


# Classificação De Emoções A Partir De Expressões Faciais com Deep Learning

Caroline Guerra and Carine Geltrudes Webber 

**Abstract**—Students with negative emotional states have greater learning difficulties, which impacts their school performance and can further worsen their emotional state, becoming a vicious cycle. Identifying emotions in the classroom can help in identifying learning difficulties. In this context, diagnostic systems, based on machine learning (*machine and deep learning*), have been used to identify cases of anxiety, distress and depression. Such diagnoses are possible by obtaining classification models, using learning mechanisms. As an example, a model based on a *deep learning* architecture, that identifies and classifies students' emotions, can help the teacher with *feedback* about their class. For the purposes of this conclusion work, the objective is to design and implement a classification model based on *deep learning* methods for classifying emotions through images of human faces. This would be the first step towards the development of a system that can infer difficulties in learning in the classroom, thus enabling the educator to approach new teaching strategies that can improve the performance of the class and their emotional states.

**Index Terms**—Artificial intelligence. Deep Learning. Emotions Classification. Depression in Classroom.

## I. INTRODUÇÃO

NA sociedade atual, o uso de tecnologias que auxiliem o ser humano no seu dia a dia estão sendo cada vez mais requisitadas. Percebe-se isso de diversas formas, desde os aplicativos de *delivery* e transporte até as melhorias nos equipamentos cirúrgicos e de saúde [1]. A tecnologia assume um papel fundamental na forma como se vive, aumentando a expectativa e a qualidade de vida humana.

Este cenário objetiva facilitar tarefas do cotidiano e melhorar o estado emocional dos seres humanos, garantindo seu bem estar. As emoções, que caracterizam o estado emocional, são respostas do organismo frente à situações, sejam elas reais ou imaginadas. O que torna uma pessoa emotiva difere de cada um, estando diretamente ligado a experiências passadas, culturais e evolutivas [2]. Contudo, estados emocionais negativos podem desencadear diversos problemas relacionados a saúde mental e a comportamentos problemáticos [3].

Saúde mental refere-se ao bem estar emocional, psicológico e social de um indivíduo. Problemas relacionados a saúde mental como estresse, ansiedade e depressão afetam a forma de agir e pensar, podendo acarretar em sérios riscos para a saúde [4]. No caso da depressão, a pessoa vivencia sentimentos como tristeza, sensação de vazio, falta de prazer e interesse

em atividades, entre diversos outros sintomas, podendo levar até mesmo ao suicídio [5]. De acordo com a Organização Mundial da Saúde, o primeiro ano da pandemia do COVID-19 aumentou em 25% os casos de permanência de ansiedade e depressão em todo o mundo [6].

Muitos problemas de saúde mental tem origem no ambiente escolar, seja em decorrência de estresse, dificuldade de aprendizagem, desmotivação por não corresponder com expectativas, entre outros fatores [7]. De acordo com Camargo, "a emoção constitui função inseparável da cognição e da aprendizagem" [8]. Uma aprendizagem correta não só faz com que o aluno aprenda a matéria, mas contribui para a regulação emocional, conhecimento consciente e estratégias cognitivas. Ambientes onde o aluno se sente com medo, ameaçado, inseguro ou desconfortável, afetam negativamente as funções de memória e tomada de decisão. Em outras palavras, um indivíduo aprende melhor quando possui uma ligação emocional com o que está sendo estudado, buscando atividades que lhe causem uma sensação de bem estar e evitando aquelas que lhe deixam mal [7], [9].

Torna-se evidente como o desempenho em sala de aula afeta a vida de um aluno, podendo comprometer não apenas suas notas mas sua saúde mental e futuro. A tecnologia surge nesse cenário como uma aliada a educação. A detecção de emoções em sala de aula pode ajudar o professor com *feedback* sobre o estado emocional dos alunos e da turma em geral, contribuindo para que este auxilie com estratégias de ensino e melhore o aprendizado da classe.

Com o avanço da área da inteligência artificial (IA), muitas tecnologias surgiram com a capacidade de aprender e lidar com uma enorme quantidade de dados e padrões. Dentre elas, as rede neurais artificiais conhecidas como *deep learning* se tornaram bastante conhecidas. Os modelos *deep learning* se baseiam em algoritmos inspirados no funcionamento do cérebro humano, sendo capazes de aprender e realizar tarefas de reconhecimento, classificação, predição entre outras. Assim, detecção de emoções está entre uma das muitas possibilidades de uso para estas redes. Elas podem utilizar diversos meios para a identificação, sejam por texto, voz, imagens, sinais fisiológicos, encefalogramas, questionários e vários outros. O reconhecimento de emoção por fala de Aouani e Ayed [10], emoção em textos de redes sociais de Peng et al. [11], sinais depressivos em redes sociais de Nascimento et al. [12] e até mesmo por respostas galvânicas de pele de Domínguez-Jiménez et al. [13] são apenas alguns dos muitos exemplos de aplicação existentes.

Este trabalho visa desenvolver um modelo de *deep learning* para auxiliar professores e alunos na tarefa de ensino-

Caroline Guerra, Área de Exatas e Engenharias, Universidade de Caxias do Sul (UCS), Caxias do Sul, Rio Grande do Sul, Brazil, e-mail: cguerra3@ucs.br

Carine Geltrudes Webber, Área de Exatas e Engenharias, Universidade de Caxias do Sul (UCS), Caxias do Sul, Rio Grande do Sul, Brazil, e-mail: cgwebber@ucs.br

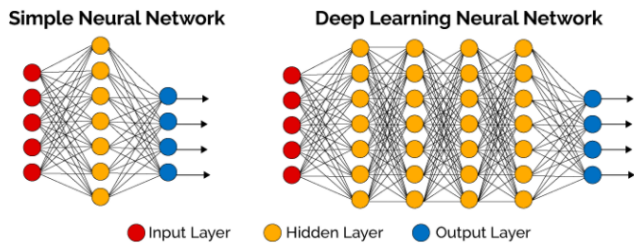


Figura 1: Exemplo de Rede Neural Simples e Redes Neural Artificial Profunda.

aprendizagem. Para isso, o modelo baseado em *deep learning* deve ser capaz de classificar as imagens em uma das sete categorias de emoções definida por Paul Ekman: felicidade, raiva, nojo, desprezo, medo, tristeza e surpresa. De acordo com o autor, de todo o espectro de emoções humanas estas sete são universais para todas as culturas ao redor do mundo, independentemente de outros fatores envolvendo as culturas específicas [2]. O objetivo é classificar tais emoções, por meio da análise de imagens expressões faciais, e aplicar o modelo em imagens de alunos dentro de uma sala de aula real, fornecendo *feedback* apropriado sobre os resultados obtidos para os professores, a fim de proporcionar reflexões sobre os estados emocionais da turma.

## II. DEEP LEARNING

*Deep learning* também são conhecidas como redes neurais artificiais profundas. Elas diferem das outras redes neurais artificiais por apresentarem mais de três camadas ocultas, o que dá a elas a capacidade de resolverem problemas mais complexos. Seu processamento de dados ocorre de forma similar ao cérebro humano, podendo realizar classificações, previsões, reconhecimento de padrões entre diversas outras tarefas. As *deep learnings* têm a capacidade de aprender sem intervenção humana, compreendendo por si só quais características extrair dos dados e, em casos de classificação, quais são as informações mais importantes que diferenciam um elemento do outro [14], [15].

Elas são organizadas em camadas, cada uma composta por diversos neurônios artificiais funcionando como um algoritmo. Os neurônios, também chamados de nós, estão interconectados uns com os outros e tal arquitetura diferencia uma rede neural da outra. No caso das redes neurais *feed-forward* a informação segue um único fluxo: da camada de entrada para a de saída. A Figure 1 [14] mostra uma comparação entre um exemplo de rede neural simples e uma *deep learning feed-forward*. Cada neurônio conecta com todos os da camada seguinte, assim sucessivamente até a camada de saída. Para as redes neurais simples, há apenas uma camada oculta, já na *deep learning* da imagem são quatro. Nela fica evidente como os nós são conectados de camada em camada, a camada anterior servindo de entrada para a camada seguinte [16].

As camadas intermediárias, aquelas localizadas entre as de entrada e saída, são chamadas de camadas ocultas. Sua quantidade varia para cada situação, servindo para otimizar a precisão da rede neural e torna-lá capaz de lidar com

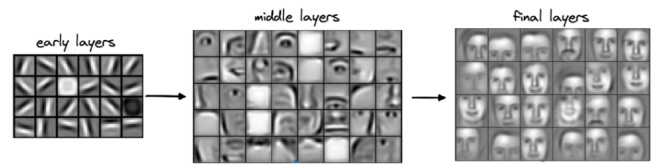


Figura 2: Exemplo de Extração de Características por Camadas.

problemas mais difíceis. Um desses casos é o de classificação de visão computacional, a Figure 2 [17] traz um exemplo de detecções de imagens de uma rede neural. Observa-se nas primeiras camadas que elementos mais simples, como as bordas das imagens, são extraídos primeiro. Conforme a rede avança, os elementos são combinados e se tornam mais complexos: nas camadas ocultas do exemplo as bordas foram combinadas com partes do rosto, pegando características como olhos ou narizes. Já nas últimas camadas fica evidente o objetivo da rede neural: a de classificação de rostos [17].

Existem diversos tipos de *deep learnings*, cada modelo apresentando uma arquitetura diferente e sendo usados em problemas diferentes. A estrutura, a forma como é treinada e as operações que realiza, tudo isso faz com que cada rede neural trabalhe de forma única umas das outras, obtendo melhor desempenho em um tipo de problema que em outro. No caso de visão computacional, as redes neurais convolucionais são as mais utilizadas por serem eficientes em classificação de imagens [18], [19].

### A. Redes Neurais Convolucionais

Uma rede neural convolucional, em inglês *convolutional neural networks* (CNN), é uma *feed-forward* que possui diversas camadas ocultas inspiradas no córtex visual humano. Os neurônios desta área do cérebro são muito sensíveis ao campo visual, respondendo a estímulos como bordas, orientação vertical ou horizontal de objetos. Isso faz com que a rede seja amplamente utilizada em problemas de visão computacional, obtendo alta acurácia em problemas de classificação de imagens e vídeos [14], [20]–[22].

Para as CNNs, imagens são entendidas como matrizes: a altura e largura das matrizes são correspondentes à da imagem original, no caso da Figure 3 [20] seriam de quatro *pixels* de altura e quatro de largura. Para entender as cores, uma CNN separa os valores por RGB. No sistema RGB, uma cor é feita da combinação de vermelho, azul e verde, onde cada um destes têm um valor entre 0 à 255 [23]. Uma imagem vermelha, por exemplo, tem valor 255 para vermelho e zero de azul e verde; Uma amarela poderia ter zero para azul e 255 para vermelho e verde. A CNN separa estes valores, criando uma matriz para cada cor, como mostra a Figure 3. Ela pega o primeiro pixel da imagem e separa os valores nas três matrizes na primeira posição, depois o segundo pixel com os valores correspondentes na segunda posição e assim por diante. Essas matrizes são chamadas de canais de cor [14], [20].

Contudo, existem diversas etapas em uma CNN para que ela consiga identificar o conteúdo de uma imagem: a imagem deve

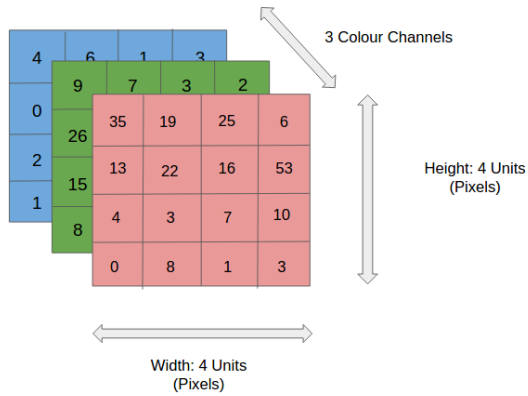


Figura 3: Exemplo de Matrizes de uma Imagem.

passar por um processo chamado convolução, depois *pooling*, então para as camadas totalmente conectadas, enfim classificando o elemento. Cada etapa é essencial para o processo, muitas vezes repetindo ou intercalando passos de convoluções e *pooling* para melhor eficiência [21].

Um elemento nem sempre está claramente visível em uma imagem. Em muitos casos, principalmente os do mundo real, os elementos podem se encontrar parcialmente ocultados, mesclados ao fundo, com diferentes iluminações, escalas, ângulos de vista e diversas outras situações adversas. Nestes casos, as imagens podem facilmente ser mal classificadas pela falta de clareza do elemento. Em decorrência disso, se torna muito importante a utilização de filtros, também conhecidos como *kernels*, cuja função é extrair ou ampliar certos fatores da imagem para torná-los mais evidentes. Alguns filtros podem destacar linhas verticais, outros remover bordas de objetos, ampliar, desfocar entre muito outros, como mostra o exemplo da Figure 4 [24], que apresenta as matrizes para detecção de bordas, *sharpen* para deixar a imagem mais nítida, *box blur* e *Gaussian blur* para desfoque, assim como uma imagem correspondente mostrando o resultado do efeito aplicado. O processo de aplicação dos filtros se chama convolução, que nomeia a primeira camada da rede [25].

Na convolução, a matriz do filtro passa sobre a imagem fazendo o produto dos valores, como exemplificado na Figure 5 [26], colocando o resultado da operação em uma nova matriz chamada de *feature map*. Na inicialização de uma rede neural, são atribuídos valores aleatórios para o filtro, a fim de deixar a rede aprender por conta própria, corrigindo os erros por *backpropagation*. O filtro se movimenta pela imagem em um determinado *stride*, que é a quantidade de *pixels* que anda de cada vez. No exemplo da figura, o *stride* tem valor 1, ou seja, vai passar de uma operação a outra indo um *pixel* para frente. Esse movimento normalmente é feito da esquerda para direita e de cima para baixo. Consequentemente, o *feature map* não terá o mesmo tamanho da matriz inicial, ficando reduzido em comparação com o original. Em casos onde haja necessidade do tamanho da saída ser igual ao de entrada, utiliza-se o *padding*. Ele vai adicionar camadas extras nas laterais da imagem, preenchendo esse espaço com o valor zero, a fim de manter o tamanho original na saída [21], [24], [25].

Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur 3 × 3 (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

Figura 4: Exemplo de Filtros.

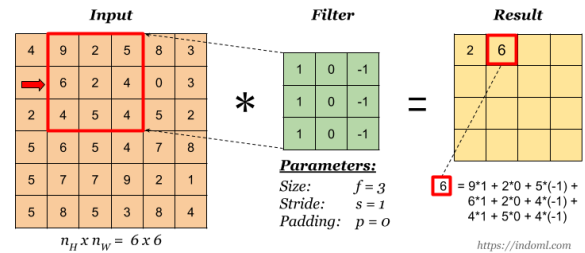
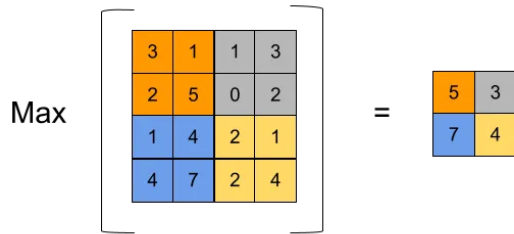


Figura 5: Operação de Convolução.

Cada filtro vai extrair diferentes características da imagem, criando um *feature map* próprio. Alguns podem armazenar os contornos do objeto, outros já selecionam elementos mais específicos como partes do rosto e assim por diante. Quanto mais profunda a camada, mais complexo o padrão [27].

Etapas de convolução são geralmente sucedidas da aplicação da função de ativação. Ela garante a não-linearidade do sistema, tornando possível que a rede aprenda problemas complexos [20], [21].

Após extraídos os *feature maps* segue-se para a etapa de *pooling*. O objetivo do *pooling* é reduzir a quantidade de informação, simplificando o escopo da camada anterior. Isso pode ser feito de diversas maneiras, a mais comum sendo o *Max Pooling*. Para tal, é definida uma área a percorrer toda a matriz, parecido com a forma que se aplicam os filtros. Neste espaço, e no caso do *Max Pooling*, é selecionado o maior valor entre os demais, passando ele à outra matriz. O objetivo é percorrer todas as quadrantes da matriz, normalmente diminuindo seu

Figura 6: Exemplo de *Max Pooling*.

tamanho pela metade. A Figure 6 [28] traz um exemplo de *Max Pooling*, onde a matriz original de 4x4 reduziu à 2x2 [20], [24], [25], [28].

Por fim, os dados vão para a camada totalmente conectada. Ela recebe a entrada e gera um vetor, cujo tamanho é o número de classes que o programa tem que escolher. Supondo um programa que classificasse dígitos, ele teria dez saídas, cada uma representando um dos dígitos de 0 à 9. Cada saída terá um valor representando a probabilidade de ser tal resposta. Utilizando o exemplo da Figure 6, caso o objetivo fosse classificar corretamente o número “6”, o resultado poderia ter uma saída alta indicando “6”, mas também probabilidades mais baixas para o “0” ou o “5”, por exemplo. A CNN consegue essa identificação analisando e determinando quais características se relacionam mais com determinada classe. Se o objetivo for a identificação do número “6”, ela terá altos valores nos mapas de ativação que possuam características como um círculo ou um risco superior por exemplo. Calculando o produto dos pesos com as camadas onde se acha mais relação, ela obtém as probabilidades mais aproximadas [21].

### B. Métricas de Avaliação

Como saber se uma rede neural está de fato produzindo resultados satisfatórios? Utilizando-se as métricas de avaliação. Estas métricas fornecem informações analisando os resultados da rede, que são usados para validar a eficácia da mesma. Para melhor entendimento, considera-se uma matriz de confusão, que é uma tabela disposta da quantidade de acertos e erros da rede, como mostra a Figure 7 [29]. Nela se encontra quatro tipos de resultados: Verdadeiro Positivo (VP) são casos onde a rede classificou corretamente os dados, como por exemplo, identificou carros em imagens que realmente possuíam carros; Verdadeiro Negativo (VN) classificou corretamente um caso negativo, como não identificar carros em imagens que realmente não possuíam carros; Falso Positivo (FP) é quando a rede classifica positivamente um resultado negativo, como a presença de um carro quando não há nenhum; e por fim, Falso Negativo (FN) são resultados classificados como negativos mas deveriam ser positivos [29].

Em problemas de múltiplas classes, a matriz segue o mesmo princípio, mas funciona um pouco diferente: a Figure 8 [30] traz um bom exemplo desse funcionamento. Nele, são três classes que o algoritmo pode classificar o objeto: como sendo uma maçã, laranja ou manga. Na horizontal são as classes verdadeiras, aquelas que as imagens realmente pertencem, e

		Detectada	
		Sim	Não
Real	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 7: Matriz de Confusão.

		True Class		
		Apple	Orange	Mango
Predicted Class	Apple	7	8	9
	Orange	1	2	3
	Mango	3	2	1

Figura 8: Matriz de Confusão Multi-classe.

na vertical as previsões do modelo. Assim, quando os nomes se encontram nos quadrados, em verde, são as classificações corretas, ou seja, aqueles que o modelo previu e acertou na classificação. Já os outros, são as previsões erradas do modelo. No exemplo da maçã, se somarmos todos os valores da sua coluna correspondente teremos a quantidade de instâncias de maçã que o *dataset* possuía, neste caso 11. Se somarmos pela linha, saberemos quantas maçãs o modelo previu que tinha, neste caso 24. Assim, sabemos que ele indicou 8 laranjas e 9 mangas como sendo maçãs. Dessa forma, fica simples verificar as classificações do modelo para qualquer classe, tantos os certos como os erros, visualizando em que classe ele os classificou.

As métricas utilizam os dados da matriz de confusão em diversos cálculos com diferentes objetivos. Acurácia informa o percentual de acertos da rede neural. Uma acurácia de 92% indica que a rede acertou 92 a cada 100 casos, indicando sua performance geral [14], [29]. Seu cálculo é feito somando os Verdadeiros Positivos e Negativos, aqueles classificados corretamente pela rede, divididos pela soma total dos resultados da seguinte forma:

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN}$$

De forma lógica, o erro é a porcentagem de casos que não foram acertados pela rede neural:

$$\text{Erro} = 1 - \text{Acurácia}$$

De outra forma, a precisão informa de todos que foram classificados como Positivos quais foram classificados corre-

tamente. É mais utilizado quando os Falsos Positivos são mais prejudiciais que os Falsos Negativos, como ao identificar quais negócios são bons investimentos: vale mais a pena não investir do que investir em uma empresa que não dará lucro [29]. Utiliza-se a seguinte fórmula:

$$\text{Precisão} = \frac{VP}{VP + FP}$$

*Recall* é muito semelhante à precisão. Por sua vez, é utilizado em casos onde os Falsos Negativos são mais prejudiciais que os Falsos Positivos. A exemplo, é preferível que a rede classifique alguns pacientes saudáveis como doentes do que o contrário [14], [29]. Calcula-se os acertos por todos os casos que devem ser Positivos, ou seja, levando em conta os Verdadeiros Positivos pela soma destes com os Falsos Negativos:

$$\text{Recall} = \frac{VP}{VP + FN}$$

Embora haja semelhança entre precisão e *recall*, são métricas que avaliam comportamentos diferentes dos resultados, ambos sendo importantes para verificar a eficiência da rede neural. Com isso em vista, o *F1-Score* faz uma média harmônica destes valores, a fim de se observar apenas uma métrica. Em médias harmônicas se o valor de alguma das variáveis está baixa acaba fazendo com que o valor do resultado final diminua significativamente. Ou seja, um *F1-Score* ruim é indicativo de que ou a precisão ou o *recall* está com valor baixo [29]. Lê-se a seguinte equação:

$$F1\text{-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

### III. TRABALHOS RELACIONADOS

O estado da arte é importante para examinar trabalhos relacionados, verificando o que já foi feito na área de pesquisa e contribuindo para o desenvolvimento do projeto. Para tal, foi realizada a revisão sistemática de trabalhos similares com o do tema proposto, dos quais tiveram seus resultados dispostos tabelas e gráficos comparativos de informações relevantes.

#### A. Revisão Sistemática

Para entender o tema de reconhecimento de emoções a partir de expressões faciais, foi realizada uma revisão de trabalhos na plataforma *Science Direct*<sup>1</sup>. Ela consta com um acervo de diversas pesquisas científicas, médicas e tecnológicas, ainda tendo disponibilidade de busca por tópicos e filtros de pesquisa. Os termos pesquisados foram:

- *Deep learning, neural net, facial/emotion recognition, anxiety, depression, image processing, mental health, FER.*

E os filtros aplicados:

- Acesso: *Open access*;
- Ano: 2020, 2021 e 2022;
- Área: *Computer science*.

A pesquisa gerou mais de 100 resultados, onde selecionou-se nove que tratavam do tema de reconhecimento de emoções

por expressões faciais. Destes selecionados para leitura, dois não tratavam dos métodos para análise de emoções a partir de imagens ou que combinavam diversos métodos, como análise de voz ou encefalograma, ficando-se apenas com sete trabalhos para estudo.

#### B. Artigos Relacionados

No trabalho de Li e Lima [31], é proposto um modelo melhorado para o reconhecimento de emoções utilizando a CNN *ResNet-50*. As imagens utilizadas como *dataset* foram tiradas por um fotógrafo, garantindo que utilizassem pessoas com diferentes idades e raça, além de incluir sete tipos de expressões faciais: felicidade, surpresa, medo, raiva, tristeza, desprezo e neutro, totalizando 700 fotos. As imagens foram tratadas com convolução para extrair os recursos da imagem e os resultados agrupados com o método *pooling*. Depois de ter tratado as imagens, a rede neural foi treinada com o *Batch Normalization* e, enfim, utilizou-se a rede neural com um *cross-validation* de dez grupos com dez imagens para cada emoção. Ao final, a melhor acurácia da rede foi de  $96.40 \pm 1.84\%$  para expressões neutras e o pior resultado sendo  $94.10 \pm 4.15\%$  para tristeza.

Ivanova e Borzunov [32] optaram por uma rede neural que utilizasse o método de Viola-Jones, que escaneia a imagem e alimenta ou não um classificador, indicando a presença ou ausência de um objeto. Depois, aplica-se diversos filtros e caso estes gerem resposta positiva, então houve reconhecimento de um rosto. Utilizaram o banco de dados *OLR* para a detecção, que consta com 400 imagens de diferentes rostos com iluminações e expressões faciais diferentes. A rede utilizada foi a *GoogleLeNet* com o acréscimo de um módulo chamado *Inception*, que combina diversos filtros convolucionais. A rede foi treinada com o banco *FER2013* da *Kaggle*<sup>2</sup>, onde um programa feito em *Python*<sup>3</sup> utilizou validação cruzada nas amostras. O teste público teve acurácia de 69% dos casos e o privado 59%. Individualmente por emoções, a maior taxa de reconhecimento foi de 88% para felicidade e a menor 57% para tristeza.

Como explicam Lebedev et al. [33], ainda não existem métodos de reconhecimento facial que possa dizer o estado psicoemocional de um ser humano. A proposta por eles enunciada, refere-se ao uso de um sistema de reconhecimento de expressões faciais que inicialmente “guarde” uma imagem de cada emoção da pessoa, já que cada ser humano experiência e expressa emoções de formas diferentes. Essa imagem seria usada como base de comparação do sistema e então, ele analisaria as expressões faciais da pessoa por semanas ou até mesmo meses, fazendo um comparativo que, ao final, quantificasse quanto tempo o sujeito está com o mesmo estado emocional. Em casos onde a pessoa esteja por um longo período de tempo com emoções negativas, poderia-se supor que ela esteja com algum problema de saúde mental e se recomendaria a buscar ajuda de um profissional da saúde.

Zhang [34] objetiva criar uma inteligência artificial para reconhecer em tempo real as expressões de alunos em sala

<sup>1</sup><https://www.sciencedirect.com>

<sup>2</sup><https://www.kaggle.com>

<sup>3</sup><https://www.python.org>

de aula. O objetivo, similar ao deste trabalho, seria para que o professor possa modificar sua aula ou a explicação do conteúdo para melhor entendimento da turma. Utilizando as imagens disponíveis no banco de dados *Fer-2013*, elas foram submetidas a um pré-tratamento com a biblioteca *OpenCV*<sup>4</sup>, padronizando-as em uma imagem cinza de 42x42 *pixels*. Uma CNN foi combinada com convolução separável em profundidade que reduz os cálculos na camada de convolução e *Global Average Pooling* para reduzir os parâmetros na camada totalmente conectada. O sistema consegue reconhecer quatro tipos de expressões, classificando-as em positivas, negativas, focadas ou surpresas, conseguindo chegar a 73% de precisão.

Identificar emoções humanas com base nas expressões faciais já é um trabalho complexo, mas identificá-las em imagens onde o rosto esteja parcialmente coberto ou com uma iluminação diferente é realmente um desafio. Kuruvayil e Palaniswamy [35] sugerem o uso de meta-aprendizagem, onde a máquina “aprende a aprender”. Utilizou-se a *Multi-task Cascaded Convolutional Networks* para detecção e extração da face das fotos com imagens do *AffectNet* e o conjunto de dados *CMU Multi-PIE*, que possuiu as imagens com as variações necessárias, excluindo aquelas que poderiam prejudicar o aprendizado, como pessoas de óculos, barba ou cabelo no rosto. *ERMOPI (Emotion Recognition using Meta-learning through Occlusion, Pose and Illumination)* foi o modelo proposto, conseguindo uma precisão de 90% para as imagens do *CMU Multi-PIE* e 68% para as do *AffectNet*.

Chiurco et al. [36] propõe o reconhecimento de emoções em tempo real. O foco foi identificar o estado dos trabalhadores durante uma tarefa colaborativa, a fim de melhorar a interação humano-robô em sistemas de montagem. Foram testadas cinco redes neurais, sendo elas a *Network*, Rede *Xception*, VGG e duas CNNs, com três conjuntos de dados, *Fer2013*, *CK+* e *LFW*. Os testes, tanto de laboratório quanto em uma fábrica real, mostraram a dificuldade dos algoritmos na identificação de emoções, mostrando como a iluminação, pose e até mesmo o ambiente de fundo impactam diretamente nos resultados. Nos testes controlados, o *DeepFace* teve a melhor performance, com 39,13% de precisão, contudo no teste do mundo real chegou apenas a 26,6%, mostrando o quanto difícil é a tarefa de reconhecimento de emoções.

Mellouk e Handouzi [37] trazem em seu trabalho um guia para reconhecimento de emoções via *deep learning*. É listado trabalhos, *datasets* com as respectivas emoções contidas nas imagens e uma lista comparativa com a acurácia. Os autores comentam como os estudos ainda precisam avançar em relação a emoções secundárias que não entram nas principais (medo, raiva, felicidade, tristeza, surpresa, nojo e neutro), além de pesquisas onde exista a combinação de mais de um meio de detecção, como áudio e imagem.

### C. Pontos de Destaque dos Artigos Selecionados

Dos artigos selecionados, Chiurco et al. [36] e Mellouk e Handouzi [37] trouxeram tabelas comparativas de trabalhos relacionados ao tema, indicando suas respectivas acurácias, arquiteturas e *datasets*. Utilizando estes dados e aqueles lidos

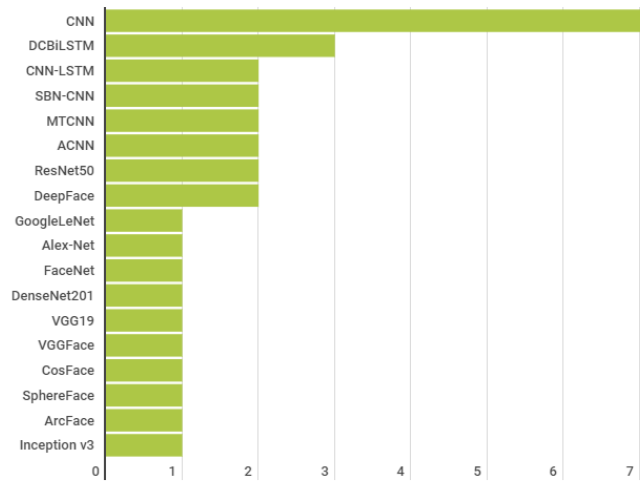


Figura 9: Quantidade de Uso das Redes Neurais.

na revisão sistemática, comparou-se os resultados de 32 artigos ao todo, notando a predominância de certos *datasets* e arquiteturas.

No que se refere as redes neurais, as CNNs são as que possuem mais ocorrências, como mostra o gráfico da Figure 9, sendo elas algum modelo pronto ou criado pelos próprios autores. Os modelos LSTM aparecem na tabela, contudo na forma de uma arquitetura híbrida, já que as redes neurais convolucionais são as que tem melhores resultados nos casos de visão computacional. Os *datasets* mais utilizados estão representados na Figure 10, sendo *CK+*, *LFW* e *Fer2013* os mais recorrentes, com 7 ocorrências para o primeiro e 4 para os outros dois. A partição “outros” são todos os *datasets* que apareceram uma única vez nos trabalhos, sendo colocados na mesma categoria para facilitar a visualização.

Com isso em vista, foi feita uma análise mais aprofundada dos trabalhos, notando-se que onze entre eles tratavam da classificação de emoções propriamente dita. Estes estão dispostos na Table I, que traz os autores do trabalho, arquitetura utilizada, *dataset*, acurácia, quantidade de emoções trabalhadas e quais são elas. A ordenação se deu da maior quantidade de emoções para a menor, seguida da maior acurácia para a menor. Assim, nota-se que destes, sete artigos tratam de seis emoções humanas mais o estado neutro - caracterizando o rosto sem nenhuma emoção - não apresentando a emoção de desprezo das sete categorias universais definidas por Paul Ekman 38. Entre esses que analisaram sete emoções, a menor acurácia obtida é de 65% e a maior de 96,40%, notando-se que quatro trabalhos ficaram acima de 90%. A importância destes resultados é observar os valores obtidos com suas respectivas arquiteturas de rede e *datasets*, sendo um comparativo para os resultados da presente pesquisa.

## IV. MATERIAIS E MÉTODOS

Nesta seção são descritos os materiais e métodos referentes a elaboração do presente trabalho.

<sup>4</sup><https://opencv.org>

Tabela I: Comparação de Trabalhos Relacionados.

Autores	Arquitetura	Dataset	Acurácia	Quantidade de emoções	Emoções
Li e Lima [31]	ResNet-50	Imagens pelos autores	96,40%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Jain, Shamsolmoali e Sehdev [39]	CNN	JAFFE	95,23%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Cheng e Zhou [40]	Alex-Net	CK+	94,40%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Jain, Shamsolmoali e Sehdev [39]	CNN	CK+	93,24%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Ivanova e Borzunov [32]	GoogleLeNet	FER2013	88,00%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Chiurco et al. [36]	DeepFace	Fer2013	72,00%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Agrawal e Mittal [41]	CNN	FER2013	65,00%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Liang et al. [42]	DCBiLSTM	Oulu-CASIA	91,07%	6	Felicidade, tristeza, medo, raiva, surpresa e nojo.
Kuruvayil e Palaniswamy [35]	MTCNN	CMU Multi-PIE	90,00%	5	Felicidade, raiva, surpresa, nojo e neutro.
Kuruvayil e Palaniswamy [35]	MTCNN	AffectNet	68,00%	5	Felicidade, raiva, surpresa, nojo e neutro.
Zhang [34]	CNN	Fer-2013	73,00%	4	Positiva, negativa, concentrado e surpreso.

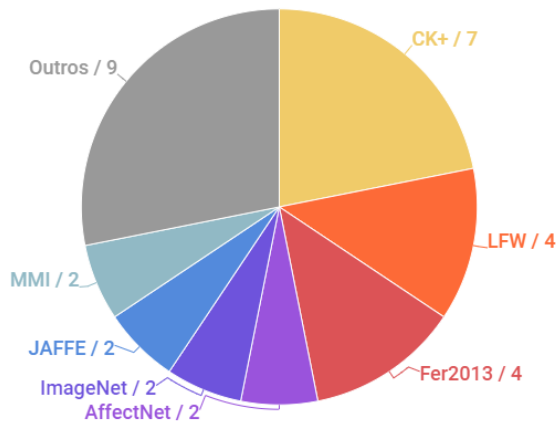


Figura 10: Quantidade de Uso dos Datasets.

### A. Método da pesquisa

O trabalho apresentado constitui uma pesquisa de natureza exploratória, a qual visa investigar, compreender e aplicar técnicas de *Machine Learning* no contexto da classificação de emoções. Como estudo de caso, escolheu-se a aplicação do trabalho no projeto SmartClass, que prevê o desenvolvimento de tecnologias para auxiliar os alunos por meio da inteligência artificial.

Estudo prévios realizados por Grando [22] e Chaves [43] identificaram que existem poucas iniciativas de sistemas que observem e monitorem as atividades dos alunos em sala de aula, com a finalidade de avaliar o interesse e a motivação com o aprendizado. Reconhece-se, contudo, que a motivação do aluno exerce um papel fundamental durante o processo de aprendizagem, não impactando apenas nele, mas também em todos que estão ao seu redor. Outros estados emocionais podem impactar nas relações em sala de aula, tais como: ansiedade e depressão. Um aluno com estado emocional de

ansiedade, que não consegue manter a concentração, pode acabar se desconcentrando e se prejudicando. Dentro deste contexto, este trabalho se concentra em compreender e identificar sentimentos e emoções que possam ser prejudiciais na vida acadêmica.

### B. Materiais

O presente trabalho visa desenvolver um modelo de *deep learning* que analisa e classifica emoções com base nas expressões faciais, em seguida aplicando o modelo em um cenário real: com imagens de alunos em três momentos dentro da sala da aula. O trabalho foi dividido da seguinte forma:

- Etapa 1: Aquisição de Imagens;
- Etapa 2: Pré-processamento de Imagens;
- Etapa 3: Aplicação do Algoritmo;
- Etapa 4: *Transfer Learning*;
- Etapa 5: Análise do Modelo.

Utilizou-se o Google Colaboratory como plataforma para o desenvolvimento do trabalho na linguagem de programação Python. Também, fez-se uso a API Keras<sup>5</sup> para o treinamento do modelo, que é uma biblioteca implementada com TensorFlow para auxiliar na tarefa de implementação da *deep learning*. O treinamento da rede foi feito utilizando as imagens disponíveis no *dataset* CK+ e as imagens dos alunos foram feitas em uma sala de aula real, na disciplina de Inteligência Computacional da Universidade de Caxias do Sul, pela professora responsável pela turma.

## V. DESENVOLVIMENTO

O desenvolvimento do trabalho foi organizado em seções para melhor abordagem dos tópicos.

<sup>5</sup><https://keras.io>



Figura 11: Exemplos de Diferentes Expressões do *Dataset CK+*.

### A. Etapa 1 - Aquisição de Imagens

No presente trabalho, houve dois momentos se trabalhando com imagens: no treinamento da rede neural, utilizando um *dataset* pronto, e com as imagens de alunos na classificação de emoções.

As imagens utilizadas no treinamento da rede neural são as disponibilizadas pelo *dataset Extended Cohn-Kanade*, mais conhecido pela sigla CK+. Ele disponibiliza mais de 500 imagens de diferentes pessoas, variando gênero e idade, em uma resolução de 640x490 ou 640x480 *pixels*. Todas estão arranjadas na mesma orientação - rosto verticalmente posicionado - na escala de cores cinza. Elas estão separadas por pastas, uma para cada emoção, sendo elas: raiva, nojo, desprezo, medo, felicidade, surpresa e tristeza [44], [45]. A Figure 11 traz alguns exemplos de imagens com as diferentes expressões disponíveis.

Separou-se 80% do *dataset* para o treino e 20% para teste com o método *hold-out*, que realizava a seleção das imagens escolhendo quantas instâncias de cada emoção selecionar.

Para o experimento, em uma sala de aula, foram utilizadas imagens de alunos durante a disciplina de Inteligência Computacional, na Universidade de Caxias do Sul (UCS). As fotos foram coletadas em três momentos durante a aula. Primeiro momento ocorreu durante a parte inicial da aula (explicação do conteúdo). O segundo momento ocorreu durante a resolução de exercícios. O terceiro momento de registro ocorreu no final da aula. As imagens foram coletadas no mês de novembro de 2022. A Figure 12 traz exemplo de algumas das capturas e registros realizados.

### B. Etapa 2 - Pré-processamento de Imagens

Para a utilização das imagens de treino do *dataset* foram feitos alguns ajustes a fim de melhor a eficiência da rede neural. O tamanho das imagens foi reduzido para 64x64 *pixels* e algumas foram invertidas horizontalmente, rotacionadas em até 20% e/ou ampliadas em até 10% para as expressões serem reconhecidas de diferentes ângulos. O exemplo da Figure 13 apresenta a mesma imagem com as diferentes alterações possíveis. Por fim, estas foram normalizadas para valores entre zero e um, a fim de aumentar a eficiência do modelo.



Figura 12: Imagens de Alunos.

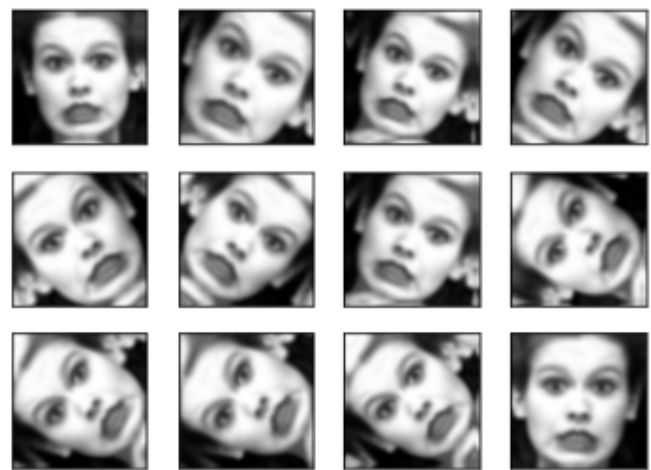


Figura 13: Exemplos de Pré-processamento das Imagens.



Figura 14: Imagens Pré-processadas dos Alunos.

Em decorrência do treinamento da rede neural ter sido feito com imagens em tons cinza, as imagens do experimento precisaram ser transformadas nessa escala de cor também. As fotos foram cortadas para deixar o rosto do aluno mais evidente e seu tamanho foi reduzido para 64x64 *pixels*, como foi feito com as imagens do *dataset*. A Figure 14 traz alguns exemplos já pré-processados, separadas por momento de aula.



### C. Etapa 3 - Aplicação do Algoritmo

Nesta etapa utilizou-se uma rede neural convolucional com a arquitetura descrita na Figure 15. As entradas são as imagens do *dataset* CK+, propagadas para as duas camadas combinadas de convolução e *Max Pooling*, até alcançar as duas camadas densamente conectadas. O modelo possui sete saídas, uma para cada emoção tratada: felicidade, tristeza, raiva, nojo, surpresa, desprezo e medo. Para a construção da arquitetura da rede utilizou-se como ponto de partida a implementação proposta por Tonon [46]. Em seu trabalho, Tonon concebeu uma rede simples para reconhecer obstáculos e objetos físicos encontrados nas calçadas. Os resultados obtidos foram muito bons, permitindo concluir-se que mesmo uma rede simples é capaz de reconhecer objetos em imagens.

A Figure 16 traz o sumário da arquitetura implementada. Nela percebe-se que a rede contém ao total dez camadas, totalizando 913.735 parâmetros para serem treinados. Os hiperparâmetros foram escolhidos depois de testes exaustivos, ao fim intercalando camadas convolucionais de função de ativação *ReLU* e *Max Pooling* de tamanho  $2 \times 2$ . A primeira camada da rede possui 32 filtros, enquanto que a segunda camada possui 64 filtros, a terceira possui 128 e a última apresenta 256 filtros. Em seguida, os dados passam para as camadas totalmente conectadas: uma camada densa de 128 neurônios e depois para outra camada densa com sete saídas. Cada uma destas saídas indica uma das emoções a ser classificada. Utilizou-se a função de ativação *Softmax* para deixar o resultado bem evidente. O modelo foi treinado por 40 épocas, com o parâmetro *batch size* igual a 32.

### D. Etapa 4 - Transfer Learning

Durante a realização dos testes com arquiteturas de redes neurais, investigou-se a possibilidade de treinamento por meio de *transfer learning* com o modelo VGG19. O modelo escolhido foi pré-treinado com as imagens do *dataset* ImageNet. Contudo, no decorrer dos testes com este modelo, percebeu-se que a rede havia sido pré-treinada com imagens coloridas (padrão RGB), o que reduziu drasticamente a acurácia dos resultados obtidos, uma vez que o *dataset* CK+ possui imagens em escala de cinza. Tendo em vista as limitações de tempo para preparação dos dados, optou-se por não utilizar as redes pré-treinadas por este motivo. Manteve-se então como alvo deste estudo a arquitetura de rede mais simples, previamente apresentada.

### E. Etapa 5 - Análise do Modelo

O modelo desenvolvido foi treinado a partir de 981 imagens de expressões faciais das sete emoções disponíveis do *dataset* CK+. Cada imagem de emoção facial estava previamente organizada em uma pasta específica. A quantidade de imagens de cada expressão seguiu a distribuição do *dataset* original:

- 135 de raiva;
- 54 de desprezo;
- 177 de nojo;
- 75 de medo;
- 207 de felicidade;

- 84 de tristeza;
- 249 de surpresa.

Para o treinamento foi utilizado o método *hold-out* estratificado, que consiste em separar os dados em duas partes: uma parte para treinar a rede neural e outra para testar sua performance [47]. Separou-se 80% das imagens para treino e 20% para os testes, não utilizando um conjunto previamente selecionado de imagens, mas sim deixando a cargo do algoritmo fazer a separação. Assim, realizou-se testes exaustivos a fim de melhorar a classificação do modelo, testando diversos tamanhos de filtros, adicionando e removendo camadas de convolução até alcançar-se o melhor modelo final.

Para avaliar o modelo e concluir se ele estava obtendo valores satisfatórios, utilizou-se as métricas de avaliação nos resultados. A maior acurácia obtida nos testes com o *dataset* foi de 90% e a menor de 65%. Tendo em vista que as imagens eram selecionadas de forma aleatória pelo algoritmo, e não com um mesmo conjunto de dados, o valor de acurácia se tornou variável, dependendo do conjunto selecionado para os testes que poderiam ser mais ou menos facilmente identificáveis pela rede neural. A matriz de confusão da Figure 17 permite visualizar os dados para cada uma das emoções específicas. No exemplo, a acurácia do modelo foi de 83,16%, sendo utilizadas 196 imagens ao todo nos testes, separadas da seguinte forma:

- 31 de raiva;
- 10 de desprezo;
- 36 de nojo;
- 20 de medo;
- 43 de felicidade;
- 7 de tristeza;
- 49 de surpresa.

As emoções possuíam números diferentes de instâncias, sendo o maior conjunto para a emoção de surpresa com 49 imagens e menor conjunto para a emoção de tristeza com 7 instâncias.

A partir da matriz de confusão, realizou-se os cálculos das métricas de avaliação para verificar a precisão, *recall* e *F1-Score* de cada emoção, como apresenta a Table II. A precisão indica a taxa de acertos, ou seja, das emoções listadas quantas o modelo classificou corretamente. Percebe-se que os melhores resultados foram para as expressões de surpresa e felicidade, com 98% e 95% respectivamente, e os piores para tristeza e desprezo com 36% e 50%. Observa-se que estas duas últimas emoções correspondem aos conjuntos de dados contendo o menor número de imagens de treino e de teste.

*Recall* indica a quantidade de classificações corretas, ou seja, quando é realmente a emoção esperada, quantas vezes o modelo a classificou corretamente. Os melhores valores foram para felicidade com 93% e raiva com 87% e os piores para desprezo, 40%, e tristeza, 57%. Por último, *F1-Score* faz a média harmônica destes valores, sendo assim, se um valor estiver muito baixo isso vai prejudicar consideravelmente o resultado desta métrica. Fica evidente que as expressões de desprezo e tristeza não obtiveram resultados muito satisfatórios, ambos com 44%, o que poderia se dar ao fato de serem as emoções com menos instâncias para treinamentos e testes da rede neural. Já felicidade e surpresa foram as melhores

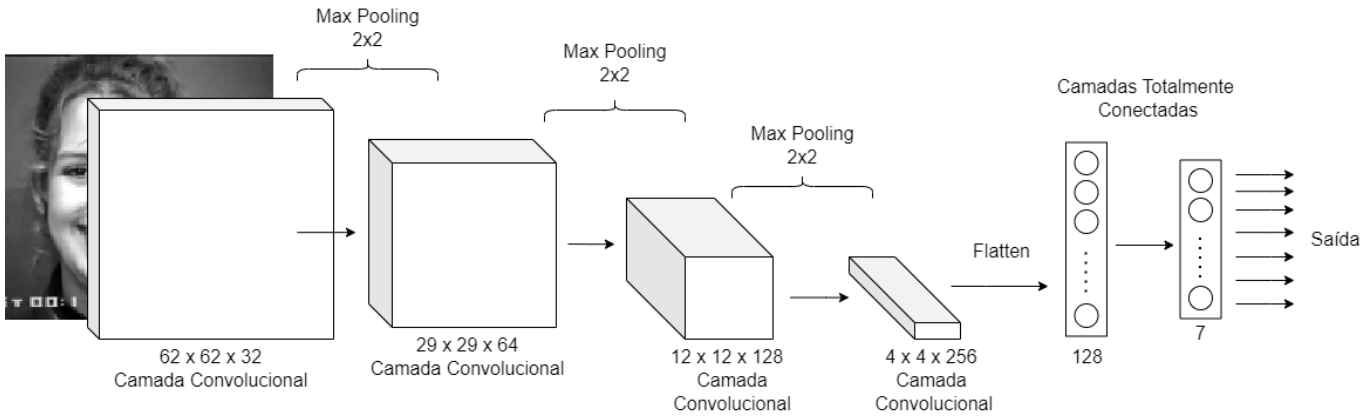


Figura 15: Arquitetura da Rede Neural Convolutacional.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 62, 62, 32)	896
max_pooling2d (MaxPooling2D)	(None, 31, 31, 32)	0
conv2d_1 (Conv2D)	(None, 29, 29, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 14, 14, 64)	0
conv2d_2 (Conv2D)	(None, 12, 12, 128)	73856
max_pooling2d_2 (MaxPooling2D)	(None, 6, 6, 128)	0
conv2d_3 (Conv2D)	(None, 4, 4, 256)	295168
flatten (Flatten)	(None, 4096)	0
dense (Dense)	(None, 128)	524416
dense_1 (Dense)	(None, 7)	903

-----

Total params: 913,735  
 Trainable params: 913,735  
 Non-trainable params: 0

Figura 16: Sumário da Rede Neural Convolutacional.

Tabela II: Resultados por Emoção.

Emoção	Precisão	Recall	F1-Score
Raiva	75%	87%	81%
Desprezo	50%	<b>40%</b>	<b>44%</b>
Nojo	81%	83%	82%
Medo	84%	80%	82%
Felicidade	95%	<b>93%</b>	<b>94%</b>
Tristeza	<b>36%</b>	57%	<b>44%</b>
Surpresa	<b>98%</b>	86%	<b>91%</b>

classificadas pelo modelo, com 94% e 91% respectivamente. Deve-se observar também que as outras emoções produziram valores para as métricas acima de 80%.

Sendo assim, percebe-se que o modelo têm maior facilidade de reconhecer as expressões de surpresa e felicidade, em decorrência da rede neural ter sido treinada com uma maior

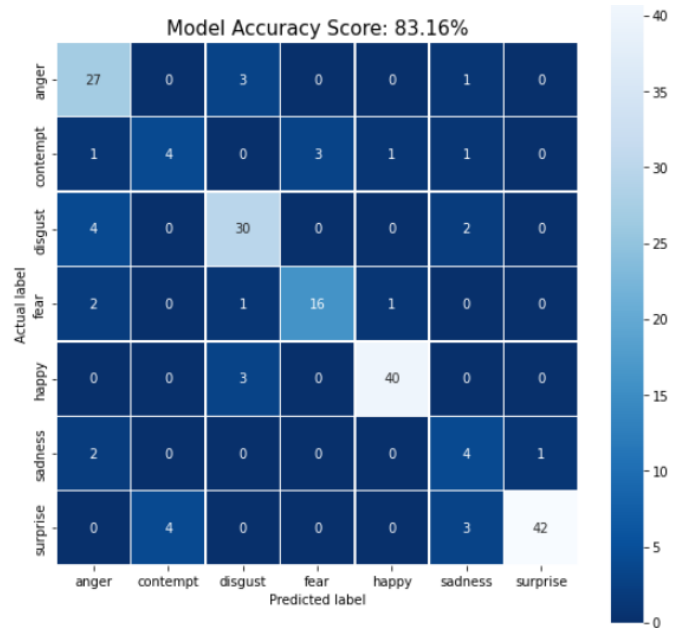


Figura 17: Matriz de Confusão do Modelo.

quantidade de instância para estas emoções. O mesmo se aplica nos casos da tristeza e desprezo, que foram as emoções com menos quantidade de instâncias no *dataset* CK+, o que pode ter prejudicado o desempenho do modelo. As outras emoções, raiva, medo e nojo, tiveram resultados parecidos entre si, ambos reconhecendo as emoções na maioria dos casos testados, o que é um fator positivo para o modelo.

## VI. EXPERIMENTO APLICADO

Para o experimento foi necessário utilizar imagens reais de alunos inseridos no ambiente de uma sala de aula. As fotografias foram feitas com 18 alunos em três momentos da classe: durante a explicação do conteúdo, na realização de exercícios e no final da aula, totalizando 61 imagens. Foi solicitado aos alunos manterem a expressão durante a captura da fotografia, a fim dos resultados serem mais naturais possíveis. Contudo, muitos rostos apresentam diferenças em relação as

imagens do *dataset* CK+, como rostos inclinados para baixo ou lado, mãos ocultando parte da face e muitos utilizando óculos, o que impacta na classificação das expressões.

As imagens foram separadas para cada momento de aula, já na escala cinza e cortadas para evidenciar os rostos. O modelo então classificou as fotos, apresentando um valor de probabilidade para cada uma das sete emoções. As figuras a seguir apresentam os resultados dessa classificação, uma para cada um dos momentos. Os valores foram arredondados para facilitar a visualização, sendo assim, em casos onde a porcentagem era muito próxima a zero o valor foi considerado como zero, ainda aparecendo a emoção correspondente ao lado já que havia uma probabilidade a ela relacionada. Nota-se que alguns alunos possuem mais de uma foto para cada momento de aula devido ao fato de seu rosto ter ficado evidente no plano de fundo de outra imagem, optando-se por utilizar o mesmo a fim de observar sua classificação. Os resultados foram comparados com a percepção da professora responsável da disciplina sobre como cada aluno aparentava estar no momento em que a foto foi tirada.

A Figure 18 mostra os resultados das imagens obtidas durante a explicação do conteúdo. A avaliação da professora indicou que os alunos estavam concentrados, fosse na explicação da matéria ou na tela do computador, evidenciando um estado de atenção em algum ponto focal. Já os resultados classificados pelo modelo indicaram muitas emoções de medo ou nojo para faces que estão mais contemplativas e não aparentando realmente um estado ansioso.

Na etapa de exercícios, com os resultados apresentados na Figure 19, a professora constatou que os alunos aparentavam estar concentrados, alguns conversando entre si, mas a maioria estava com o rosto sério, compenetrados na atividade e em silêncio. Observa-se muitas classificações de medo, surpresa e desgosto por parte do modelo, enquanto que a professora não observou tais emoções.

No final da aula os alunos estavam mais felizes, como observado pela professora e evidente em vários rostos da Figure 20. Contudo, na maioria dos casos o modelo não classificou como tal, principalmente em imagens com sorrisos mais sutis.

A partir destas avaliações, destacam-se alguns pontos importantes: a maioria das classificações foram de surpresa, medo e nojo, ao passo de que a professora não notou tal comportamento durante a aula. Os estudantes de fato aparentavam estarem calmos e/ou concentrados. Percebe-se, com estas evidências, a necessidade de que o modelo deveria ter sido treinado com faces neutras para momentos em que o aluno está mais relaxado ou concentrado, sem apresentar um estado de emoção intensa e/ou temerosa. De fato, em ambiente escolar, as emoções são mais moderadas, como a atenção, relaxamento ou distração, sentimentos que não constavam no treinamento do modelo, o que sem dúvida prejudicou a comparação com a opinião da professora.

Nota-se ainda que o modelo não percebe emoções sutis como os sorrisos discretos de felicidade, evidente nas classificações do final da aula. Vale destacar que, em ambiente escolar, os alunos raramente apresentam emoções intensas, normalmente aparentando expressões mais discretas e contidas

e não tão expansivas. Deve-se levar em consideração que a rede neural foi treinada com as imagens do *dataset* CK+ em que as faces mostram emoções exageradas, que não foi o caso encontrado dentro da sala de aula. Dessa forma, pontua-se como trabalhos futuros que modelo inclua estados mais discretos de emoção, apresentando faces mais serenas de expressões faciais, como tranquilidade e atenção por exemplo, e não apenas treinado com emoções intensas e acentuadas.

## VII. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

A sociedade atual está sempre em busca por tecnologias que facilitem o dia-a-dia e bem estar geral. Na área da saúde, isso é ainda mais evidente seja com aparelhos médicos, exames e cirurgias sendo feitos de formas mais rápidas e menos invasivas [48]. Com a quantidade crescente de casos de estresse, ansiedade e depressão, o tema de saúde mental tem estado cada vez mais relevante, não apenas na área da saúde como também na de tecnologia. Essas doenças se tornam cada vez mais comuns, a depressão afetando cerca de 3,8% da população mundial, 5% só entre adultos [5].

Na escola, isso não é diferente. O estresse, a preocupação, desmotivação, dificuldade de aprendizado, tudo isso e mais gera ansiedade entre os alunos, podendo evoluir para casos mais sérios de saúde mental. O aprendizado está diretamente relacionado ao emocional dos seres humanos; Quanto mais emocionalmente confiantes e estáveis, melhor é a capacidade de raciocinar e entender, e o contrário também se aplica. Para alunos em que o ambiente escolar se mostra um lugar hostil, a capacidade de aprendizado é afetada, aumentando sentimentos negativos que podem gerar consequências psicológicas [7].

Detectar emoções negativas em sala de aula pode ser visto como uma forma de auxiliar aos alunos e educadores no ambiente escolar. Para o professor, tal ferramenta serve de suporte para saber como os alunos se sentem em relação ao que está sendo ensinado: emoções negativas podem indicar que a matéria pode estar complicada ou confusa, sugerindo uma necessidade de mudança no método de ensino. Para o aluno é uma forma de se comunicar indiretamente sobre o que está sentindo no momento. O objetivo é auxiliar ambos os lados, informando os professores dos estados emocionais de sua turma.

Visando tal objetivo, o trabalho partiu de uma revisão sistemática da literatura, de cunho exploratório, a fim de entender o tema proposto. Logo após as pesquisas, passou-se para a implementação de uma ferramenta de *deep learning* para analisar imagens de expressões faciais e catalogá-las de acordo com as sete emoções humanas básicas. Entre as arquiteturas de *deep learning* existentes, a rede neural convolucional possui maiores taxas de sucesso relacionadas a problemas de classificação de imagens.

Foi selecionado o *dataset* CK+ para o treinamento da rede neural, que contém imagens com expressões faciais para as emoções de raiva, nojo, desprezo, medo, felicidade, surpresa e tristeza. A Table III compara o modelo desenvolvido pelo autor com dois trabalhos similares da revisão sistemática. Tais trabalhos também fizeram uso do *dataset* CK+ e analisaram sete emoções, com a diferença destes tratarem da expressão



Figura 18: Classificações dos Alunos Durante a Explicação do Conteúdo.

neutra e o autor da expressão de desprezo no lugar. Nota-se que a melhor acurácia do modelo proposto foi de 90%, ficando em terceiro lugar comparado com os outros dois trabalhos - o melhor deles chegando a 94,40%.

Após a etapa de treinamento e testes, o modelo foi aplicado com casos do mundo real: com fotos tiradas em três períodos dentro de uma sala de aula: durante a explicação do conteúdo, na realização de exercícios e no final da aula. Tais imagens foram classificadas pela rede e suas respostas comparadas com as avaliações feitas pelo professor.

Percebeu-se nos testes, utilizando as métricas de avaliação, que a rede neural conseguiu bons resultados, indicando um bom desempenho por parte do modelo. Para o contexto da sala de aula, entende-se que um trabalho futuro importante a

ser realizado é a inclusão de emoções pertinentes ao contexto da sala de aula. Neste ambiente, os alunos se mostram concentrados e atentos. O modelo baseado no *CK+* trata destas emoções como medo, desprezo, surpresa ou nojo. Notou-se então que a rede neural precisa ser aprimorada, incluindo emoções relacionadas ao ambiente de ensino-aprendizagem, que reflitam o estado neutro, a tranquilidade e atenção, já que é um ambiente controlado onde as emoções são mais contidas e não apresentadas de forma intensa.

#### REFERENCES

- [1] J. Lorenzetti, L. de Lima Trindade, D. E. P. de Pires, and F. R. S. Ramos, "Tecnologia, inovação tecnológica e saúde: uma reflexão necessária," *Texto & Contexto - Enfermagem*, vol. 21, pp. 432–439, June 2012.



Figura 19: Classificações dos Alunos Durante os Exercícios.

Tabela III: Comparação com o Modelo do Autor.

Autores	Arquitetura	Dataset	Acurácia	Quantidade de emoções	Emoções
Cheng e Zhou [40]	Alex-Net	CK+	94,40%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
Jain, Shamsolmoali e Sehdev [39]	CNN	CK+	93,24%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e neutro.
O Autor.	CNN	CK+	90,00%	7	Felicidade, tristeza, medo, raiva, surpresa, nojo, e <b>desprezo</b> .

- [2] P. E. Group, "What are emotions?," 2022. <https://www.paulekman.com/universal-emotions/>, accessed 2022-05-16.
- [3] L. A. N. ROBERT L. LEAHY, DENNIS TIRCH, *Regulação Emocional em Psicoterapia: Um Guia para o Terapeuta Cognitivo-Comportamental*. Editora Artmed, 02 2013.
- [4] CDC, "About mental health," 2021. <https://www.cdc.gov/mentalhealth/learn/index.htm>, accessed 2022-06-12.
- [5] WHO, "Depression," 2021. <https://www.who.int/news-room/fact-sheets/detail/depression>, accessed 2022-06-12.
- [6] WHO, "Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide," 2022. <https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide>, accessed 2022-06-12.
- [7] V. d. Fonseca, "Importância das emoções na aprendizagem: uma abordagem neuropsicopedagógica," *Revista Psicopedagogia*, vol. 33, pp. 365 – 384, 00 2016.
- [8] D. de Camargo, *As emoções & a escola*. 06 2022.
- [9] J. C. Souza, A. A. Hickmann, A. Asinelli-Luz, and G. M. Hickmann, "A influência das emoções no aprendizado de escolares," *Revista Brasileira de Estudos Pedagógicos*, vol. 101, Aug. 2020.
- [10] H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251–260, 2020. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.
- [11] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, and S. Yu, "A survey on deep learning for textual emotion analysis in social networks," *Digital Communications and Networks*, 2021.
- [12] R. da Silva Nascimento, P. Parreira, G. N. D. Santos, and G. P. Guedes, "Identificando sinais de comportamento depressivo em redes sociais," in *Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, Sociedade Brasileira de Computação - SBC, July 2018.
- [13] J. Domínguez-Jiménez, K. Campo-Landines, J. Martínez-Santos, E. Delahoz, and S. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical Signal Processing and Control*, vol. 55, p. 101646, 2020.



Figura 20: Classificações dos Alunos no Final da Aula.

- [14] D. S. Academy, “Deep learning book,” 2022. <https://www.deeplearningbook.com.br>, accessed 2022-04-28.
- [15] IBM, “Deep learning,” 2020. <https://www.ibm.com/cloud/learn/deep-learning>, accessed 2022-06-16.
- [16] DeepAI, “Feed forward neural network,” 2022. <https://deepai.org/machine-learning-glossary-and-terms/feed-forward-neural-network>, accessed 2022-05-18.
- [17] P. Antoniadis, “Hidden layers in a neural network,” 2022. <https://www.baeldung.com/cs/hidden-layers-neural-network>, accessed 2022-06-17.
- [18] F. V. Veen, “The neural network zoo,” 2016. <https://www.asimovinstitute.org/neural-network-zoo/>, accessed 2022-05-18.
- [19] D. Ceccon, “Os tipos de redes neurais,” 2020. <https://iaexpert.academy/2020/06/08/os-tipos-de-redes-neurais/>, accessed 2022-05-19.
- [20] G. Alves, “Entendendo redes convolucionais (cnns),” 2018. <https://medium.com/neuronio-br/entendendo-redes-convolucionais-cnns-d10359f21184>, accessed 2022-06-02.
- [21] A. Deshpande, “A beginner’s guide to understanding convolutional neural networks,” 2022. <https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/>, accessed 2022-05-17.
- [22] A. GRANDO, “Análise facial de emoções utilizando redes neurais no contexto de uma sala de aula inteligente,” 2020.
- [23] TechTerms, “Rgb,” 2019. <https://techterms.com/definition/rgb>, accessed 2022-06-02.
- [24] V. Rodrigues, “Conhecendo a visão do computador: Redes neurais convolucionais,” 2019. <https://vitorborbarodrigues.medium.com/conhecendo-a-visão-do-computador-redes-neurais-convolucionais-e1c2b14bf426>, accessed 2022-06-03.
- [25] R. M. Marcius Lima, Eduardo Rubik, “Introdução ao reconhecimento de imagens,” 2020. <https://lamfo-unb.github.io/2020/12/05/Captcha-Break/>, accessed 2022-06-02.
- [26] B. Prijono, “Student notes: Convolutional neural networks (cnn) introduction,” 2018. <https://indoml.com/2018/03/07/student-notes-convolutional-neural-networks-cnn-introduction/>, accessed 2022-06-04.
- [27] A. P. J. Dwiyanoro, “The evolution of computer vision techniques on face detection, part 2,” 2018. <https://medium.com/nodeflux/the-evolution-of-computer-vision-techniques-on-face-detection-part-2-4af3b22df7c2>, accessed 2022-06-05.
- [28] K. Transfer, “Explain pooling layers: Max pooling, average pooling, global average pooling, and global max pooling,” 2021. <https://androidkt.com/explain-pooling-layers-max-pooling-average-pooling-global-average-pooling-and-global-max-pooling/>, accessed 2022-06-05.
- [29] V. Rodrigues, “Métricas de avaliação: acurácia, precisão, recall... quais as diferenças?,” 2019. <https://vitorborbarodrigues.medium.com/métricas-de-avaliação-acurácia-precisão-recall-quais-as-diferenças-c8f05e0a513c#:~:text=Recall%2FRevocação%2FSensibilidade%3A%20entre,harmática%20entre%20precisão%20e%20recall,> accessed 2022-05-23.
- [30] J. Mohajon, “Confusion matrix for your multi-class machine learning model,” 2020. <https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826>, accessed 2022-12-04.
- [31] B. Li and D. Lima, “Facial expression recognition via resnet-50,” *International Journal of Cognitive Computing in Engineering*, vol. 2,

- pp. 57–64, 2021.
- [32] E. Ivanova and G. Borzunov, “Optimization of machine learning algorithm of emotion recognition in terms of human facial expressions,” *Procedia Computer Science*, vol. 169, pp. 244–248, 2020. Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society), held August 15-19, 2019 in Seattle, Washington, USA.
- [33] G. Lebedev, E. Zhovnerchuk, I. Zhovnerchuk, and A. Moskovenko, “Remote recognition of human emotions using deep machine learning of artificial neural networks,” *Procedia Computer Science*, vol. 176, pp. 1517–1522, 2020. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.
- [34] T. Zhang, “Face expression recognition based on deep learning,” *Journal of Physics: Conference Series*, vol. 1486, p. 042048, apr 2020.
- [35] S. Kuruvayil and S. Palaniswamy, “Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning,” *Journal of King Saud University - Computer and Information Sciences*, 2021.
- [36] A. Chiurco, J. Frangella, F. Longo, L. Nicoletti, A. Padovano, V. Solina, G. Mirabelli, and C. Citraro, “Real-time detection of worker’s emotions for advanced human-robot interaction during collaborative tasks in smart factories,” *Procedia Computer Science*, vol. 200, pp. 1875–1884, 2022. 3rd International Conference on Industry 4.0 and Smart Manufacturing.
- [37] W. Mellouk and W. Handouzi, “Facial emotion recognition using deep learning: review and insights,” *Procedia Computer Science*, vol. 175, pp. 689–694, 2020. The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC), The 15th International Conference on Future Networks and Communications (FNC), The 10th International Conference on Sustainable Energy Information Technology.
- [38] P. E. Group, “Are facial expressions universal?,” 2022. <https://www.paulekman.com/resources/universal-facial-expressions/>, accessed 2022-05-16.
- [39] D. K. Jain, P. Shamsolmoali, and P. Sehdev, “Extended deep neural network for facial emotion recognition,” *Pattern Recognition Letters*, vol. 120, pp. 69–74, 2019.
- [40] S. Cheng and G. Zhou, “Facial expression recognition method based on improved vgg convolutional neural network,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 07, p. 2056003, 2020.
- [41] A. Agrawal and N. Mittal, “Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy,” *The Visual Computer*, vol. 36, pp. 405–412, Jan. 2019.
- [42] D. Liang, H. Liang, Z. Yu, and Y. Zhang, “Deep convolutional bilstm fusion network for facial expression recognition,” *The Visual Computer*, vol. 36, no. 3, pp. 499–508, 2020.
- [43] R. R. CHAVES, “Redes neurais deep learning aplicadas ao reconhecimento facial,” 2018.
- [44] J. R. H. Lee, L. Wang, and A. Wong, “Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition,” 2020.
- [45] 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,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, IEEE, June 2010.
- [46] A. E. G. Tonon, “Modelo de visão computacional para auxílio a deficientes visuais na locomoção urbana,” Master’s thesis, Universidade de Caxias do Sul (UCS), 12 2021.
- [47] E. Allibhai, “Hold-out vs. cross-validation in machine learning,” 2018. <https://medium.com/@ejjaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f>, accessed 2022-11-04.
- [48] L. Lopes, “Medicina robótica tornou cirurgias mais seguras e menos invasivas,” 2019. <https://jornal.usp.br/atualidades/medicina-robotica-tornou-cirurgias-mais-seguras-e-menos-invasivas/>, accessed 2022-06-16.