UNIVERSIDADE FEDERAL DE OURO PRETO
ESCOLA DE MINAS
COLEGIADO DO CURSO DE ENGENHARIA DE CONTROLE E AUTOMAÇÃO - CECAU

LUCIANO ERMELINDO DE ALMEIDA

ESTRATÉGIA MULTIMODELO DE REDES NEURAIS CONVOLUCIONAIS PARA CLASSIFICAÇÃO DE EXPRESSÕES FACIAIS

MONOGRAFIA DE GRADUAÇÃO EM ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Ouro Preto, 2022

LUCIANO ERMELINDO DE ALMEIDA

ESTRATÉGIA MULTIMODELO DE REDES NEURAIS CONVOLUCIONAIS PARA CLASSIFICAÇÃO DE EXPRESSÕES FACIAIS

Monografia apresentada ao Curso de Engenharia de Controle e Automação da Universidade Federal de Ouro Preto como parte dos requisitos para a obtenção do Grau de Engenheiro de Controle e Automação.

Orientador: Prof. Ph.D. Guillermo Cámara Chávez

Coorientador: Prof. Dr. Agnaldo José da Rocha Reis

Ouro Preto
Escola de Minas – UFOP
2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

A447e Almeida, Luciano Ermelindo De.
Estratégia multimodelo de redes neurais convolucionais para classificação de expressões faciais. [manuscrito] / Luciano Ermelindo De Almeida. - 2022.
45 f.: il.: color., gráf., tab.. + Matriz de confusão.

Orientador: Prof. Dr. Guillermo Cámara Chávez.
Coorientador: Prof. Dr. Agnaldo José da Rocha Reis.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Escola de Minas. Graduação em Engenharia de Controle e Automação .

1. Expressão facial. 2. Rede Neural -Convulacionais. 3. Fusão- sistema de votação suave. 4. Redes neurais-computação. I. Chávez, Guillermo Cámara. II. Reis, Agnaldo José da Rocha. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 681.5

Bibliotecário(a) Responsável: Angela Maria Raimundo - SIAPE: 1.644.803

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Luciano Ermelindo de Almeida

Estratégia Multimodelo de Redes Neurais Convolucionais para Classificação de Expressões Faciais

Monografia apresentada ao Curso de Engenharia de Controle e Automação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Engenheiro de Controle e Automação

Aprovada em 04 de janeiro de 2022

Membros da banca

Prof. Dr. - Guillermo Cámara Chávez - Orientador (Universidade Federal de Ouro Preto)
Prof.Dr. - Agnaldo José da Rocha Reis - Coorientador (Universidade Federal de Ouro Preto)
Prof. Dr. - Rodrigo César Pedrosa Silva - (Universidade Federal de Ouro Preto)
Prof. Me. - Pedro Henrique Lopes Silva - (Universidade Federal de Ouro Preto)

Guillermo Cámara Chávez, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 09/02/2022

Documento assinado eletronicamente por **Guillermo Camara Chavez**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 09/02/2022, às 16:57, conforme horário oficial de Brasília, 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 http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0277868** e o código CRC **2EDA0FF3**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho a Deus, a quem estendendo a mão me deixo ser guiado,
dedico também à republica dusmininu segunda família,
e também à minha primeira família, mãe, Nilma, Valentina e Petrus, obrigado por tudo.

AGRADECIMENTOS

Agradeço primeiramente à Deus por me ensinar a continuar caminhando. Agradeço também ao professor Guillermo por insistir na qualidade do resultado, e ao professor Agnaldo pelas excelentes aulas que me inspiraram a entender o potencial das redes neurais. E por fim, porém não menos importante, agradeço à UFOP e ao departamento de controle e automação pela excelência no ensino.

Cum mente et malleo. Viva a Escola de Minas!

"Com ele falo face a face, claramente, e não por enigmas; e ele vê a forma do Senhor."
(Números 12:8)

RESUMO

O reconhecimento de emoções desempenha um papel promissor no sistema de interação homem-máquina. Diferentes técnicas tem sido desenvolvidas para a solução deste desafio, dentre elas as arquiteturas de redes neurais convolucionais estão entre as mais eficientes. Este trabalho apresenta uma solução multimodelo, onde diferentes redes neurais convolucionais são treinadas de forma individual e a fusão das redes acontece no nível de decisão utilizando um sistema de votação que garante a diminuição dos erros das classificações de cada rede. Ao todo quatro CNN's foram treinadas na base de dados FER2013+, que é uma base de dados bem diversificada, as quatro redes utilizaram a técnica de *fine tuning* cada uma com seu respectivo modelo base, a saber, VGG16, VGG19, Xception e o InceptionV3, desta maneira cada uma das redes adquire capacidades diferentes de reconhecer as expressões, garantindo a eficiência da fusão. Cada uma das quatro redes obtiveram acurácias individuais de 81,4%, 81,1%, 79,8% e 80,5%, para as oito expressões de emoções faciais universais, felicidade, surpresa, tristeza, raiva, nojo, medo, desprezo e neutro, fazendo a fusão dos quatro modelos a acurácia atinge 84,2%, para a fusão contendo três modelos a acurácia é de 83,8% e o sistema multimodelo com duas arquiteturas o resultado é de 82,4%.

Palavras-chaves: Expressão facial. Rede Neural Convolucional (CNN). Multimodelo. Fusão com sistema de votação suave. FER2013+.

ABSTRACT

Emotion recognition plays a promising role in the human-machine interaction system. Different techniques have been developed to solve this challenge, among them the convolutional neural network architectures are among the most efficient. This work presents a multi-model solution, where different convolutional neural networks are individually trained and the fusion of networks takes place at the decision level using a voting system that guarantees the reduction of errors in the classifications of each network. Altogether four CNNs were trained in the FER2013+ database, which is a well-diversified database, the four networks used the technique defines tuning each one with its respective base model, namely, VGG16, VGG19, Xception, and InceptionV3, in this way each of the networks acquires different capabilities to recognize expressions, ensuring the efficiency of fusion. Each of the four networks had individual accuracies of 81.4%, 81.1%, 79.8%, and 80.5% for the eight expressions of universal facial emotions, happiness, surprise, sadness, anger, disgust, fear, contempt, and neutral, making the fusion of the four models the accuracy reaches 84.2%, for the fusion containing three models the accuracy is 83.8% and the multi-model system with two architectures the result is 82.4%.

Key-words: Facial expression. Convolutional Neural Network (CNN). Multi-model. soft-voting-fusion. FER2013+.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAM	Active Appearance Models
AD	Action Descriptors
AU	Action Unit
CK+	Cohn-Kanade extendido - database
CNN	Rede Neural Convolucional
DISFA	Denver Intensity of Spontaneous Facial Action
FACS	Facial Action Coding System
FER	Facial Emotion Recognition
FER+	Facial Emotion Recognition plus
GB	Gross behavior
JAFFE	The Japanese female facial expression - Database
MLCNN	Multinível CNN
MV	Movements
SVM	Support vector machine
VC	Visibility codes
ViT	Vision transformer

SUMÁRIO

1 INTRODUÇÃO

1.1 contextualização

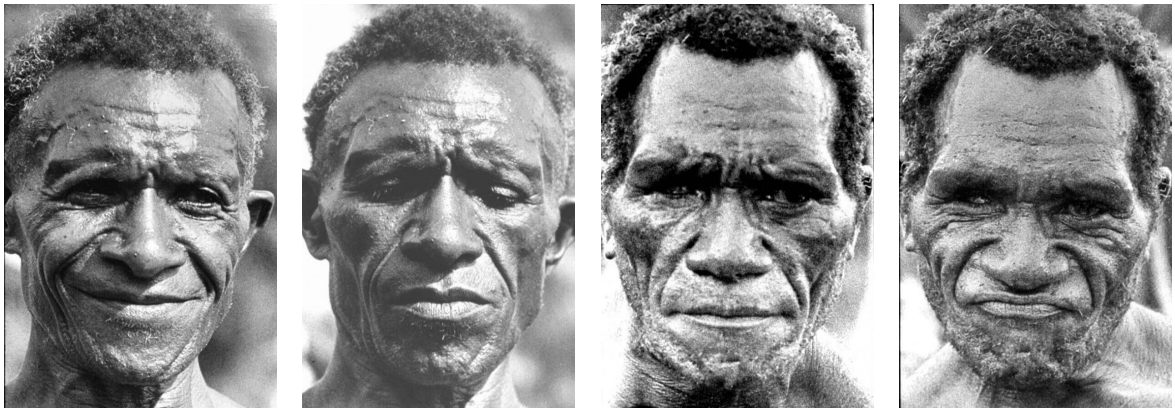
De acordo com (EKMAN, 2006) uma atenção cientificamente séria às expressões faciais só começa em 1872 com o trabalho do naturalista britânico Charles Darwin no seu célebre livro *The expression of the emotion in man and animals*, que ele escreve após ter visitado oito partes do mundo, Africa, América, Austrália, Bornéu, China, Índia, Malásia e Nova Zelândia. Neste livro, Darwin descreve ter encontrado as mesmas expressões de emoções nestas terras distantes das que ele conhecia na sua terra natal, a Inglaterra. Darwin se mostra extasiado pela uniformidade nos padrões das expressões corporais encontrados ao redor do mundo.

Dentre todas as expressões corporais, abordaremos neste trabalho a mais relevante delas, as expressões faciais. Esta universalidade das expressões faciais foi arduamente analisada pelo Dr. Paul Ekman, eleito pela *American Psychological Association* (APA) um dos 100 psicólogos mais importantes do século XX e o maior pesquisador sobre as expressões faciais (EKMAN, 1999). E, embora ele demonstre ter encontrado certas expressões faciais de emoções específicas de diferentes culturas, condicionadas pelas características próprias daquele meio, há também um conjunto pancultural de expressões que independem de qualquer fator externo.

O fato de existirem expressões que são encontradas em todas as culturas do mundo pode ser ilustrado com a história do "homem de Nova Guiné", onde o Dr. Ekman narra o encontro com um homem que vivia em uma cultura isolada e pré-letrada, usando instrumentos de pedra e que nunca tinha visto forasteiros antes (EKMAN, 1972). O pesquisador pediu-lhe que mostrasse como seria o seu rosto se, os seus amigos tivessem vindo lhe visitar, Figura 1a. Se o seu filho acabara de morrer, Figura 1b. Se ele estivesse prestes a lutar, Figura 1c. Se ele tivesse pisado em um porco morto fedorento, Figura 1d. As respostas do homem de Nova Guiné coincidem com a de todas as demais culturas estudas pelo pesquisador, e, como é fácil perceber, com a própria cultura ocidental.

Ao todo o Dr. Ekman encontrou, além das quatro expressões demonstradas pelo homem de Nova Guiné, outras três emoções universais que podem ser codificadas nos músculos faciais (EKMAN, 1999). A saber, as sete emoções básicas são: alegria, tristeza, raiva, nojo, medo, surpresa e desprezo, esta última analisada por (MATSUMOTO; EKMAN, 2004). Vale destacar aqui que o Dr Armindo Freitas Magalhães, Ph.D em Neurociência e Psicologia Biocomportamental, defende em seus mais recentes trabalhos, como em (FREITAS-MAGALHÃES, 2020), a adição da dor como a oitava emoção básica.

Existem duas principais maneiras de construção de um sistema computacional capaz de descobrir as emoções faciais:



(a) Felicidade (b) Tristeza (c) Raiva (d) Nojo

Figura 1 – O homem de Nova Guiné
Fonte: Ekman (1972)

A primeira maneira é por meio dos modelos psicológicos de codificação de ação facial (*Facial Action Coding System - FACS*) onde é possível construir uma emoção baseado em ativações de diferentes músculos da face e em movimentos faciais específicos mas que não estão associados a um músculo isolado, cada um desses movimentos menores é chamado de *Action Units-AU* (ativação de músculo facial específico), *Action Descriptors-AD* (sem ação muscular específica, como por exemplo morder os lábios, bufar de bochecha) e *Moviments-M* (movimento de cabeça, movimento dos olhos etc). Por exemplo, na Figura 2 podemos notar a diferença entre um falso sorriso de um sorriso verdadeiro, perceba que o sorriso verdadeiro é formado pela AU6 + AU12, onde a AU6 é a contração do músculo orbicular dos olhos e a AU12 dos músculos zigomático maior, já no falso sorriso existe a falta da *action unit AU6*. Ao todo são descritos 40 *Action units*, (AU) 7 *Action descriptors* (AD), 41 *Movements* (MV) (11 movimentos de cabeça, 8 de olhos e 22 da língua), 9 *Gross behavior* (GB) e 5 *Visibility codes* (VC) (FREITAS-MAGALHÃES, 2020), de modo que, em cada emoção existem combinações distintas destas microexpressões. A tabela com os 102 códigos do FACS pode ser visto no anexo A.

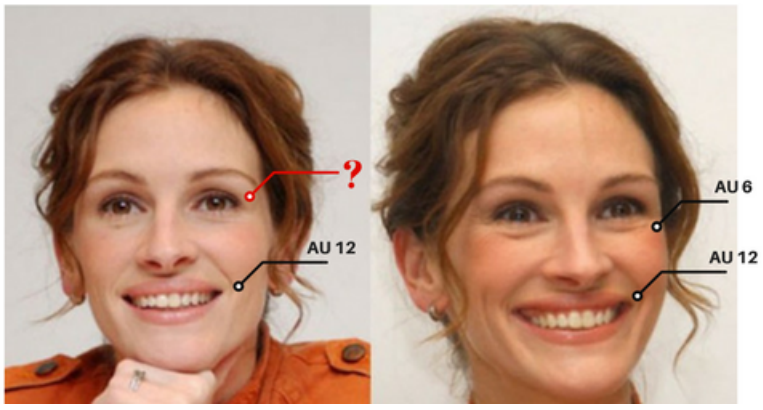


Figura 2 – Sorriso falso versus sorriso verdadeiro
Fonte: David Leucas (2018)

A segunda maneira de construir um sistema de detecção de emoções humanas é por meio de classificadores treinados com imagens que contenham as expressões de emoções almejadas. Diversas técnicas podem ser usadas para a realização deste treinamento, alguns exemplos são as redes neurais convolucionais - CNN (MOLLAHOSSEINI; CHAN; MAHOOR, 2016), (KHAIREDDIN; CHEN, 2021), *Support Vector Machine* (ASHRAF et al., 2009) - SVM , Algoritmo de Viola-Jones (AGRAWAL; KHATRI, 2015), dentre outras.

1.2 Objetivos gerais

Desenvolver um método para o reconhecimento de expressões faciais via estratégia multimodelo de rede neural convolucional.

1.3 Justificativa do trabalho

A interação entre o homem e o computador é um ramo de extrema importância para a evolução de interfaces cada vez mais claras, ágeis e amigáveis. Podemos afirmar que evoluções como os sistemas operacionais baseados em interfaces gráficas, as atuais telas sensíveis ao toque, dentre outras inúmeras evoluções, tornaram os sistemas computacionais mais próximas e atraentes para todos os públicos.

Mas, por vezes, os usuários ainda se sentem frustrados com os sistemas computacionais. Um estudo de 1999 amplamente divulgado pela *Concord Communications* nos EUA descobriu que 84% dos gerentes da área de suporte técnico pesquisados disseram que os usuários admitiram se envolver em comportamento "violento e abusivo"em relação aos computadores (PICARD, 1999).

Uma maneira de aumentar a proximidade entre o homem e a máquina é criando interfaces capazes de interpretar as emoções humanas. Se analisarmos com cuidado perceberemos que os seres humanos naturalmente expressam suas emoções para as máquinas, porém as máquinas não conseguem naturalmente interpreta-las, tornando impossível esse tipo de comunicação tão espontânea da espécie humana.

Nem todas os sistemas computacionais se beneficiariam deste tipo de capacidade de interação, mas diversos setores podem se beneficiar, como por exemplo, a telemedicina, a educação, o entretenimento dentre tantos outros.

As emoções humanas podem ser expressas de diversas formas, como por exemplo, pela tonalidade da voz, respiração, ritmo cardíaco. Porém, neste trabalhos daremos enfoque nas expressões de emoções faciais, pois como sabemos 23% da comunicação humana se dá pelas expressões faciais (CUVE, 2015).

1.4 Organização do Texto

O Capítulo 2 trata da fundamentação teórica, abordando as principais técnicas encontradas na literatura capazes de reconhecer as expressões de emoções faciais. Elas serão organizadas em dois grupos, às técnicas que necessitam da extração de características de forma manual para alimentar camadas de aprendizado de máquina tradicionais e das técnicas que utilizam as redes neurais convolucionais, e as técnicas que não precisam ser alimentadas diretamente com estas características, pois são capazes de aprender estes atributos diretamente dos dados da imagem de forma autônoma. O Capítulo 3 fornece detalhes da solução idealizada neste trabalho. A seguir, no Capítulo 4, trata dos experimentos, dos resultados e observações obtidas. E no Capítulo 5, conclusões e trabalhos futuros são descritos.

2 REVISÃO DA LITERATURA

2.1 Bases de dados

A utilização de bases de dados contendo um grande número de imagens abarcando expressões faciais é primordial para este trabalho. Dependendo da qualidade desta base de dados o treinamento da rede neural pode adquirir maior ou menor habilidade de classificar as emoções de forma exata.

Existe uma grande quantidade de trabalhos que trazem bases de dados com expressões faciais, a maioria destes trabalhos contém imagens produzidas em ambiente controlado, algumas por meio de atores que são solicitados a posar para as emoções desejadas, e outros em que os indivíduos são estimulados a sentirem as emoções por meio de vídeos ou mesmo trilhas sonoras. Também existem bases de dados produzidos por imagens coletadas nos sites de busca da internet.

Deve-se fazer distinção também entre as bases de dados contendo imagens classificadas por emoções, intensidade de emoções e microexpressões (de acordo com o código FACS, ver anexo A). Há também uma série de base de dados que apresentam vídeos, contendo ou não arquivos de áudio, que podem ser utilizados para classificadores multicanais (RINGEVAL et al., 2015) e (SCHULLER et al., 2011).

A seguir será destacado algumas das bases de dados mais relevantes:

Um dos primeiros trabalhos foi desenvolvido por uma equipe da Universidade Kyushu, no Japão, denominado: *The Japanese Female Facial Expression* (JAFFE) (LYONS et al., 1998). Esta base de dados contém 213 imagens de 10 mulheres japonesas que posaram para 6 expressões faciais universais (tristeza, surpresa, felicidade, medo, raiva, e nojo) além de fotos com expressão neutra. Perceba que as expressões de desprezo e dor só foram introduzidos como emoções universais em trabalhos mais recentes (MATSUMOTO; EKMAN, 2004) e (FREITAS-MAGALHÃES, 2020).

A base de dados Cohn-Kanade (KANADE; COHN; TIAN, 2000), e mais tarde Cohn-Kanade estendido (LUCEY et al., 2010), também são formadas por imagens posadas. No caso do Cohn-Kanade estendido (CK+) são 593 sequências de imagens de 123 indivíduos diferentes que posaram para expressões neutras e para 7 expressões básicas (tristeza, surpresa, felicidade, medo, raiva, e nojo e desprezo). Tanto CK+ quanto JAFFE contém imagens faciais de alta qualidade e sem obstruções (como óculos, boné, etc). Um exemplo das imagens contidas na base de dados CK+ pode ser visto na Figura 3.

Em (VALSTAR et al., 2006) notamos que existe uma importante diferença entre expressões posadas e expressões espontâneas, essas diferenças podem ser notadas tanto pela configuração dos músculos faciais (microexpressões), quanto pela intensidade destas microex-

Figura 3 – Exemplo das oito expressões faciais de CK+
Fonte: Lucey et al. (2010) - Editada pelo autor

pressões. Por este motivo algumas bases de dados contém imagens registradas de indivíduos que passam por algum tipo de estímulo com o objetivo de expressarem, espontaneamente, as expressões estudadas. Um exemplo é o DISFA (MAVADATI et al., 2013), que possui 4845 frames de vídeos gravados de 27 indivíduos diferentes enquanto estes recebiam estímulos audiovisuais, estes frames não estão codificados por emoções e sim por 12 AUs e a respectiva intensidade de cada uma destas microexpressões, como pode ser visto na Figura 4.

Figura 4 – Exemplo das 12 AU codificadas em DISFA
Fonte: Mavadati et al. (2013) - Editada pelo autor

O EmotionNet, desenvolvido em (BENITEZ-QUIROZ; SRINIVASAN; MARTINEZ, 2016), é um exemplo de uma base de dados mais diversificadas, contendo imagens com grande quantidade indivíduos, variações de postura, ângulo de cabeça, luz, obstruções parciais. A EmotionNet contém mais de 1 milhão de imagens capturadas da internet por meio de buscas por palavras associadas ao termo 'sentimento' do artigo *Word Net* (MILLER, 1995). Estas imagens, depois de passarem por filtros que excluem imagens repetidas ou imagens que não contém rostos, passam por um algoritmo que codifica estas imagens em 12 AUs junto com suas intensidades e por 23 expressões de emoções básicas e compostas. Aqui vale destacar que as expressões compostas são aquelas que possuem microexpressões de mais de uma emoção, por exemplo, se uma imagem foi anotada como tendo as AUs 1, 2, 12 e 25, é codificado como feliz e surpreso, a codificação completa pode ser vista na Tabela 1. Vale lembar também que as imagens da EmotionNet continuam codificadas pelas palavras do *Word Net* usadas para realizar a pesquisa. Uma desvantagem da EmotionNet é que as imagens são classificadas automaticamente por um algoritmo desenvolvido pela equipe, apenas 10% (100.000) das imagens foram manualmente classificadas por especialistas. A acurácia do algoritmo classificador é de 80%.

Tabela 1 – Lista das emoções básicas e compostas da base de dados EmotionNet

Fonte: Benitez-Quiroz, Srinivasan e Martinez (2016)

Categoria	AUs	Categoria	AUs
Happy	12, 25	Sadly disgusted	4, 10
Sad	4, 15	Fearfully angry	4, 20, 25
Fearful	1, 4, 20, 25	Fearfully surpd.	1, 2, 5, 20, 25
Angry	4, 7, 24	Fearfully disgd.	1, 4, 10, 20, 25
Surprised	1, 2, 25, 26	Angrily surprised	4, 25, 26
Disgusted	9, 10, 17	Disgd. surprised	1, 2, 5, 10
Happily sad	4, 6, 12, 25	Happily fearful	1, 2, 12, 25, 26
Happily surpd.	1, 2, 12, 25	Angrily disgusted	4, 10, 17
Happily disgd.	10, 12, 25	Awed	1, 2, 5, 25
Sadly fearful	1, 4, 15, 25	Appalled	4, 9, 10
Sadly angry	4, 7, 15	Hatred	4, 7, 10
Sadly surprised	1, 4, 25, 26	-	-

Outra base de dados importante é a FER2013(*face emotion recognition*). Inicialmente esta base de dados foi parte do trabalho dos pesquisadores Pierre-Luc Carrier e Aaron Courville, mas tomou notoriedade após ser usado em uma competição de reconhecimento de expressões faciais organizado por *Ian Goodfellow, Dumitru Erhan, e Yoshua Bengio durante o ICML workshop* organizado pela empresa Kaggle em 2013 (GOODFELLOW et al., 2013). Existem muitos trabalhos publicados que utilizam esta base, o que a torna estado da arte na área de reconhecimento de emoções. Outra vantagem é que este conjunto de dados é muito diverso, contendo uma boa quantidade de personagens e de variações de luz e obstruções, tornando-a uma base de dados bem realista. A principal desvantagem é que esta base contém erros que vão deste imagens mau categorizadas até imagens corrompidas, para resolver este problema foi desenvolvida uma nova classificação chamada de FER2013+ (BARSOUM et al., 2016). A FER2013+ possui as mesmas imagens que a FER2013 porém a classificação foi melhorada, as

imagens mau categorizadas foram corridas, as imagens corrompidas foram sinalizadas como tal e foi adicionado a expressão de desprezo, já que algumas imagens apresentavam esta expressão, neste trabalho será usada a FER2013+.

2.2 Algoritmos Existentes

Muitos trabalhos têm sido realizados com intuito de detectar e reconhecer emoções, é possível fazer distinção entre dois grupos de algoritmos: No primeiro agrupamento colocamos os algorítimos baseados em *HandCrafted features* que são derivados de características encontradas na imagem, como brilho, cor, arestas, etc e que utilizam métodos tradicionais de aprendizado de máquina alimentados por estas características. E um segundo grupo que utiliza CNN (Rede neural convolucional) que não precisam ser alimentadas diretamente com estas características, pois são capazes de 'aprender' estes atributos diretamente dos dados da imagem de forma autônoma.

2.2.1 Algoritmos com extração de características *HandCrafted* (manufaturadas)

Como exemplo de algoritmo do primeiro grupo podemos ver, (ASHRAF et al., 2009) onde é desenvolvido um classificador binário baseado em SVM que faz distinção entre vídeos contendo expressões faciais de dor e vídeos sem expressão faciais de dor. Os autores utilizam a técnica *Active Appearance Models*(AAM), descrita pela primeira vez por (COOTES; EDWARDS; TAYLOR, 1998), que representa o rosto por meio de um modelo estatístico de sua forma e aparência (Figura 5), de modo que, cada indivíduo possua o seu próprio modelo base, maximizando assim a correspondência entre o modelo e a imagem de entrada.

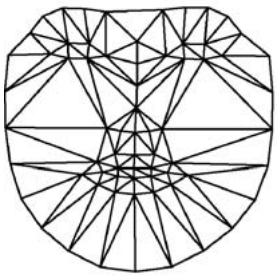


Figura 5 – *Active Appearance Models*(AAM) com 68 vértices
Fonte: Cootes, Edwards e Taylor (1998)

Já em (MOHAMMADI; FATEMIZADEH; MAHOOR, 2014) a técnica usada é o *Sparse representation* que utiliza um dicionário contendo um determinado vetor de características. Então, aplicando um dos *sparse solvers*, pode-se encontrar a representação *sparse* de um novo vetor de características e daí classificar com base nesta representação. Esta técnica é bem simples, porém promissora em relação ao tratamento de imagens com oclusões e ou ruídos.

Também pode-se destacar (LIU; WECHSLER, 2002) e (SHAN; GONG; MCOWAN, 2009) que utilizam respectivamente *Gabor filter* e *Local Binary Patterns*, além destas, existem várias outras técnicas *HandCrafted* não descritas nesta revisão.

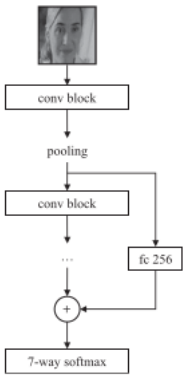
2.2.2 Algoritmos com extração de características via redes neurais convolucionais

Com relação às redes neurais convolucionais não é necessário se preocupar com a etapa de extração de características, já que a rede tem a capacidade de aprender estas características automaticamente, sendo assim, basta preocupar-se com a arquitetura da rede, de modo que o aprendizado não fique nem sobreajustado (overfitting) nem subajustado (underfitting).

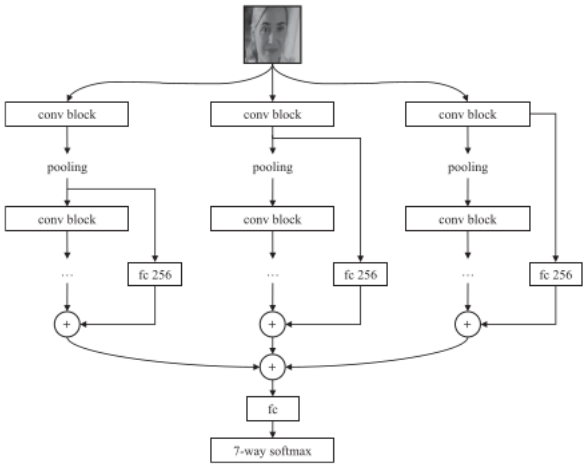
Em (MOLLAHOSSEINI; CHAN; MAHOOR, 2016) é proposto a criação de um rede convolucional contendo dois elementos, o primeiro consiste em duas camadas CNN tradicionais formadas por uma camada convolucional seguida de uma camada de max pooling, em seguida o autor acrescenta dois módulos com a arquitetura do estilo *Inception*. A arquitetura *Inception* foi desenvolvida por pesquisadores da Google (SZEGEDY et al., 2015). Tendo como resultado 66.4% de acurácia na base de dados FER2013.

Já em (KHAIREDDIN; CHEN, 2021) é utilizado como referência a arquitetura VGG16 ganhadora da competição ILSVR(Imagenet) em 2014. O trabalho utiliza a técnica de *fine tuning*. Nesta técnica o modelo VGG16 é inicialmente congelado e uma nova camada é adicionada no topo do modelo, desta maneira, durante o treinamento os pesos anteriormente aprendidos pela rede consagrada serão aproveitados. Em seguida o modelo VGG16 é descongelado e uma nova seção de treinamento acontece. O resultado obtido na base FER2013 foi de 73%.

Em (HAI-DUONG et al., 2019) é utilizado uma técnica chamada de rede neural convolucional multinível(MLCNN), esta técnica tira vantagem de características que são adquiridas por camadas intermediárias, ver Figura 6a. Existem muitas variantes da técnica MLCNN, de modo que o mapa de recursos adquirido pode vir de camadas de *pooling*, ou de camadas convolucionais dentre outras possibilidades. Pensando nisso (HAI-DUONG et al., 2019) cria um conjunto de três variações da técnica MLCNN, ver figura 6b conectando-os em uma arquitetura unificada, conseguindo um resultado de 74% na base FER2013.



(a) CNN Multinível



(b) Unificação de três redes CNN Multinível

Figura 6 – Modelo CNN Multinível(MLCNN)
Fonte: Hai-Duong et al. (2019)

A contribuição da pesquisa apresentada por (SAABNI; SCHCLAR et al., 2020) é a utilização da técnica de *fine tuning* utilizando a combinação de três modelos consagrados em uma grande rede convolucional. Depois de combinar esses modelos em paralelo, ele adiciona outras camadas. Nesse trabalho é feita uma comparação do resultado na base FER2013 e na sua atualização FER2013+. A acurácia obtida na FER2013 foi de 74.4% e na FER2013+ foi de 85.1%. Porém é necessário destacar que o trabalho utiliza técnicas *HandCrafted* para detectar as faces frontais antes de treinar a rede neural, e desta maneira acaba excluindo imagens com angulações de cabeça ou obstruções, como a imagem da Figura 7 tornando o resultado incapaz de reconhecer imagens com obstrução parcial ou angulação de cabeça.

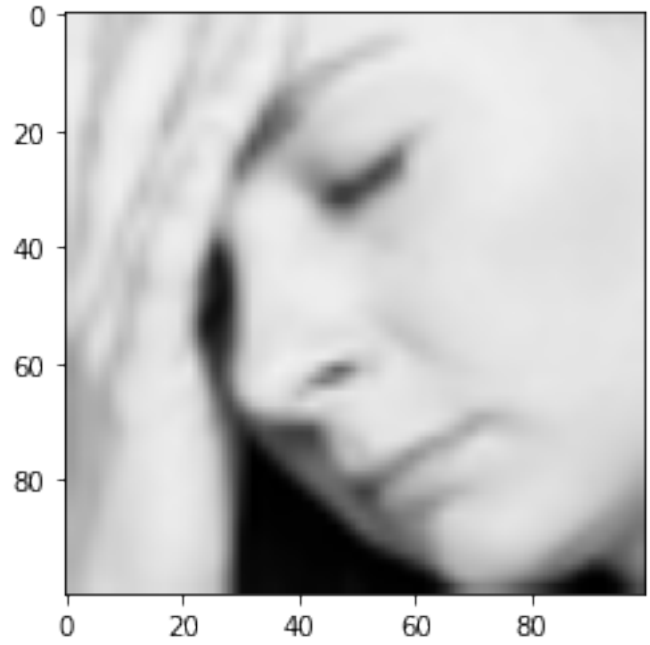


Figura 7 – Face com angulação e obstrução parcial
Fonte: Base de dados FER2013

Já em (LING; LIANG; YANG, 2021) é utilizado a técnica *Vision transformer* (ViT). Uma rede transformer é um modelo de aprendizado de máquina que imita a atenção cognitiva, dando importâncias maiores para determinados dados e menores para outros. O efeito aumenta as partes importantes dos dados de entrada e esmaece o resto, o pensamento é que a rede deve dedicar mais poder de computação a essa parte pequena, mas importante dos dados. Qual parte dos dados é mais importante do que outras depende do contexto e é aprendido por meio dos dados de treinamento por gradiente descendente. No caso da *Vision transformer*, onde os dados de entrada são uma imagem seria muito custoso computacionalmente realizar a análise para cada pixel, por isso a imagem é dividida em seções menores. O resultado deste método na base FER2013+ foi de 85,05%. O trabalho também realizou, para efeito de comparação, experimentos com a rede neural convolucional ResNet e conseguiu resultados de 83,02% com a ResNet50 e 81,02% ResNet18. Na Figura 8 podemos ver um exemplo de como podem se comportar as divisões da imagem por meio de uma *Vision transformer* (ViT).

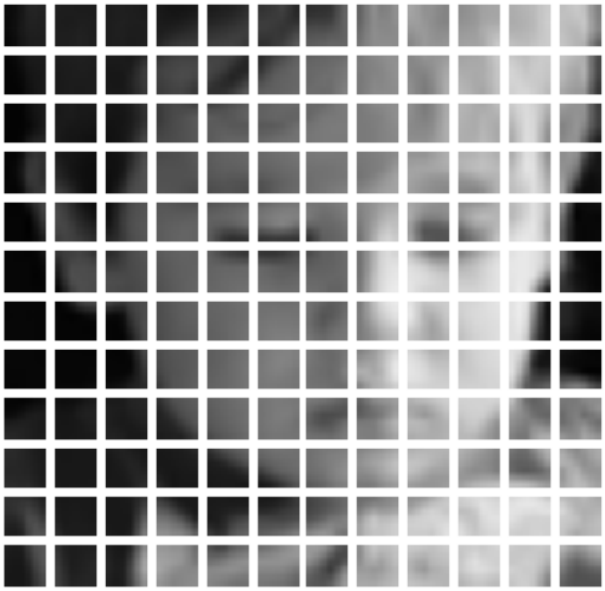


Figura 8 – Exemplo de divisões da imagem para a utilização do ViT
Fonte: Elaborada pelo autor

Comparando-se os dois grupos, pode ser observado que os trabalhos que utilizam CNN são mais abrangentes, sendo portanto mais eficientes em relação às variações de luz da cena, tipos de câmera, resoluções da imagem, planos de fundo, ângulos da cabeça, etnias, etc.

3 METODOLOGIA

O método multimodelo desenvolvido neste trabalho, ver Figura 9, é inspirada no trabalho (YÜZKAT; ILHAN; AYDIN, 2021) e consiste na fusão no nível de decisão de quatro modelos de rede neural convolucional. No primeiro momento os modelos individuais são treinados de maneira isolada, como os modelos são diferentes eles desenvolvem recursos de classificação diferentes. Na etapa final a classificação individual de cada modelo é somada utilizando a técnica *soft-voting-fusion* descrita por (YÜZKAT; ILHAN; AYDIN, 2021). Esta estratégia multimodelo é capaz de driblar momentos de falha que acontecem em um dos modelos de maneira particular mas que não se repetem nos demais, tornando o resultado final mais robusto. Essa melhora na qualidade final acontece pois, como os modelos individuais possuem recursos diferentes de classificação, uma determinada imagem que é erroneamente classificada por uma das redes pode ser corretamente classificada pelas demais, desta maneira a fusão de todas as classificações tende a obter um melhor resultado que as redes individuais.

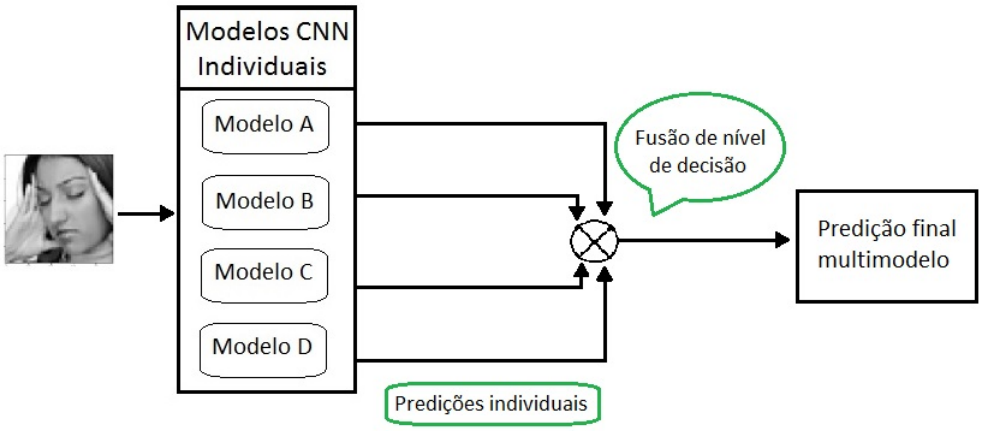


Figura 9 – Arquitetura multimodelo
Fonte: Elaborada pelo autor

3.1 Arquiteturas individuais

Todas as 4 arquiteturas individuais utilizaram a técnica conhecida como *fine tuning*. O *fine tuning* se aproveita de modelos pré-treinados que ajudam bastante, economizando tempo e recursos de hardware, já que estes modelos, que são estado da arte, são altamente conceituados e possuem estruturas gigantescas. Treinar estes modelos do zero poderia ser uma tarefa de dias mesmo em uma GPU potente. E neste caso foram utilizados os modelos VGG16, VGG19, Xception e o InceptionV3, pois estes apresentaram os melhores resultados individuais.

São inseridos algumas camadas adicionais, a saber: camadas de preprocessamento capazes de gerar ligeiras modificações nas imagens gerando uma maior quantidade de dados,

técnica conhecida como aumento de dados (*Data augmentation*); uma camada de *dropout*, para mitigar o efeito de sobre-ajuste; e a camada *dense* de classificação final para as 8 classes.

Para realizar o *fine tuning*, inicialmente parte do modelo pré-treinado é congelado, o que impede que todos os pesos aprendidos anteriormente pela rede sejam perdidos, então a rede é treinada por algumas épocas (ciclos de treinamento), o critério para a quantidade de épocas de treinamento usada foi a *EarlyStopping* (parada precoce). Esta primeira etapa de treinamento permite que as camadas adicionais 'acordem' e consigam aprender a reconhecer as diferenças entre as emoções, utilizando para isso parte do que foi aprendido pela rede VGG19 quando treinada pelos seus criadores. Esta etapa de treinamento, conhecida como aquecimento (*warm up*), alcança um resultado ainda insuficiente. Posteriormente é descongelado todas as camadas da rede e uma nova etapa de treinamento acontece, etapa esta de refinamento que dá o nome à técnica.

A escolha dos modelos pré-treinados também foi alvo de vários testes, onde foram utilizados diversos modelos estado da arte, como o ResNet50 (HE et al., 2016), EfficientNetB0 (TAN; LE, 2019), dentre outros. Da mesma maneira as camadas adicionais também foram alvos de estudos chegando-se às melhores acurácias individuais.

3.2 Fusão dos modelos

A principal contribuição deste trabalho é a utilização da fusão dos modelos no nível de decisão, onde as previsões de cada rede são confirmadas ou refutadas pelas previsões das demais, este método aumenta o desempenho geral, já que cada rede possui características distintas e durante o treinamento adquirem capacidades de classificação diferentes.

Note que, cada modelo trabalha individualmente na classificação da imagem e a predição final se dá por um sistema de votação entre os resultados de cada modelo. O trabalho de (YÜZKAT; ILHAN; AYDIN, 2021) indica duas possibilidades para realizar esta votação, chamadas de votação forçada (*hard-voting*) e votação suave (*soft-voting*). A votação forçada usa apenas a previsão de cada rede como valor para a eleição da melhor escolha, exemplo, imagine que os modelos A, B e C indiquem uma imagem como tendo uma expressão de felicidade, já o modelo D indica a emoção de surpresa, neste caso a felicidade seria a predição final, já que possui três votos contra um. A segunda técnica é a votação suave, neste caso são somadas todas as probabilidades encontradas pelo modelo para cada uma das oito classes, um exemplo pode ser visto na Tabela 2 que classifica a imagem mostrada na Figura 10, perceba como cada modelo atribui certas probabilidades para cada uma das classes e como estes valores são somadas, a predição final será a que possuir o maior valor.

A técnica escolhida para este trabalho será a votação suave, pois além de ter apresentado melhores resultados na literatura não possui uma limitação que a votação forçada apresenta, que é a alta possibilidade de empate, e neste caso demandaria alguma técnica complementar para o

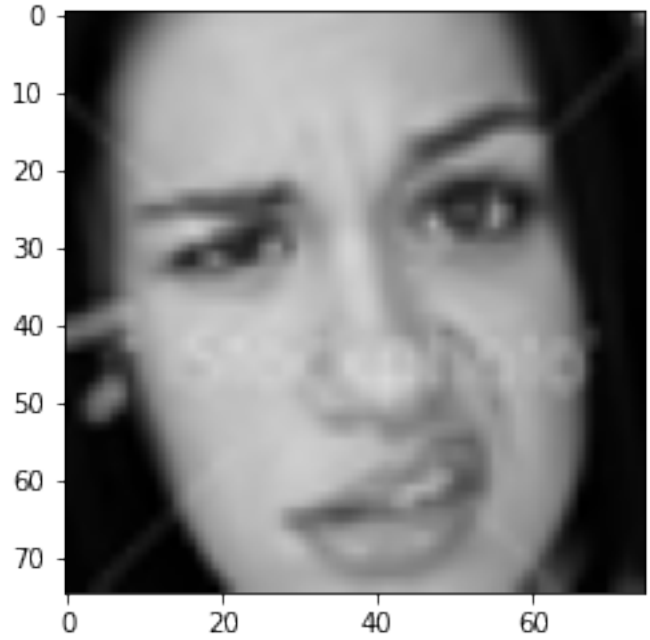


Figura 10 – Face com expressão de desprezo
Fonte: Base de dados FER2013

Tabela 2 – Análise da técnica *soft-voting-fusion*

	Modelo A	Modelo B	Modelo C	Modelo D	*soft-fusion*
Neutro	**50,5%**	**74,8%**	0,3%	2,6%	128,3
Felicidade	1,0%	0,6%	0,1%	0,0%	1,7
Surpresa	0,0%	0,1%	0,0%	0,0%	0,1
Tristeza	11,5%	14,7%	0,0%	1,7%	28,0
Raiva	0,1%	0,3%	0,0%	0,0%	0,4
Nojo	0,5%	0,2%	1,3%	2,0%	4,0
Medo	0,0%	0,1%	0,0%	0,0%	0,1
Desprezo	36,4%	9,3%	**98,2%**	**93,5%**	**237,4**
Predição	Neutro	Neutro	Desprezo	Desprezo	**Desprezo**

desempate.

3.3 Desbalanceamento entre as classes

Como foi dito anteriormente, existe uma grande diferença entre as quantidades de imagens por classe, ver Tabela 3. Isto é um problema pois o resultado do treinamento será enviesado, ou seja, ele tende a classificar os novos dados como sendo da classe que possui mais dados. Existem várias maneiras de lidar com este problema, neste trabalho foi abordado a *class_weight* (peso de classe) uma das técnicas mais eficazes para este caso. Esta técnica penaliza, durante o treinamento, o loss (erros) nas amostras de classes com o peso diferentes de 1. Ou seja, um peso de classe mais alto significa que você deseja colocar mais ênfase nesta classe. Desta maneira colocamos pesos de forma proporcionalmente maior para aquelas classes

que possuem menor quantidade de exemplos.

4 RESULTADOS

4.1 Base de dados

O banco de dados escolhido para este trabalho foi o FER2013+ que pode ser encontrado com facilidade por ser *open-source*. A Figura 11 traz alguns exemplos da versão anterior FER2013, perceba como existem algumas imagens inválidas, por causa disso o treinamento para esta base é bem desafiador e, mesmo com a análise humana o desempenho não passa dos 65,5% como pode ser visto em (GOODFELLOW et al., 2013), além das imagens inválidas também existem muitas imagens com erros de classificação. Por causa destes erros, foi escolhido para este trabalho a nova anotação chamada FER2013+ já que ela corrige as falhas anteriores, ver (BARSOUM et al., 2016), perceba na Figura 12 a diferença entre as duas classificações.

Figura 11 – Amostras do FER2013 com as respectivas expressões faciais organizadas em sete colunas: raiva, nojo, medo, felicidade, neutro, tristeza e surpresa
Fonte: Hai-Duong et al. (2019)

Esta base contém 35.887 imagens em escala de cinza e dimensões 48×48. Na FER2013+ as imagens são organizadas nas 8 emoções vistas em (MATSUMOTO; EKMAN, 2004) e também separa em uma classe específica as imagens que não podem ser classificadas e as que não representam faces, ver Tabela 3. Como pode ser notado, outro desafio desta base é o desbalanceamento entre as classes, observe a vultosa diferença entre a emoção de nojo e desprezo das demais emoções, observe também que na nova classificação as faces neutras possuem 12.906 imagens o que representa quase 36% de todas as imagens da base.

Figura 12 – Alguns exemplos dos rótulos FER2013 vs FER2013+ (FER2013 superior, FER2013+ inferior):

Fonte: Barsoum et al. (2016)- Editada pelo autor

Tabela 3 – Classes FER2013+

Classe	Qtd de imagens
Raiva	3111
Nojo	248
Medo	819
Felicidade	9355
Tristeza	4371
Surpresa	4462
Desprezo	216
Neutro	12906
Sem classificação	222
Sem face	177

4.2 Treinamento das arquiteturas de aprendizado individuais

O primeiro passo para a realização deste trabalho foi treinar os modelos individuais, cada um dos quatro modelos criados foi baseado na técnica de *fine tuning* a partir de 4 arquiteturas estado da arte. O modelo A utilizou a arquitetura VGG16, desenvolvida por (SIMONYAN; ZISSERMAN, 2014) este é um dos famosos modelos submetidos ao ILSVRC (*ImageNet Large Scale Visual Recognition Challenge*) em 2014 e atingiu 92,7% de precisão nos 5 primeiros testes no ImageNet, que é um conjunto de dados de mais de 14 milhões de imagens e 1000 classes. O modelo B utilizou a arquitetura VGG19, uma evolução do VGG16, o modelo C se beneficiou dos recursos aprendidos pela arquitetura Xception e o modelo D usou o InceptionV3.

Cada uma das redes aprende recursos distintos e conseguem reconhecer características diferentes sobre as expressões faciais, na Tabela 4 vemos a performance de cada uma das redes para cada uma das emoções além do resultado geral. Perceba como cada rede adquire

capacidades diferentes, o modelo B, por exemplo, não conseguiu reconhecer a emoção de desprezo em nenhuma das imagens do conjunto de teste. De uma maneira geral a média de precisão das 4 redes é de 80,7%, sendo que o melhor resultado individual foi do modelo A com uma acurácia de 81,4%. Note que a acurácia total é a media ponderada das acurácias por classe, já que a base de dados é desbalanceada.

Tabela 4 – Performance dos modelos individuais

Classe	Modelo A	Modelo B	Modelo C	Modelo D
Neutro	80,0%	80,0%	81,0%	78,2%
Felicidade	92,3%	89,8%	90,0%	91,8%
Surpresa	85,7%	87,3%	86,6%	82,6%
Tristeza	69,8%	66,4%	72,5%	71,9%
Raiva	79,3%	71,8%	70,2%	74,7%
Nojo	42,3%	35,3%	58,6%	50,0%
Medo	54,6%	58,9%	60,4%	64,6%
Desprezo	16,7%	0,0%	21,4%	40,0%
Total	81,4%	79,8%	81,1%	80,5%

4.2.1 Hiperparâmetros e Aumento de dados

Em todos os modelos foram utilizados a técnica de *data augmentation* para aumentar a quantidade de dados existentes, foram usados os processos de espelhamento horizontal e vertical, rotação com fator de 10% e preenchimento *wrap*, translação com fator de 10% na largura e no comprimento, normalização e contraste.

Com relação ao otimizador: Adam e SGD (Estocástico Gradiente Descendente) obtiveram bons resultados, sendo o SGD o que obteve a melhor performance. Como parâmetros utilizou-se a taxa de aprendizado inicial de 0,001 com decaimento de 0,0005 e *momentum* de 0,9.

Já o tamanho de lote (*batch size*), o melhor resultado alcançado empiricamente foi de 200. O número de épocas foi definido de forma variável utilizando, para isso, a técnica de *EarlyStopping* que observa o aprendizado e finaliza o treinamento no momento em que um parâmetro predefinido deixa de melhorar por uma quantidade de épocas estipulada. No caso, foi observado a taxa de perda (*loss*) do conjunto de validação por 18 épocas.

4.3 Fusão no nível de decisão

Em seguida foi realizado a fusão das predições particulares de cada uma dos modelos individuais, para isto foi utilizado a técnica de *soft-voting-fusion* descrita em (YÜZKAT; ILHAN; AYDIN, 2021) onde são somadas as probabilidades individuais de cada uma das redes. O que acontece é que, no caso de classificações individuais errôneas, o esquema de fusão baseado em probabilidade, *soft-voting-fusion*, pode corrigir o erro de acordo com a soma das outras

previsões da rede que classificaram a amostra com maior precisão. Na Tabela 2 vê-se um exemplo deste cálculo para a imagem da Figura 10 que apresenta uma expressão facial de desprezo. Analisando-se os resultados, pode-se notar que o modelo A e o modelo B entendem que esta face representa uma emoção neutra, embora tanto o modelo A quanto o B indiquem que existe certa probabilidade de se tratar de expressão de desprezo (36,4% e 9,3% respectivamente). A proposito, pode-se constatar que o modelo B é o que tem maior dificuldade de reconhecer a expressão de desprezo (observe na tabela 4 como este modelo não consegue reconhecer nenhuma das expressões de desprezo do conjunto de testes). Por outro lado, os modelos C e D reconheceram a expressão de desprezo com grandes probabilidades (98,2% e 93,5% respectivamente). Dessa maneira, o somatório das probabilidades garante a correta classificação da emoção.

Com a finalidade de atestar o efeito da fusão dos modelos, realizou-se três diferentes experimentos, no primeiro experimento combinaram-se os modelos A e B, no segundo experimento utilizou-se os modelos A, B e C, e por fim, no último experimento fundiram-se os quatro modelos individuais desenvolvidos neste trabalho.

4.4 Análise dos experimentos

O primeiro experimento, realizando o *soft-voting-fusion* com os modelos A e B alcançou a acurácia total de 82,4%, veja a Tabela 5, perceba que, como a base é desbalanceada as classes com maior quantidade de imagens são mais relevantes para o resultado geral que é a media ponderada dos resultados para cada classe. Note que, de maneira geral a performance melhorou em relação aos modelos individuais, embora, na comparação por emoções o resultado das emoções de raiva 77,7% e desprezo 0% tenham ficado abaixo do modelo A que foi respectivamente de 79,3% e 16,7% mas acima ou igual ao modelo B que foi respectivamente de 71,8% e 0%, já todas as outras emoções obtiveram a sua pontuação acrescida em relação aos dois modelos bases.

O segundo experimento utiliza as arquiteturas A,B e C e alcançou um resultado de 83,8% de acurácia e o terceiro experimento utilizando as predições de todos os quatro modelos individuais alcançou um resultado de 84,2%. É interessante observar que nem todas as classes obtiveram melhora, a classe de desprezo, por exemplo, teve o seu melhor resultado no segundo experimento com a fusão de três modelos.

No gráfico da Figura 13, que mostra a acurácia geral de todos os modelos individuais e dos três experimentos multimodelos realizados, é possível perceber que todos os três experimentos multimodelo se sobressaíram em relação aos modelos individuais, e que, quanto maior a quantidade de modelos associados mais eficaz se tornou a classificação. O primeiro experimento com a fusão no nível de decisão dos modelos A e B obteve um resultado 2,6% superior ao modelo B e 1% superior ao modelo A, sendo que entre os modelos isolados o A é o que mais se destacou. O segundo experimento com a fusão dos modelos A, B e C obteve resultado 2,4% superior ao modelo A e o terceiro experimento com a fusão dos modelos A, B, C e D com o resultado de 84,2% superior ao modelo A em 2,8%.

Tabela 5 – Performance multimodelo

Classe	multimodelo AB	multimodelo ABC	multimodelo $ABCD$
Neutro	80,5%	81,4%	81,5%
Felicidade	92,5%	93%	93,6%
Surpresa	87,9%	87,1%	87,6%
Tristeza	73,2%	78,9%	79,9%
Raiva	77,7%	78,4%	79,9%
Nojo	47,3%	56,5%	47,8%
Medo	64,6%	63,7%	68,1%
Desprezo	0%	40%	25%
Total	82,4%	83,8%	84,2%

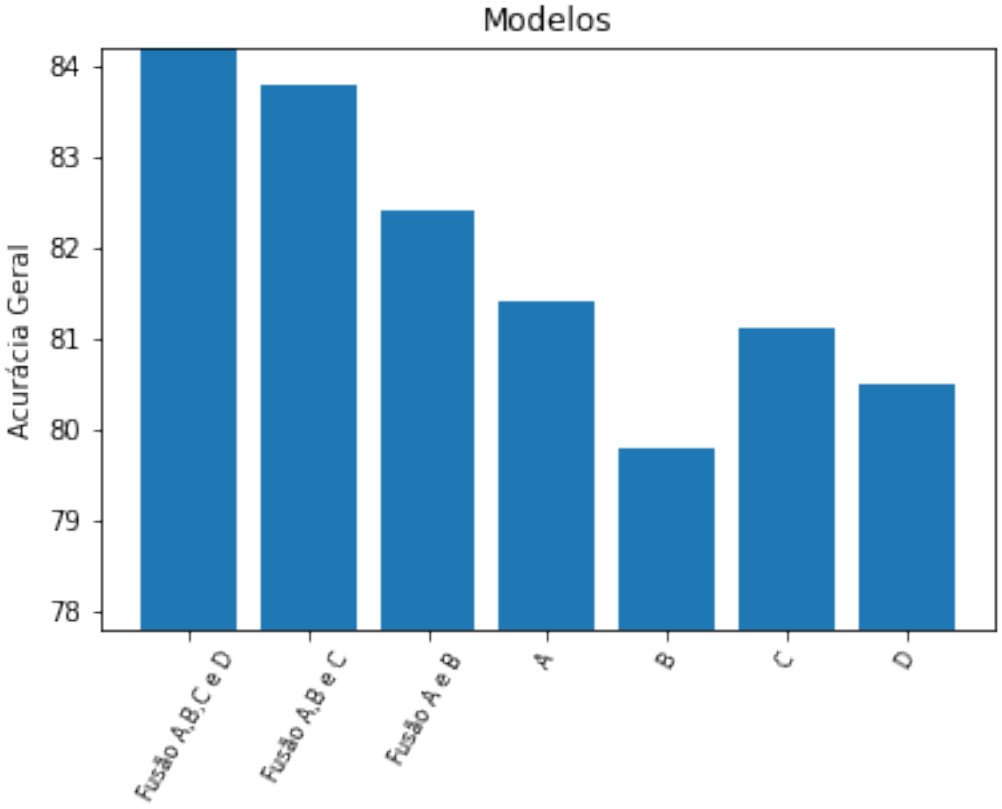


Figura 13 – Comparação entre modelos individuais e técnica multimodelo
Fonte: Elaborada pelo autor

Outro ponto a se destacar é a diferença na acurácia entre cada uma das 8 classes. Como a base de dados é desbalanceada era esperado uma diferença entre os resultados por classe, mesmo utilizando as técnicas para tratar as classes desbalanceadas, o que de fato aconteceu como pode ser visto na Tabela 5.

A emoção facial melhor classificada foi a felicidade com 93.6% de acurácia no multimodelo com 4 arquiteturas. No gráfico da Figura 14 é possível observar que todas as estratégias multimodelo se sobressaíram em relação aos modelos individuais.

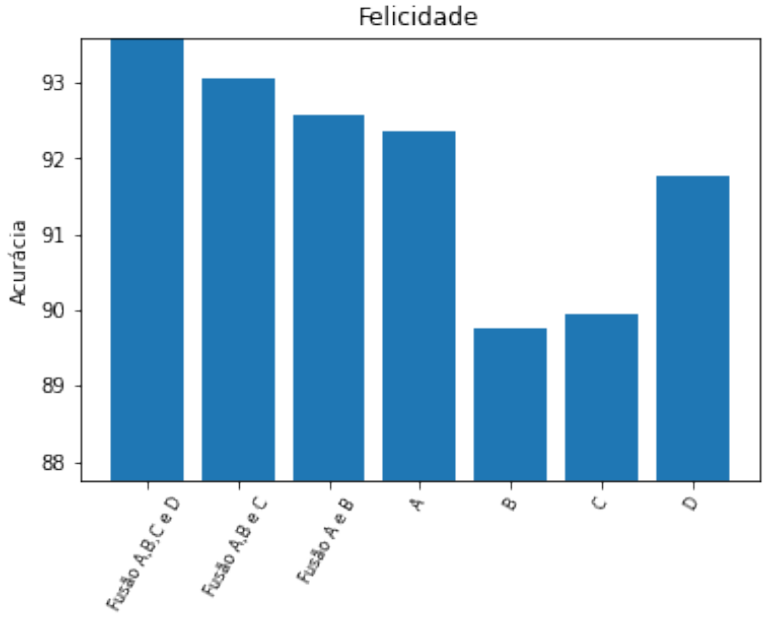


Figura 14 – Comparação entre modelos individuais e técnica multimodelo para a emoção de Felicidade

Fonte: Elaborada pelo autor

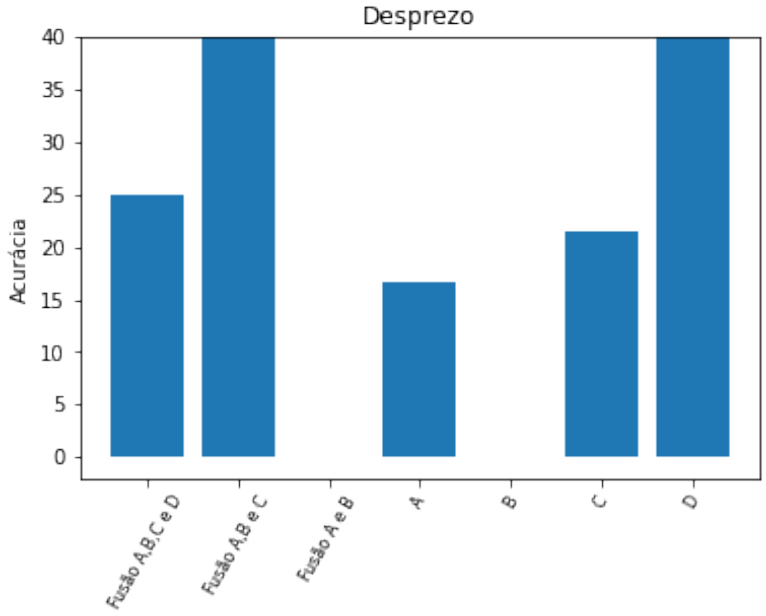


Figura 15 – Comparação entre modelos individuais e técnica multimodelo para a emoção de Desprezo

Fonte: Elaborada pelo autor

A classe que teve o resultado menor foi a emoção de desprezo, é interessante notar que esta também é a emoção mais difícil de ser identificada pelo ser humano, ver (MATSUMOTO; EKMAN, 2004). Se analisarmos o gráfico da Figura 15 observamos uma fato inesperado e que merece uma análise mais detalhada, no experimento multimodelo 2 com a fusão das arquiteturas A, B e C a acurácia foi de 40% para a emoção de desprezo, porém, quando adicionamos o

modelo D na fusão e, sabendo que o modelo D é o que melhor performou individualmente na classificação da emoção analisada, espera-se que o resultado melhore, contudo, o que acontece na prática é uma piora do resultado. Uma investigação mais profunda mostrou que o modelo D embora apresente um bom resultado na classificação não possui boa confiança no seu resultado. Observe a Tabela 6a onde uma imagem com a expressão de desprezo é erroneamente classificada pelo modelo D, porém, corretamente classificada no modelo C, no multimodelo ABC a emoção é corretamente classificada, mas adicionando o erro do modelo D no multimodelo $ABCD$ o erro do modelo D é somado aos erros dos outros modelos e a performance deteriora-se. Já na Tabela 6b observa-se um exemplo do modelo D classificando corretamente a imagem com expressão de desprezo, porém observa-se uma baixa probabilidade na classificação correta 61,8% e uma probabilidade de 27,9% de ser uma expressão de nojo, desta maneira, a classificação correta do modelo D não garante a correta classificação do multimodelo $ABCD$.

Tabela 6 – Análise da emoção de desprezo

(a) Imagem com emoção de desprezo classificada **erroneamente** no modelo D

Classe	A	B	C	D	Fusão ABC	Fusão ABCD
Neutro	**53,7%**	**39,8%**	1,2%	**64,1%**	94,7	**158,9**
Felicidade	0,2%	0,7%	0,0%	0,5%	0,9	1,4
Surpresa	0,3%	0,5%	0,2%	0,1%	1,1	1,1
Tristeza	14,2%	38,5%	0,6%	14,6%	53,4	68,0
Raiva	8,2%	16,3%	0,2%	0,2%	24,7	24,9
Nojo	3,5%	1,6%	0,4%	1,3%	5,6	6,8
Medo	1,5%	1,1%	0,0%	0,3%	2,6	2,9
Desprezo	18,3%	1,4%	**97,3%**	19,0%	**117,0**	136,0
Predição	Neutro	Neutro	**Desprezo**	Neutro	**Desprezo**	Neutro

(b) Imagem com emoção de desprezo classificada **corretamente** no modelo D

Classe	A	B	C	D	Fusão ABC	Fusão ABCD
Neutro	33,3%	4,2%	0,1%	6,0%	37,7	43,7
Felicidade	0,2%	0,0%	0,0%	0,2%	0,2	0,5
Surpresa	0,0%	0,0%	0,0%	0,0%	0,0	0,0
Tristeza	2,5%	9,2%	0,0%	0,9%	11,7	12,7
Raiva	16,6%	**64,1%**	1,4%	3,1%	82,1	85,3
Nojo	12,7%	19,1%	**97,3%**	27,9%	**129,1**	**157,0**
Medo	0,0%	0,0%	0,0%	0,0%	0,1	0,1
Desprezo	**34,6%**	3,3%	1,1%	**61,8%**	39,0	100,8
Predição	**Desprezo**	Raiva	Nojo	**Desprezo**	Nojo	Nojo

Esta dificuldade com a classificação da expressão facial de desprezo tem muita relação com a baixa quantidade de imagens categorizadas com esta classe, afinal, existem apenas 216 exemplos como pode ser visto na Tabela 3, e por consequência no conjunto de teste existem apenas 32 imagens, já que o conjunto de teste possui 15% da base. Na matriz de confusão para

o multimodelo $ABCD$ vista na Figura 16 percebe-se como a expressão de desprezo é prevista apenas em quatro momentos com apenas um acerto. Na maioria das vezes as imagens contendo expressões de desprezo são consideras como neutras.

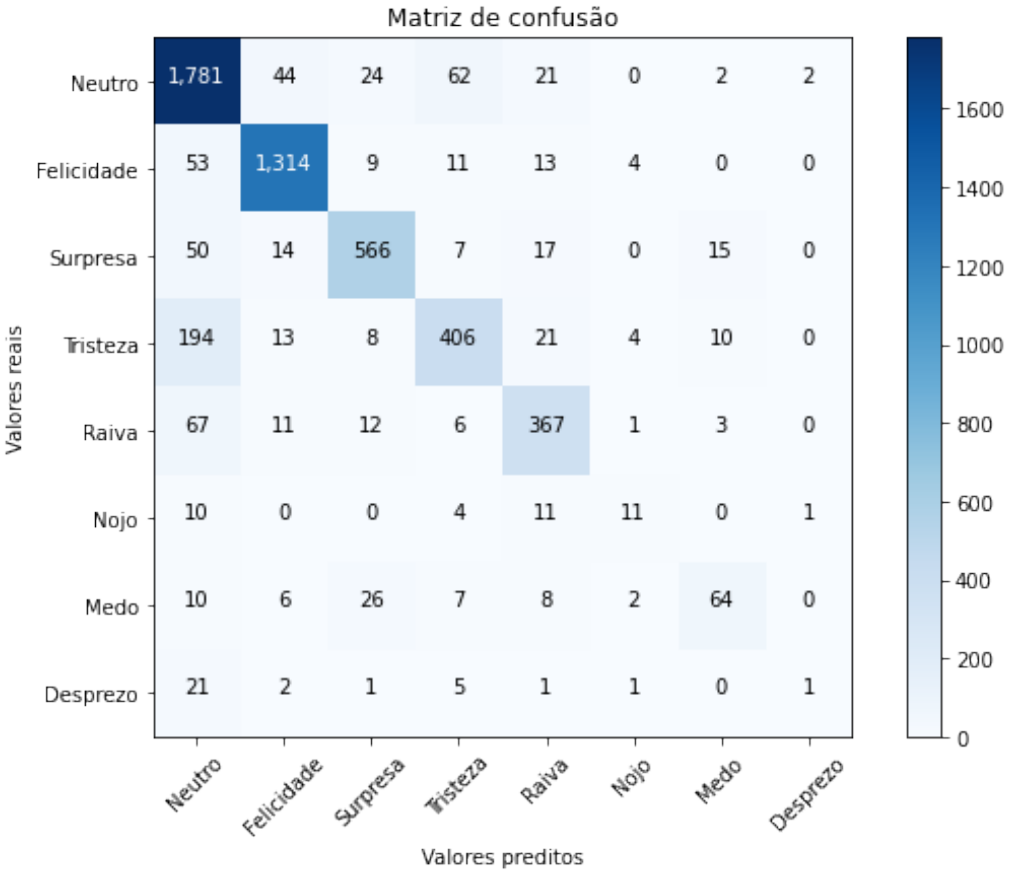


Figura 16 – Matriz de confusão para o multimodelo $ABCD$
Fonte: Elaborada pelo autor

4.5 Comparação com os melhores resultados na base de dados FER2013+

Como visto, o melhor resultado geral obtido neste trabalho foi de de 84,2% de acurácia, resultado em linha aos trabalhos que utilizam a base de dados FER2013+, ver Figura 17. Observe que, na revisão bibliográfica, os melhores trabalhos encontrados foram: ViT com um resultado de 85,05% / ResNet50 83,02% / ResNet18 81,02% vistos em (LING; LIANG; YANG, 2021), e o melhor modelo foi o desenvolvido por (SAABNI; SCHCLAR et al., 2020) com o *fine tuning* utilizando três ConvNets em paralelo obtendo um resultado de 85,1%, em linha ao trabalho desenvolvido aqui.

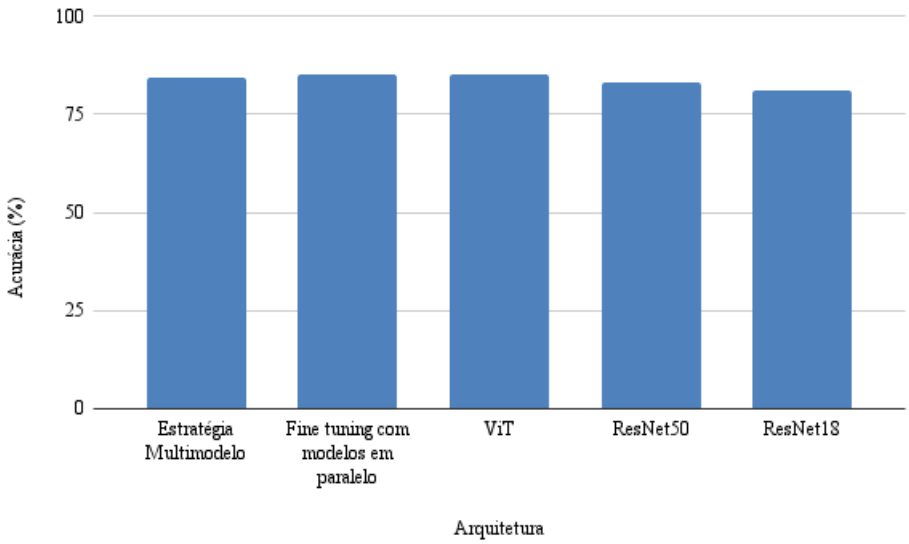


Figura 17 – Comparação de desempenho na base FER2013+
Fonte: Compilação do autor
[1]Fine tuning com modelos em paralelo - Saabni, Schclar et al. (2020)
[2]ViT - Ling, Liang e Yang (2021)
[3]ResNet50 - Ling, Liang e Yang (2021)
[4]ResNet18 - Ling, Liang e Yang (2021)

5 CONCLUSÃO

O reconhecimento de expressões faciais é, sem dúvida, muito poderoso e revolucionário. Demonstrou-se neste trabalho que as redes neurais convolucionais são um caminho de atuação para a criação de um sistema capaz de compreender as emoções humanas.

O sistema desenvolvido aqui trouxe uma acurácia de 84,2% para 8 emoções universais na base de dados FER2013+ que é uma base extremamente diversificada, com imagens de vários atores, de diversas idades, sexos, nacionalidades, além de oclusões parciais e angulação de rosto, todas estas características tornam a base difícil de ser treinada, o que aproxima o trabalho de aplicações reais, já que em aplicações reais todos os tipos de personagens com todo tipo de obstrução (como bonés, óculos, barba, dentre outros) e em todo tipo de ambiente, poderiam utilizar o sistema.

O desbalanceamento entre as classes também é uma grande dificuldade e que precisa ser melhor tratada em trabalhos futuros. A baixa quantidade de imagens contendo expressões de desprezo e nojo prejudicaram a performance na classificação destas expressões.

Contudo, a técnica multimodelo é potencialmente instigante, já que podendo ser acrescentado outros modelos na fusão ou mesmo melhorando a qualidade individual de cada modelo o resultado final pode ser substancialmente melhorado.

Embora este resultado seja promissor, é importante ressaltar que o doutor Ekman expõe a diferença entre perceber um expressão de emoção e saber o que a pessoa está pensando, ou seja, o que desencadeou aquele sentimento, e que, embora este modelo tenha demonstrado um caminho para ser seguido ainda estamos distantes de colocar os sistemas computacionais em pé de igualdade com os humanos no entendimento da linguagem natural.

REFERÊNCIAS

AGRAWAL, S.; KHATRI, P. Facial expression detection techniques: based on viola and jones algorithm and principal component analysis. In: IEEE. *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. [S.l.], 2015. p. 108–112. Citado na página 15.

ASHRAF, A. B. et al. The painful face–pain expression recognition using active appearance models. *Image and vision computing*, 2009. Elsevier, v. 27, n. 12, p. 1788–1796, 2009. Citado 2 vezes nas páginas 15 e 20.

BARSOUM, E. et al. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. [S.l.: s.n.], 2016. p. 279–283. Citado 3 vezes nas páginas 19, 28 e 29.

BENITEZ-QUIROZ, C. F.; SRINIVASAN, R.; MARTINEZ, A. M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 5562–5570. Citado na página 19.

COOTES, T. F.; EDWARDS, G. J.; TAYLOR, C. J. Active appearance models. In: SPRINGER. *European conference on computer vision*. [S.l.], 1998. p. 484–498. Citado na página 20.

CUVE, H. C. J. Expressões faciais das emoções e micro-expressões: Tendências e contributos da psicologia moderna. 2015. 2015. Citado na página 15.

David Leucas. *FACS – Medindo as emoções na face*. 2018. Disponível em: https://clue-lab.com.br/2018/01/04/facs-medindo-as-emocoes-na-face. Acesso em: 27 Julho 2021. Citado na página 14.

EKMAN, P. Expressions of emotion'. In: UNIVERSITY OF NEBRASKA PRESS. *Nebraska Symposium on Motivation*. [S.l.], 1972. v. 19, p. 207. Citado 2 vezes nas páginas 13 e 14.

EKMAN, P. Facial expressions. *Handbook of cognition and emotion*, 1999. New York, v. 16, n. 301, p. e320, 1999. Citado na página 13.

EKMAN, P. *Darwin and facial expression: A century of research in review*. [S.l.]: Ishk, 2006. 1–2 p. Citado na página 13.

FREITAS-MAGALHÃES, A. *Facial Action Coding System 3.0-Manual de Codificação Científica da Face Humana*. [S.l.]: Leya, 2020. Citado 4 vezes nas páginas 13, 14, 17 e 41.

GOODFELLOW, I. J. et al. Challenges in representation learning: A report on three machine learning contests. In: SPRINGER. *International conference on neural information processing*. [S.l.], 2013. p. 117–124. Citado 2 vezes nas páginas 19 e 28.

HAI-DUONG, N. et al. Facial emotion recognition using an ensemble of multi-level convolutional neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 2019. v. 33, n. 11, p. 1940015, 2019. Citado 2 vezes nas páginas 21 e 28.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado na página 25.

KANADE, T.; COHN, J. F.; TIAN, Y. Comprehensive database for facial expression analysis. In: IEEE. *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. [S.l.], 2000. p. 46–53. Citado na página 17.

KHAIREDDIN, Y.; CHEN, Z. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint arXiv:2105.03588*, 2021. 2021. Citado 2 vezes nas páginas 15 e 21.

LING, X.; LIANG, J.; YANG, J. A self-attention based method for facial expression recognition. In: *2021 7th International Conference on Computing and Artificial Intelligence*. [S.l.: s.n.], 2021. p. 191–195. Citado 3 vezes nas páginas 22, 35 e 36.

LIU, C.; WECHSLER, H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 2002. IEEE, v. 11, n. 4, p. 467–476, 2002. Citado na página 20.

LUCEY, P. et al. A complete dataset for action unit and emotion-specified expression. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. [S.l.: s.n.], 2010. p. 94–101. Citado 2 vezes nas páginas 17 e 18.

LYONS, M. J. et al. The japanese female facial expression (jaffe) database. In: *Proceedings of third international conference on automatic face and gesture recognition*. [S.l.: s.n.], 1998. p. 14–16. Citado na página 17.

MATSUMOTO, D.; EKMAN, P. The relationship among expressions, labels, and descriptions of contempt. *Journal of personality and social psychology*, 2004. American Psychological Association, v. 87, n. 4, p. 529, 2004. Citado 4 vezes nas páginas 13, 17, 28 e 33.

MAVADATI, S. M. et al. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 2013. IEEE, v. 4, n. 2, p. 151–160, 2013. Citado na página 18.

MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. ACM New York, NY, USA, v. 38, n. 11, p. 39–41, 1995. Citado na página 19.

MOHAMMADI, M. R.; FATEMIZADEH, E.; MAHOOR, M. H. Pca-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, 2014. Elsevier, v. 25, n. 5, p. 1082–1092, 2014. Citado na página 20.

MOLLAHOSSEINI, A.; CHAN, D.; MAHOOR, M. H. Going deeper in facial expression recognition using deep neural networks. In: IEEE. *2016 IEEE Winter conference on applications of computer vision (WACV)*. [S.l.], 2016. p. 1–10. Citado 2 vezes nas páginas 15 e 21.

PICARD, R. W. Affective computing for hci. In: CITESEER. *HCI (1)*. [S.l.], 1999. p. 829–833. Citado na página 15.

RINGEVAL, F. et al. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In: *Proceedings of the 23rd ACM international conference on Multimedia*. [S.l.: s.n.], 2015. p. 1335–1336. Citado na página 17.

SAABNI, R.; SCHCLAR, A. et al. Facial expression recognition using combined pre-trained convnets. *Comput. Sci. Inf. Technol.*, 2020. v. 95, p. 95–106, 2020. Citado 3 vezes nas páginas 22, 35 e 36.

SCHULLER, B. et al. Avec 2011–the first international audio/visual emotion challenge. In: SPRINGER. *International Conference on Affective Computing and Intelligent Interaction*. [S.l.], 2011. p. 415–424. Citado na página 17.

SHAN, C.; GONG, S.; MCOWAN, P. W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 2009. Elsevier, v. 27, n. 6, p. 803–816, 2009. Citado na página 20.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2014. Citado na página 29.

SZEGEDY, C. et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 1–9. Citado na página 21.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2019. p. 6105–6114. Citado na página 25.

VALSTAR, M. F. et al. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In: *Proceedings of the 8th international conference on Multimodal interfaces*. [S.l.: s.n.], 2006. p. 162–170. Citado na página 17.

YÜZKAT, M.; ILHAN, H. O.; AYDIN, N. Multi-model cnn fusion for sperm morphology analysis. *Computers in Biology and Medicine*, 2021. Elsevier, v. 137, p. 104790, 2021. Citado 3 vezes nas páginas 24, 25 e 30.

ANEXO A – OS 102 CÓDIGOS DA F-M FACS 3.0

Fonte : (FREITAS-MAGALHÃES, 2020)

Action Units

Face Superior

1	Elevação da parte interna da sobrancelha
2	Elevação da parte externa da sobrancelha
3	Depressão do ângulo médio das sobrancelhas, contração da glabela
4	Baixar as sobrancelhas
5	Elevação da pálpebra superior
6	Levantamento das bochechas
7	Tensão das pálpebras
8	Contração do temporal
41	Queda das pálpebras
42	Estreitamento da abertura das pálpebras
43	Olhos fechados
44	Estrabismo
45	Piscar
46	Piscadela
47	Dilatação da pupila
48	Contração da pupila

Face Inferior

9	Franzimento do nariz
10	Elevação do lábio superior
11	Acentuação da prega nasolabial
12	extrator do canto do lábio
13	Ascensão e inchaço das bochechas
14	Retração dos lábios e estreitamento das comissuras
15	Diminuição do ângulo da boca
16	Depressão do lábio inferior
17	Elevação do queixo

18 Contração dos lábios e arredondamento fechado em frente da boca
19 Contração dos lábios e arredondamento aberto em frente da boca
20 Estiramento horizontal dos lábios
22 Lábios em posição de funil
23 Contração dos lábios
24 Apertar os lábios
28 Sucção dos lábios para dentro
38 Dilatação das narinas
39 Contração das narinas

Abertura do maxilar e separação dos lábios

25 Separação dos lábios
26 Queda do mento
27 Abertura da boca

Pescoço

21 Tensão no pescoço
36 Contração e flexão do esternomastoideo
37 Contração do esternotireóideo

Action descriptors

29 Elevação da mandíbula
30 Mandíbula para o lado
31 Apertar a mandíbula
32 Morder
33 Soprar
34 Bufar
35 Sucção das bochechas

Movements

Movimento de cabeça

51 Virar para a esquerda
52 Virar para a direita
53 levantar a cabeça
54 Baixar a cabeça
55 Inclinar para a esquerda
56 Inclinar para a direita
57 Cabeça para a frente, precedido da combinação 17 + 24
58 Para trás
59 Cabeça para a frente e para trás, precedido da combinação 17 + 24
60 Cabeça da esquerda para a direita ou vice versa, precedido da combinação 17 + 24
83 Cabeça para a cima e virada para a esquerda ou para a direita. Ocorre em conjunto com AU14.

Movimento dos olhos

61 Olhos para a esquerda
62 Olhos para a direita
63 Olhos para a cima
64 Olhos para a baixo
65 Olhos em direções opostas
66 Olhos cruzados
68 Olhos para cima, para o lado e volta à posição fixa. Ocorre em conjunto com AU14
69 Olhar fixo em conjunto com as AUs 4,5,7 e 14 isolados ou em combinação

Movimento de língua

100 Mostrar a língua interna
101 Língua interna esquerda
102 Língua interna direita
103 Mostrar a língua externa
104 Língua externa esquerda
105 Língua externa direita
106 Língua externa para cima
107 Língua externa para baixo
108 Língua interna enrolada para cima
109 Língua interna enrolada para baixo
110 Língua interna dobrada
111 Língua externa enrolada
112 Língua externa enrolada para cima
113 Língua externa enrolada para baixo

114 Língua externa dobrada
115 Morder a língua
116 Limpar lábio superior com a língua da esquerda para a direita
117 Limpar lábio superior com a língua da direita para a esquerda
118 Limpar lábio inferior com a língua da esquerda para a direita
119 Limpar lábio inferior com a língua da direita para a esquerda
120 Protuberância da bochecha esquerda com a língua
121 Protuberância da bochecha direita com a língua

Gross Behavior

40 Cheirar
50 Discursar
80 Engolir
81 Mastigar
82 Encolher os ombros
84 Mexer a cabeça de um lado para o outro
85 Mexer a cabeça para cima e para baixo
91 Flash
92 Flash parcial

Visibility Codes

70 Sobrancelhas não visíveis
71 Olhos não visíveis
72 Face inferior não visível
73 Face inteira não visível
74 Não é possível codificar