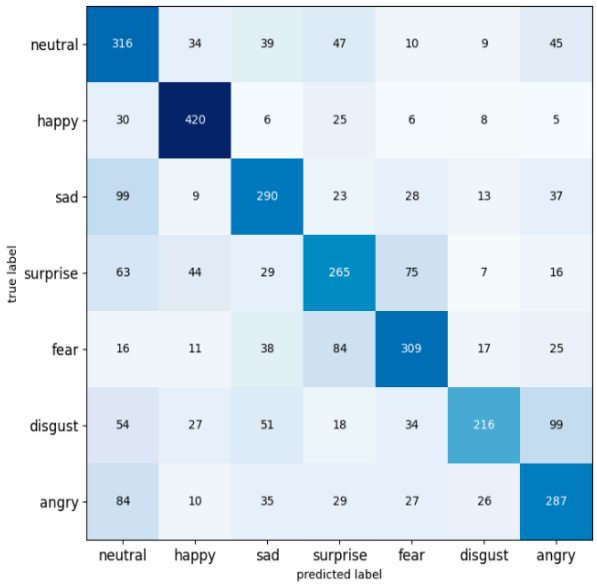


Fig. 5. Confusion matrix for the fusion model using the DenseNet121