Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 07/Oct/2022; 1st round notif.: 14/Oct/2022; New version: 09/Nov/2022; 2nd round notif.: 26/Jan/2023;
Camera ready: 27/Jan/2023; Edition review: 13/Apr/2023; Available online: 28/Apr/2023; Published: 28/Apr/2023;

Detecção por face de emoções de aprendizagem: abordagem baseada em redes neurais profundas e fluxo de emoções

Title: Detection by face of learning emotions: an approach based on deep neural networks and on the emotions flow

Pablo Werlang
PPG em Computação Aplicada (PPGCA)
Universidade do Vale do Rio dos Sinos (UNISINOS)
ORCID: 0000-0003-0564-1336
pablowerlang@edu.unisinos.br

Patrícia A. Jaques
PPG em Informática (PPGInf)
Universidade Federal do Paraná (UFPR)
PPG em Computação (PPGC)
Universidade Federal de Pelotas (UFPEL)
ORCID: 0000-0002-2933-1052
patricia@inf.ufpr.br, patricia.jaques@inf.ufpel.edu.br

Resumo

O reconhecimento automático de emoções através da face possui o potencial de tornar a interação com um computador uma experiência mais natural. Em especial nos ambientes inteligentes de aprendizagem, a detecção das emoções beneficia diretamente os estudantes ao usar as suas informações afetivas para perceber suas dificuldades, adaptar a intervenção pedagógica e engajá-lo. Este artigo apresenta um modelo de aprendizado de máquina capaz de reconhecer, por vídeos da face, as emoções engajamento, confusão, frustração e tédio, experimentadas pelos estudantes em seções de interação com ambientes de aprendizagem. O modelo proposto se utiliza de redes neurais profundas para realizar a classificação em uma destas emoções, extraindo características estatísticas, temporais e espaciais dos vídeos fornecidos para treinamento, incluindo movimento dos olhos e movimentos musculares face. O trabalho possui como principal diferencial a consideração do fluxo das emoções como entrada, ou seja, a sequência em que as emoções são manifestas. Diversas configurações de modelos de aprendizado profundo de máquina foram testadas, e suas eficiências comparadas ao estado da arte. Os resultados trazem evidências que considerar a sequência de emoções de aprendizagem dos estudantes como entrada nos modelos melhora a efetividade desses algoritmos. Utilizando o treinamento na base de dados DAiSEE, o ganho de desempenho na métrica F1 foi de $26,27\%$ *(de* $0,5122$ *para* $0,6468$*) quando incluído o histórico de emoções no modelo.*

Palavras-chave: *reconhecimento de emoções, redes neurais profundas, emoções no aprendizado*

Abstract

Automatic face recognition of emotions has the potential of turning the human-computer interaction an increasingly natural experience. Especially in intelligent learning environments, emotion detection benefits the students by directly using their affective information to perceive their difficulties, adapt the pedagogic intervention and engage them. The present article presents a model capable of recognizing by face the emotions commonly experienced by students in interaction sections with learning environments: engagement, confusion, frustration, and boredom. The proposed model uses deep neural networks to classify one of these emotions, extracting statistical, temporal, and spatial features from the videos provided for training, including eye and facial movements. This work's main contribution is to take into account the flow of emotions (the sequence of emotions in the order that they are experienced by a student) as a mean for increasing emotion detection accuracy. We tested several model configurations and their efficiency compared to the state of art models. Results show that taking into account the learning emotions sequence as models' input improves those algorithms' effectiveness. Training the model on the DAiSEE dataset, we achieved 26.27% F1 improvement (from 0.5122 to 0.6468) when including the emotions' history in the model.

Keywords: *emotion recognition, deep learning, learning emotions*

Cite as: Werlang, P., Jaques, P. A. (2023). Detecção por face de emoções de aprendizagem: abordagem baseada em redes neurais profundas e fluxo de emoções. Revista Brasileira de Informática na Educação – RBIE, v. 31, 174-204. DOI: 10.5753/rbie.2023.2936.

1 Introdução

No passado, as emoções tiveram seu papel relegado ao contraponto do raciocínio humano, porém, atualmente, há o entendimento que elas são parte do processo cognitivo e importante ferramenta no processo de avaliação de experiências e tomadas de decisão (Lazarus, 1982). Tendo em vista sua importância, é esperado que sistemas que almejem uma interação mais eficiente com o usuário busquem identificar e responder às emoções do usuário. E essa ciência das emoções do usuário é especialmente importante aos sistemas dedicados ao ensino e aprendizagem. Eles necessitam ter ciência das emoções que seus usuários experienciam, bem como saber como agir para instigar as emoções mais úteis no contexto da aprendizagem.

Dado este contexto, a possibilidade da detecção das emoções de maneira automática nos ambientes inteligentes de aprendizagem é algo muito útil, pois possibilita um processamento e resposta imediata por parte do sistema ao estímulo emocional detectado. Um grande número de estudos relatam o reconhecimento automático de emoções através de diversas modalidades (Reis, Jaques, & Isotani, 2018) como, por exemplo, expressões faciais e movimento dos olhos (Yang, Wang, Peng, & Qiao, 2018), condutividade da pele (Subramanian et al., 2016), batimentos cardíacos (Nardelli, Valenza, Greco, Lanata, & Scilingo, 2015), ondas cerebrais (Ackermann, Kohlschein, Bitsch, Wehrle, & Jeschke, 2016), áudio da voz (Liu, Tang, Lv, & Wang, 2018), *logs* de uso do sistema (Morais, Kautzmann, Bittencourt, & Jaques, 2019), dentre outros.

Dentre todas as modalidades de detecção automática de emoções, a que mais se assemelha ao modo como humanos realizam esta tarefa, é a detecção através de informações visuais. Seres humanos utilizam principalmente o sentido da visão para detectar perigos e intenções em outros seres. Consequentemente, para reconhecimento das emoções, também utilizam naturalmente meios detectáveis visualmente, como expressões faciais, movimentos do corpo, e sinais gestuais (Kreifelts, Wildgruber, & Ethofer, 2013). Sendo assim, a detecção das emoções através de aspectos visuais é naturalmente um caminho promissor a seguir quando se criam algoritmos na tentativa de realizar tal feito.

A detecção automática de emoções através da face tornou-se possível através dos avanços nas técnicas de reconhecimento facial (Sariyanidi, Gunes, & Cavallaro, 2014). Para realizar o reconhecimento facial em imagens, algoritmos, chamados de classificadores, detectam pontos importantes da face, chamados *landmarks*. A partir destes pontos, o contorno da face é separado do resto da imagem e obtém-se a face alinhada (*aligned face*, do inglês). Após o reconhecimento facial, a extração de características e classificação das emoções é realizada.

Para a implementação do classificador, uma das técnicas frequentemente utilizadas é rede neural. As redes neurais imitam a maneira que os neurônios em um cérebro se comunicam e aprendem novas informações, e elas se utilizam de imensas quantidades de dados fornecidos como exemplo para aprenderem padrões. As redes neurais cresceram em popularidade nos últimos tempos e hoje conseguem aprender padrões complexos graças ao (Marcus, 2018): (1) aumento da capacidade computacional, que permitiu que redes mais complexas e com mais camadas sejam criadas, e (2) aumento da quantidade de bases de dados para treinamento, pois as redes neurais necessitam de muita informação para aprender os padrões. Estes dois fatores, somado a sucessos na resolução de problemas, como do gradiente de desaparecimento, culminaram no crescimento do uso das redes neurais profundas: redes neurais cuja abordagem para construção do modelo está

na utilização de muitas camadas, ou profundidade da rede, ao invés da largura da mesma. Este tipo de rede, capaz de identificar padrões abstratos mais complexos, tornou possível a resolução de uma série de problemas para os quais não haviam sido encontradas soluções anteriormente.

Grande parte das pesquisas em detecção automática de emoções através da expressão facial realizam a classificação em uma das seis emoções básicas (Ekman, 1999): alegria, tristeza, raiva, medo, surpresa e nojo. Conforme a teoria das emoções básicas, indivíduos em diferentes contextos culturais e geográficos expressam estas emoções da mesma maneira. Porém, de acordo com D'Mello and Calvo (2013), se tratando de situações de aprendizagem, e em especialmente em ambientes de aprendizagem, essas emoções acontecem raramente. Emoções como engajamento, confusão, frustração e tédio são muito mais frequentes, e quando comparadas às emoções básicas, se manifestam neste tipo de ambiente em uma razão de 1:5. Essas emoções são geralmente rotuladas de emoções de aprendizagem, emoções cognitivas ou emoções acadêmicas (Pekrun, 2011; Ocumpaugh, 2015; D'Mello & Calvo, 2013).

Entre as emoções de aprendizagem, o engajamento é a emoção mais desejável, pois indica que o aluno está motivado e assimilando informações. Quando uma aluna engajada se depara com discrepâncias entre as informações apresentadas e seu conhecimento, ela entra no estado de confusão, que, de acordo com D'Mello, Lehman, Pekrun, and Graesser (2014), lhe leva a refletir sobre as informações, fazendo-a entrar em um estado de desequilíbrio cognitivo. Caso ela resolva com sucesso o estado de confusão, ela retorna ao estado de engajamento, e caso a confusão perdure por muito tempo, ela pode entrar no estado de frustração, e futuramente em tédio caso a frustração se desenvolva por muito tempo. Portanto, é importante para o processo de aprendizagem do aluno que ele resolva o estado de confusão, entrando no estado de engajamento. D'Mello et al. (2014) demonstram que a confusão é um estado afetivo que pode auxiliar no processo de aprendizado, e Fredrickson (1998) demonstra como as emoções positivas influenciam positivamente o repertório de ações e pensamento das pessoas.

O presente trabalho foi desenvolvido de modo a realizar a detecção e a classificações dessas emoções relacionadas à aprendizagem (engajamento, confusão, frustração e tédio) através do uso de vídeos das faces de alunos em situação de aprendizagem, seja lendo conteúdos no computador ou interagindo com um ambiente de aprendizagem. Embora já existam trabalhos consolidados na detecção de emoções por face, esses trabalhos são voltados ao reconhecimento das emoções básicas, incomuns em ambientes de aprendizagem. Conforme demonstrado por D'Mello and Calvo (2013), outras emoções, como engajamento, confusão, frustração e tédio, são presenciadas com muito mais frequência no contexto de aprendizagem. No entanto, a tarefa da detecção destes estados afetivos é muito mais complexa que a detecção de emoções básicas. Isto se dá pelo fato que as emoções presentes no aprendizado são expressas de maneira muito mais sutil. Tendo isso em vista, há a necessidade de uma abordagem diferenciada para realização da mesma.

Os autores abordam neste trabalho esta questão através da exploração da relação temporal entre as emoções. Estudos mostraram que existe uma probabilidade de transição entre as emoções de aprendizagem (D'Mello & Graesser, 2012; D'Mello et al., 2014), ou seja, estando o estudante numa emoção x de aprendizagem, existe maior ou menor probabilidade dele transitar para uma emoção y do que para as outras emoções. Por exemplo, quando frustrada ou entendiada existe menor probabilidade de uma estudante voltar a se engajar. Esses padrões de transições poderiam ser aprendidos por uma rede neural profunda se ela possuir como entrada a sequência de emoções vivenciadas pelo aprendiz.

Para o trabalho em questão, a captura de informações se dá por dispositivos de gravação de vídeo (*webcam*) presentes nos computadores em que foram utilizados os ambientes de aprendizagem pelos estudantes para assistir videoaulas ou resolver problemas. Os vídeos gerados pela gravação do rosto dos estudantes são processados a fim de que características importantes para a classificação das emoções sejam extraídas. Por fim, a classificação das emoções é realizada através da implementação de diversos modelos de redes neurais profundas, onde existe um classificador para cada uma das quatro emoções: engajamento, confusão, tédio e frustração. Redes neurais convulacionais e LSTM foram treinadas com vídeos de face e rótulos de emoções de três bases diferentes com estudantes em situações de aprendizagem. Os modelos foram testados com validação cruzada. A eficiência dos modelos são comparados entre si e aos classificadores de outros modelos existentes que possuem a proposta de detecção de emoções presentes em ambientes de aprendizagem, como, por exemplo, o engajamento.

O trabalho proposto possui como **objetivo principal** obter acurácia do reconhecimento de emoções acadêmicas por algoritmos de aprendizagem profunda melhor que o estado da arte para as emoções manifestadas durante situações de aprendizagem, sendo elas engajamento, confusão, frustração e tédio. Além de considerar informações visuais comumente empregadas por esses modelos, tais como pose da cabeça, esse trabalho também usa no treinamento das redes neurais classificadoras a temporalidade das emoções dos estudantes para melhorar a acurácia da detecção das emoções de aprendizagem. A sequência de emoções, na ordem em que elas são manifestas pelo estudante, são fornecidas como entrada para redes profundas temporais, para que os modelos aprendam a inferir padrões de transição entre emoções e assim possam melhorar a acurácia da detecção. Dessa forma, esse trabalho possui dois objetivos específicos:

- Obter precisão do reconhecimento de emoções acadêmicas por algoritmos de aprendizagem profunda melhor que o estado da arte para o reconhecimento de emoções por face para estas emoções.
- Verificar a influência da temporalidade das emoções (sequencia que as emoções foram sentidas pelos estudantes) na precisão do reconhecimento de emoções acadêmicas por algoritmos de aprendizagem profunda.

Esse artigo está organizado da seguinte maneira. A Seção 2 apresenta os trabalhos relacionados, comparando suas características principais a fim de destacar a originalidade do trabalho proposto. A Seção 3 detalha o modelo de detecção de emoções por face desenvolvido, explicando desde seus conceitos gerais, as decisões de projeto, e detalhes das arquiteturas utilizadas. Por fim, a Seção 4 discute os resultados encontrados no trabalho conforme os objetivos iniciais e a Seção 5 apresenta as principais conclusões.

2 Trabalhos Relacionados

Para a seleção dos trabalhos relacionados, foi primeiramente empregada a técnica de revisão *Snowball*, inspirada na metodologia proposta por Goodman (1961). Sua aplicação na seleção de artigos científicos é dada através da leitura de alguns trabalhos inicialmente selecionados e, a partir das referências presentes nestes, são buscados novos artigos que sejam do interesse da pesquisa sendo

realizada. Este processo pode possuir quantos níveis forem necessários, formando assim uma rede de buscas que visam complementar a seleção dos artigos e incluir referências importantes (Heckathorn, 1997).

A partir da leitura dos trabalhos mais relevantes da área de detecção de emoções presentes no aprendizado, obtidos a partir de um levantamento inicial informal realizado pelos autores, foi possível elaborar uma *string* de busca[1] com os termos mais relevantes para o trabalho proposto. Foram realizadas pesquisas nas bibliotecas digitais Google Scholar, ACM Digital Library e IEEE Xplore Digital Library[2], reconhecidas por serem bases de pesquisa acadêmica de publicações científicas.

A seleção foi realizada buscando atender além dos critérios do assunto da pesquisa, filtrados na *string* de busca, um ou mais dos seguintes critérios: número de citações, classificação do periódico ou conferência da publicação, recenticidade do trabalho, uso de bases de dados públicas e/ou acessíveis através de autorização, bem como trabalhos que relacionam seus resultados usando as mesmas bases de dados. Foram selecionados 51 artigos pelos seus títulos, e após a leitura de seus resumos, 15 foram selecionados para leitura integral. Seis trabalhos deste grupo foram considerados mais relevantes à presente proposta, descritos na Seção 2.1.

2.1 Detecção automática de emoções acadêmicas por face

D'Mello and Calvo (2013) demonstram que, em situações de aprendizagem, emoções não básicas como engajamento, confusão, frustração e tédio são encontradas com muito maior frequência que as emoções básicas, em uma razão de 5:1, e parecem ter um impacto muito maior na aprendizagem. Por este motivo, trabalhos que detectam emoções em ambientes de aprendizagem costumam se utilizar destas outras emoções, chamadas de acadêmicas ou de aprendizagem, no auxílio ao processo de aprendizagem do aluno.

Os trabalhos de Gupta e colegas (2016) divulgam a criação da base de dados DAiSEE (explicada adiante) e desenvolvem um classificador das emoções engajamento, confusão, frustração e tédio, servindo como ponto de origem para outros trabalhos que utilizam a mesma base de dados. Como um dos poucos trabalhos que realizam a classificação das quatro emoções acadêmicas, os autores utilizam a ferramenta FaceAlign, que invoca funções da OpenCV (Bradski, 2000) para extrair as faces alinhadas das imagens. A partir deste pré-processamento, são extraídas as características LBP-TOP e HOG das imagens, bem como é realizado o ajuste fino da rede VGG (Simonyan & Zisserman, 2014) para a extração de características espaciais da imagem. Eles experimentam diversos modelos, desde o ajuste fino de redes Inception (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), treinamento de uma rede C3D (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015) e o treinamento de uma rede neural convolucional recorrente (CNN+LSTM). Seu melhor resultado foi proveniente desta última configuração (embora não mencione a topologia), onde obteve acurácia 57,9% para o engajamento, 53,7% para o tédio, 72,3% para a confusão e 73,5 para a frustração.

Dentre as emoções presentes no contexto do aprendizado, o engajamento é de longe a emoção que mais possui publicações a respeito. O trabalho de Dewan, Lin, Wen, Murshed, and Uddin

[1] String de busca usada: (engagement OR confusion OR frustration OR boredom) AND (detection OR recognition OR prediction OR classification) AND deep learning

[2] https://scholar.google.com, https://dl.acm.org e https://ieeexplore.ieee.org, respectivamente.

(2018), por exemplo, realiza a detecção do engajamento em alunos utilizando classificadores DBN (*Deep Belief Network*, do inglês) (Hinton, Osindero, & Teh, 2006). No trabalho, os autores utilizam também a base de dados DAiSEE, porém utilizam as informações das outras emoções como sinais para o engajamento. Por exemplo, eles utilizam os rótulos de frustrado e entediado como sendo não engajado. As redes DBN são utilizadas de maneira não supervisionada para a extração de características das imagens, e, após o ajuste fino das mesmas, a classificação do engajamento é realizada de forma supervisionada. O trabalho treinou as redes para dois e três níveis de engajamento. No primeiro tipo, a rede classificava imagens extraídas dos vídeos da base de dados em **engajado** e **não engajado**, obtendo acurácia de 93,23%, enquanto obteve acurácia de 87,25% quando classificando as emoções em **não engajado**, **moderadamente engajado** e **totalmente engajado**. Os autores relataram utilizar validação cruzada de cinco partes ($k = 5$) para o treinamento, porém não mencionam se respeitaram o princípio da independência dos participantes, o que pode ocasionar imagens de participantes do grupo de teste terem vazado para o grupo de treinamento.

Outro trabalho que utiliza estados afetivos diversos como sinais de detecção de engajamento é o de Nezami et al. (2018). Os autores treinaram uma rede convolucional usando a base de dados FER2013 (Kaggle, 2013; Goodfellow et al., 2013) para reconhecer expressões faciais, e após eles realizaram o ajuste fino da rede para realizar a classificação do engajamento com dados de uma base própria. A detecção facial foi realizada utilizando a ferramenta Dlib-ml (King, 2009), e a rede foi treinada com uma topologia semelhante à VGG, utilizando oito camadas convolucionais (tamanhos 64, 64, 128, 128, 256, 256, 512 e 512), seguida de três totalmente conectadas (tamanhos 4096, 4096, 1024). Os rótulos de saída da base própria utilizada continha informações comportamentais indicando se o aluno estava realizando a tarefa ou não, e emocionais, que indicavam se o aluno estava confuso, entediado ou satisfeito. Portanto, para a detecção do engajamento, os autores tiveram que associar os rótulos às informações binárias de engajado ou não engajado e a acurácia obtida pela detecção foi de 72,38%.

Tipicamente, redes neurais que realizam tarefas de classificação por imagem expressam resultados de saída através de classificação binária, que é quando a classificação de classes resulta nos rótulos **sim** ou **não**, ou classificação categórica, quando o resultado é expresso em um percentual de confiança que a rede possui para a classificação de cada uma das possíveis classes. No entanto, existem situações que a resposta para o problema modelado pode ser expressa em um domínio contínuo. No contexto das emoções, estas situações se encaixam quando a rede não está tentando prever quais emoções estão representadas nas características extraídas e fornecidas na entrada, mas sim com qual **intensidade** que determinada emoção está sendo expressa. Este tipo de problema pode ser resolvido através da modelagem de um regressor, usando redes neurais. Quando classificando características através de regressão, ao invés de acurácia (percentual de acertos comparado da classificação), a métrica utilizada é uma medida de erro, como, por exemplo, o erro médio quadrático.

Algumas bases de dados são preparadas para ambos os tipos de modelagem. Das bases usadas para treinamento de emoções na aprendizagem, a DAiSEE e a EmotiW são dois exemplos que possuem rótulos expressos em intensidades. De acordo com Gupta, D'Cunha, et al. (2016), a utilização de intensidades no processo de anotação de emoções é muito útil, pois mesmo que determinados trabalhos não utilizem informações em formato contínuo, ainda se pode realizar classificação com múltiplas classes expressando as diferentes intensidades da emoção ou até mesmo se decidir simplificar os rótulos, como no trabalho (Dewan et al., 2018).

Os algoritmos que competem na modalidade *Engagement Prediction in the Wild* da EmotiW usam como métrica o erro médio quadrático, como é caso dos trabalhos (Kaur, Mustafa, Mehta, & Dhall, 2018; Thomas, Nair, & Jayagopi, 2018; Yang et al., 2018). O primeiro trabalho (Kaur et al., 2018) foi desenvolvido pelos criadores da base de dados EmotiW, e é um modelo que serve como ponto de partida para os competidores. Nele, os autores extraem as características faciais LBP-TOP, movimento dos olhos, e postura da cabeça utilizando a ferramenta OpenFace (Baltrusaitis, Zadeh, Lim, & Morency, 2018). O regressor é construído utilizando uma rede neural com uma camada LSTM de tamanho 32, seguida de três camadas totalmente conectadas de tamanhos 128, 128 e 100. O modelo obtém um erro médio quadrático (MSE) de $0,1$.

Dos que participaram na edição de 2018 da competição EmotiW, pode-se destacar o trabalho (Thomas et al., 2018), que obteve a segunda colocação. Ele utiliza uma rede convolucional unidimensional temporal (Lea, Flynn, Vidal, Reiter, & Hager, 2017) (TCN - *Convolutional Temporal Network*, do inglês) para realizar a regressão, isto é, ao invés da entrada da camada convolucional receber informações sobre os píxeis em uma imagem, ela recebe um vetor unidimensional com as características extraídas. Esta rede TCN possui três camadas de tamanho 24 cada e difere de uma rede CONV+LSTM ao realizar as operações de convolução ao longo do tempo camada a camada, semelhante às redes do tipo C3D. Para a extração de características, os autores utilizam o OpenFace para extrair movimento dos olhos, postura da cabeça e *Action Units*. A rede obteve um MSE $0,0792$.

Por último, o modelo criado por Yang et al. (2018) cria quatro redes para realizar a detecção da intensidade do engajamento nos alunos. A primeira se utiliza de características extraídas do movimento dos olhos e da postura da cabeça através da ferramenta OpenFace. A segunda extrai as características da postura do corpo utilizando a ferramenta OpenPose. A terceira extrai o LBP-TOP das faces alinhadas obtidas através do algoritmo MTCNN. A quarta utiliza uma rede do tipo C3D para extrair características espaciais. As três primeiras redes são conectadas a duas camadas do tipo LSTM para a extração das características temporais. A rede C3D, porém, como já é construída para lidar com sequências temporais em imagens, não possui camadas recorrentes. Todas as redes possuem três camadas totalmente conectadas na saída, que é onde a predição de cada uma delas ocorre. A fusão dos modelos é realizada utilizando pesos iguais para cada modalidade, e o melhor desempenho obtido pelo modelo no trabalho foi $0,0626$ como MSE.

2.2 Comparação dos trabalhos relacionados com o trabalho realizado

Diante dos conceitos e dos trabalhos apresentados anteriormente, cabe ressaltar que dos trabalhos que realizam a detecção automática de emoções por vídeos da face, foram encontrados poucos trabalhos que realizam a classificação das características faciais em emoções não-básicas presentes em situações de aprendizagem, com exceção do engajamento. Mais especificamente, a detecção das emoções confusão, frustração e tédio esteve presente em somente dois dos trabalhos analisados. Estas quatro emoções são as emoções predominantemente presenciadas quando alunos estão em situações de aprendizagem (D'Mello & Calvo, 2013). Além de sua correlação com o aprendizado, cada uma destas emoções carrega informações importantes sobre o processo de aprendizagem do aluno. Como a maioria dos trabalhos foca somente nas emoções básicas, e dos trabalhos que tratam de emoções em situação de aprendizagem praticamente todos se detém à detecção do engajamento, percebe-se aí a oportunidade da exploração das outras emoções acadêmicas para a

obtenção de um importante conhecimento sobre o aprendizado.

Levando também em conta a correlação entre cada uma destas emoções, e seu fluxo de uma para a outra (D'Mello & Graesser, 2012), é justificável a investigação sobre a manifestação destas emoções em uma escala temporal. No entanto, nenhum trabalho relacionado usou a informação temporal das emoções, ou seja, a sequência de emoções na ordem em que foram manifestas pelo estudante, no treinamento de modelos de detecção de emoções por face. Dessa forma, outro diferencial do trabalho envolve o uso da sequência de emoções dos estudantes para melhorar a capacidade de detecção e previsão do modelo construído.

3 Modelo Proposto

Foi realizado neste trabalho a detecção automática das emoções de aprendizagem (ou acadêmicas) **engajamento, confusão, tédio e frustração** através de vídeo da face dos estudantes. Tais emoções, comumente presenciadas em situações de aprendizagem, possuem forte relação com o engajamento, a eficiência do aprendizado, e o sucesso do aluno (D'Mello & Calvo, 2013). Sua detecção é bastante complexa, pois os traços que demonstram sua presença são muito mais sutis que os presentes nas emoções básicas e sua identificação visual não é tão clara quanto das emoções básicas (Whitehill, Serpell, Lin, Foster, & Movellan, 2014). Por conta disso, existe a necessidade da utilização de uma abordagem que consiga lidar com essas diferenças, como, por exemplo, características auxiliares como movimento dos olhos, importantes indicativos de foco e atenção por parte do aluno, e encontradas, por exemplo, quando os estudantes estão engajados.

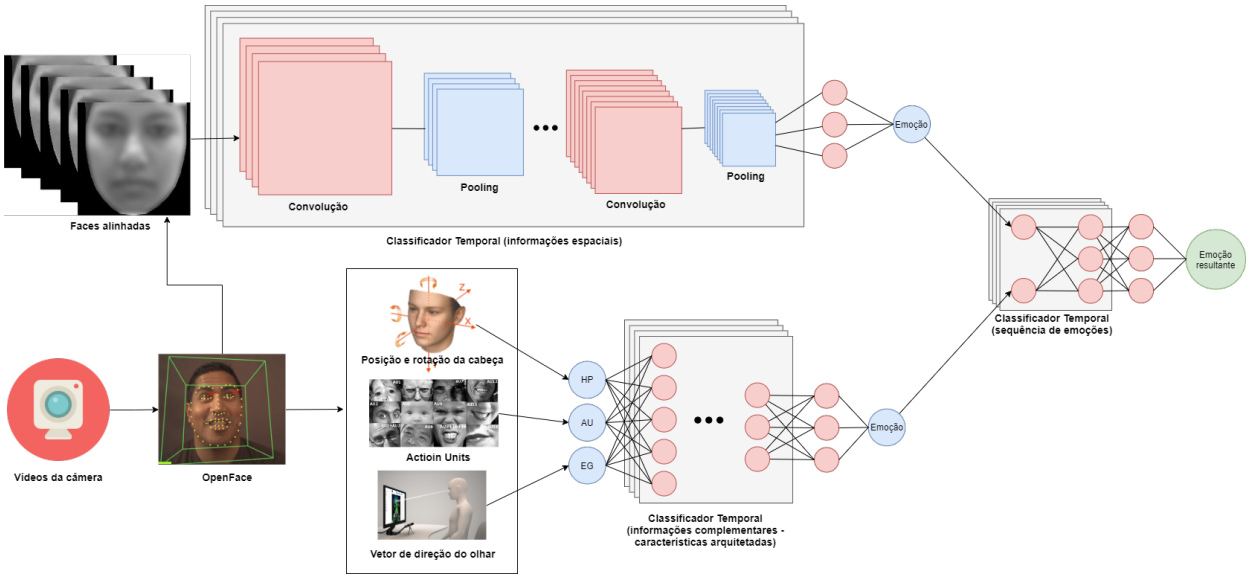
A detecção é realizada automaticamente a partir de vídeos gravados da face dos alunos. Estes vídeos mostram as expressões faciais dos estudantes durante sua interação com ambientes de aprendizagem, obtidos por câmeras (*webcam*) acopladas. A detecção das emoções a partir dos vídeos é realizada através da combinação do uso de ferramentas de detecção e extração de características faciais com o desenvolvimento de um classificador de emoções de aprendizagem que considera tanto a informação temporal da sequência de imagens em um vídeo, bem como a sequência de emoções na ordem em que expressas pelos estudantes. Este classificador utiliza dois conjuntos de informações: (i) informações provenientes de uma rede neural de características extraídas de vídeos, e (ii) informações provenientes de uma rede neural de características arquitetadas.

A rede de características extraídas dos vídeos (i) recebe como entrada os píxeis dos quadros dos vídeos contidos nas bases de dados e, através de camadas convolucionais e recorrentes, obtém como saída durante o processo de treinamento características abstratas que são utilizadas pelo classificador para determinar se um trecho de vídeo dado como entrada contém ou não uma pessoa manifestando uma das emoções de aprendizagem.

A rede de características complementares (ii) é uma rede neural recorrente que utiliza como entrada informações de *movimento dos olhos*, *posição da cabeça* e *action units* obtidas de algoritmos de aprendizagem de máquina. Estas características são fornecidas ao modelo diretamente por valores em arquivos, e a rede recorrente é usada para extrair novas características a partir das relações temporais entre os quadros das imagens.

A Figura 1 demonstra uma visão geral do modelo desenvolvido. Um detalhamento maior

Figura 1: Etapas do módulo de reconhecimento de emoções do modelo implementado.



de cada um dos elementos que o compõe podem ser observados na Seção 3.3.

3.1 Bases de dados

Para a realização do treinamento das redes é necessária a seleção de bases de dados que contenham informações rotuladas segundo o que a rede está se propondo a classificar, dado que para essa tarefa se usa algoritmos de aprendizado supervisionado. O reconhecimento de emoções através de informações visuais, por ser uma área que vêm produzindo uma quantidade razoável de trabalhos, possui consequentemente diversas bases de dados relacionadas. Os primeiros trabalhos propuseram bases compostas de imagens em ambientes controlados, como na Extended Cohn-Kanade (CK+) (Lucey et al., 2010). Outros, na tentativa de melhorar a generalidade de seus modelos, propuseram bases em ambientes não controlados, como a SFEW (Dhall, Goecke, Lucey, & Gedeon, 2011b), composta de imagens estáticas de vídeos da base AFEW (Dhall, Goecke, Lucey, & Gedeon, 2011a), que, por sua vez, contém vídeos de filmes, extraídos da internet, onde pessoas estão atuando e expressando emoções durante os trechos recortados. Trabalhos como de (Dhall et al., 2011a) usam vídeos, ao invés de imagens, para realizar a captura de informações temporais. Outros trabalhos (Li & Deng, 2018) fazem um estudo ainda mais aprofundado sobre bases de dados de emoções básicas, fazendo uma revisão das principais bases utilizadas e suas diferenças, bem como algoritmos utilizados para reconhecimento de expressões faciais, e principais trabalhos da área. Como no presente trabalho foi realizado o reconhecimento de emoções não básicas, estas bases não podem ser utilizadas para o treinamento. Poucas bases específicas para o treinamento de emoções presentes na aprendizagem existem até o momento, e menor ainda é a quantidade das bases públicas. Em virtude da escassez de bases específicas para este objetivo, os autores do presente trabalho usaram todas as bases encontradas que tenham vídeos de pessoas em situação de aprendizagem rotulados com alguma das quatro emoções presentes na aprendizagem: **Engajamento**, **Confusão**, **Frustração** e **Tédio**.

Diante dos requisitos mencionados, são utilizadas para a realização do treinamento das re-

des três bases de dados encontradas. A primeira, chamada *Dataset for Affective States in E-Environments* - DAiSEE (Gupta, D'Cunha, et al., 2016), produzida pelo grupo de pesquisa da *Indian Institute of Technology Hyderabad* e tornada pública, se propõe a auxiliar pesquisas de aprendizado de máquina que buscam resolver problemas relacionados ao engajamento em diversas situações, como, por exemplo, aprendizagem, propagandas, saúde, aprendizado, veículos autônomos, dentre outros. A base contém 9.068 trechos de 10 segundos de vídeos de 112 participantes (32 mulheres e 80 homens) com idades entre 18 e 30 anos. Os vídeos foram gravados em condições diversas. Foram utilizados seis categorias de ambientes e três níveis de iluminação. As anotações dos vídeos foram realizadas através de uma plataforma colaborativa e a confiança das anotações foi garantida por anotações redundantes e remoção de anotadores não confiáveis através de uma análise com especialistas do instituto. Os participantes assistiram os vídeos em frente ao computador e, para cada trecho do vídeo de duração de 10 segundos, anotaram as emoções **engajamento**, **confusão**, **frustração** e **tédio**, bem como suas intensidades. Os anotadores tiveram que relatar para cada uma das quatro emoções, níveis de intensidade entre 1 (muito baixo) e 4 (muito alto).

A segunda base de dados utilizada é a base *Engagement Prediction in the Wild - EmotiW 2018* (Kaur et al., 2018), que se propõe a fornecer uma base de dados para ampliar a compreensão a respeito de problemas relacionados ao aprendizado, tipicamente vistos em estudantes, como perda de interesse, fadiga, tédio, etc. A base consiste em vídeos de aproximadamente cinco minutos que mostram a reação de estudantes assistindo a filmes educacionais. Foram gravados 195 vídeos de 78 participantes (25 mulheres e 53 homens), com idades entre 19 e 27 anos. Os vídeos foram gravados em diferentes tipos de ambiente, iluminação e postura dos participantes. A base possui rótulos que expressam o nível de engajamento de cada vídeo, em uma escala de 0 até 3, onde 0 representa que o participante está totalmente desengajado e 3 que ele está altamente engajado. Foram utilizados cinco anotadores para criar os rótulos da base, onde a confiabilidade da informação fornecida por cada um se baseou na concordância entre os anotadores para os rótulos fornecidos.

Para a complementação do treinamento, a terceira base de dados utilizada provém de experimentos realizados localmente a partir da interação de alunos com um sistema de aprendizagem. Os vídeos para sua composição foram obtidos por estudos de pesquisas desenvolvidos com o sistema tutor inteligente PAT2Math (Jaques et al., 2013), um ambiente inteligente de aprendizagem que assiste estudantes enquanto resolvem equações de primeiro grau passo a passo. Os dados são provenientes do trabalho de Morais and Jaques (2022), que realizaram a gravação de 230 vídeos de 55 alunos (29 feminino e 26 masculino) durante o uso do sistema PAT2Math, dos quais 30 vídeos de alunos diferentes foram usados nesse trabalho. Eles tem entre 12 e 13 anos e são de duas turmas do sétimo ano do ensino fundamental de uma escola privada na grande Porto Alegre. Esses alunos utilizaram o STI PAT2Math durante 10 sessões no laboratório da escola, uma vez por semana, com duração de 40 minutos. Todos os alunos participantes dessa coleta entregaram o Termo de Consentimento Livre e Esclarecido (validado pelo Comitê de Ética da universidade dos autores) assinado por um responsável. Todo o processo de coleta dos dados e anotação das emoções seguiu o protocolo EmAP-ML (Morais et al., 2019) e as emoções expressas pelos estudantes nos vídeos foram anotadas por três anotadores humanos treinados.

3.2 Pré-processamento

O pré-processamento é uma importante fase do treinamento de algoritmos de aprendizado de máquina. Nesta fase, os dados usados como entrada dos modelos devem ser tratados para que o modelo possa aprender adequadamente. Algoritmos que realizam o treinamento são extremamente sensíveis aos dados recebidos, e certas técnicas já se mostraram eficazes em transformar os dados de entrada de maneira que preservem as informações relevantes enquanto reduzam a complexidade ou alterem a formatação das mesmas.

Caracteriza-se como pré-processamento técnicas de normalização dos dados, *sampling*, remoção de informações com baixa correlação às variáveis desejáveis no treinamento, tratamento de desbalanceamento, correção de falhas, ruído, ou informação faltante nos dados de entrada, adequação de dimensões das informações, dentre outros.

3.2.1 Tratamento dos vídeos

Para trabalhar com os vídeos presentes nas bases de dados, alguns ajustes foram feitos para garantir um melhor resultado na classificação das emoções e um menor tempo de treinamento das redes. Entendendo que a manifestação da emoção é um fenômeno que possui determinado tempo de duração, os vídeos foram quebrados em trechos de dois segundos de duração. Este período específico foi escolhido dentre diversos outros períodos testados por duas razões principais: vídeos de mais curta duração possuem menos requisito computacional para realização do treinamento e também pelo fato de que para o problema analisado, esta janela de tempo foi a que mostrou melhor desempenho em questão da evolução da função de perda. Os rostos presentes nos vídeos foram detectados utilizando o algoritmo MTCNN, implementado pela ferramenta OpenFace. A partir destes rostos detectados, as faces alinhadas foram extraídas, e as imagens resultantes foram redimensionadas para possuírem tamanho 224x224 píxeis e um canal de cor (escala de cinza). A decisão pelo tamanho das imagens é baseada no tamanho mínimo necessário para realizar o ajuste fino de uma rede da topologia ResNet (He, Zhang, Ren, & Sun, 2016). Conforme visto na Seção 3.1, como cada vídeo da base de dados DAiSEE possui 10 segundos de duração, cada vídeo foi quebrado em 150 quadros, pois se optou por utilizar a taxa de atualização de 15 quadros por segundo pelo motivo de ser uma taxa de atualização mínima em que ainda existe uma certa fluidez do vídeo. Como existem 9.068 trechos de vídeo, o resultado foi a obtenção de 1.360.200 imagens. A base PAT2Math possui um único vídeo por cada aluno (ao total, são 30 alunos). Embora cada vídeo possua duração variável (variando de 30 a 45 minutos), os rótulos estão disponíveis para apenas cinco minutos de cada vídeo. Consequentemente, somente pode ser utilizado cinco minutos de cada vídeo do PAT2Math, o que resultou na extração de 214.602 imagens. A base EmotiW possui 195 vídeos, um para cada indivíduo, com duração aproximada entre 4 e 6 minutos cada, e destes vídeos foram extraídos 2.723.342 imagens. A Tabela 1 relaciona as bases com o número de vídeos presentes em cada uma, o tamanho dos vídeos, e o número de imagens resultantes da extração das faces alinhadas dos vídeos.

3.2.2 Balanceamento de Classes

Uma característica presente em todas as bases de dados existentes que lidam com as emoções no aprendizado é o desbalanceamento das classes, que é quando o número de exemplos em cada

Tabela 1: Comparativo entre imagens resultantes da extração dos quadros dos vídeos das bases de dados utilizadas.

Base	**Vídeos**	**Duração(s)**	**Imagens extraídas**
DAiSEE	9068	10	1.360.200
PAT2Math	30	60-120	214.602
EmotiW	196	240-360	2.723.342

classe (cada emoção detectada, neste trabalho) se diferem consideravelmente. No caso do presente trabalho, nas três bases empregadas, a emoção engajamento possui um severo desbalanceamento em relação à confusão, tédio e frustração. De fato, mesmo ao analisar cada emoção individualmente, de um ponto de vista de classificação binária (emoção presente em comparação a emoção ausente), as quatro emoções possuem desbalanceamento em todas as bases observadas, onde o engajamento possui mais casos positivos, e as outras três emoções mais casos negativos. Este fato se deve à natureza do ato da gravação de pessoas lendo ou assistindo conteúdos específicos em frente ao computador. Nesta situação, desde que as seções não sejam demasiadamente longas, as pessoas costumam apresentar um engajamento predominante em relação às outras emoções. Por exemplo, a base DAiSEE possui 94,15% dos vídeos com rótulos de intensidade dois ou três para emoção engajamento (valores variam entre zero e três), enquanto somente 4,83% da mesma intensidade para emoção frustração. Da mesma forma, na base EmotiW, 72,45% dos vídeos são rotulados como de engajamento e tendo intensidade entre 0,66 ou 1 (valores variam entre 0 e 1). O desbalanceamento de classes é um problema bastante comum na mineração de dados e aprendizagem profunda, tendo já sido relatado por diversos trabalhos (Longadge & Dongre, 2013; Wang et al., 2016; Krawczyk, 2016), que igualmente abordam alternativas específicas para alguns casos.

Para combater o desbalanceamento de classes nas bases de dados, existem algumas técnicas que podem ser implementadas. Tais técnicas se baseiam em três abordagens: (i) a manipulação dos dados de entrada, (ii) elaboração de algoritmos que manipulam o viés da classificação, e (iii) a utilização de características complementares aos dados de entrada de maneira que auxilie o classificador.

As técnicas de *oversampling* e *undersampling* são exemplos da abordagem (i) e consistem em alterar a quantidade de dados usados no treinamento da rede neural. Para aplicar o *undersampling*, reduz-se a quantidade de exemplos de entrada das classes mais representadas para equilibrar a proporção entre todas as classes da base de dados. O critério para exclusão dos exemplos mais representados pode ser aleatório, ou seguindo algum princípio com a finalidade de manter a diversidade das informações das classes afetadas. A grande desvantagem do *undersampling* é que, ao remover exemplos, perdem-se informações valiosas que poderiam ser utilizadas para qualificar o aprendizado do algoritmo.

Ao utilizar o *oversampling*, informações das classes menos representadas precisam ser multiplicadas. Este efeito pode ser obtido de diversas maneiras, dentre elas a escolha aleatória de exemplos para serem replicados, ou a replicação de exemplos seguindo algum critério que garanta diversidade das informações, de maneira semelhante à aplicada no *undersampling*. Todavia, este tipo de abordagem frequentemente conduz ao *overfitting* da rede, pois a mesma aprende a identificar os padrões específicos dos exemplos fornecidos (Elrahman & Abraham, 2013). Um método que visa reduzir este problema é a geração de dados sintéticos, que consiste em utilizar os exemplos das classes menos representadas para a geração de novos exemplos, diferentes dos originais.

Para realizar esta geração, são normalmente aplicados algoritmos que realizam o deslocamento, rotação ou algum tipo de modificação da imagem original.

Outra técnica que visa combater o desbalanceamento dos dados é a aplicação de pesos ao aprendizado, representando a abordagem supracitada (ii). Embasando-se no princípio que para certos problemas é esperado que hajam poucas ocorrências de certas classes, e que as eventuais ocorrências das mesmas devem ser identificadas inequivocadamente, como em diagnósticos médicos de doenças severas ou detecção de fraudes em documentos, é natural o raciocínio que o algoritmo de aprendizagem atribua importâncias distintas a cada classe. Desta forma, a função de perda de uma rede neural penaliza de maneira mais severa os erros de classificação das classes menos frequentes comparativamente às outras classes.

Um desafio a esta abordagem é a definição de critérios para a atribuição dos pesos das classes. Uma técnica com ampla utilização é a atribuição dos pesos conforme a proporção inversa dos exemplos de cada classe. Esta técnica possui como limitações sua adequação exclusiva à proporção do conjunto de treinamento, que caso não possua uma correspondência adequada à representação das classes de exemplos da realidade, não necessariamente fará com que o modelo atinja um bom desempenho (W. Huang, Song, Li, Hu, & Xie, 2013). Em seu trabalho, W. Huang et al. (2013) ainda citam um modelo alternativo que utiliza um algoritmo evolutivo para a determinação dos pesos.

Por fim, outra alternativa à resolução do problema do desbalanceamento dos dados é a utilização de características complementares extraídas das bases de dados, que segue a abordagem (iii). Este método está embasado na premissa que a obtenção de outras características não relacionadas com as principais enriquecem as informações de treinamento, tornando mais precisa a classificação dos exemplos menos frequentes. Estas características complementares provém de algoritmos cujas teorias não estão relacionadas entre si, ou ainda da obtenção de dados de outras fontes, como, por exemplo, áudio ou ondas cerebrais.

Os autores utilizaram no presente trabalho todas as três técnicas em seus modelos, embora nem todas as técnicas foram usadas em todos. Após definidas a quantidade de segundos a ser utilizada para cada trecho de vídeo e a quantidade de amostras de imagens por segundo de vídeo, foi implementado o método de *random undersampling* para remover exemplos da classe mais representada. Para este método, determinou-se uma proporção máxima que os vídeos com rótulos da classe mais representada apareceriam. Então, foram removidos vídeos aleatoriamente do grupo de treinamento que possuía rótulos da classe mais representada até que a proporção desejada fosse atingida.

Nos modelos de segunda geração foram realizadas algumas tentativas para configurações de *undersampling*, todas entre 66:100 e 100:100. Este valor representa a proporção de exemplos da classe menos representada para a mais representada. Nos modelos das gerações três e quatro, este método não foi empregado por não ter sido notado melhoria significativa em seu uso. Uma possível explicação para isto é o fato de ter reduzido drasticamente o conjunto de dados disponíveis para treinamento.

Para o *oversampling*, duas abordagens diferentes foram utilizadas. A primeira, semelhante ao método utilizado no *undersampling*, pegou os vídeos da categoria menos representada e os selecionou aleatoriamente para replicação até que a proporção desejada fosse obtida. Na segunda, foram criados exemplos sintéticos através do algoritmo SMOTE (Chawla, Bowyer, Hall, & Ke-

gelmeyer, 2002), sendo bastante utilizado para a tarefa. Após a aplicação de ambas as técnicas, o desbalanceamento do conjunto de treinamento foi significantemente reduzido.

Nos modelos de segunda geração foram realizadas algumas tentativas para configurações de *oversampling*, todas entre 13:100 e 25:100. Pela mesma razão que nos casos de *undersampling*, o *oversampling* não foi empregado nas gerações subsequentes de modelos.

Outra técnica utilizada foi a aplicação de pesos de treinamento. Das três técnicas, a aplicação de pesos foi a que mostrou melhores resultados, tendo sido utilizada até o modelo final. Foram realizadas tentativas de aplicação de pesos em diversas formas, desde valores fixos entre 3:1 e 50:1 até valores inversamente proporcionais à representatividade das classes no conjunto de treinamento. Este último se mostrou mais efeito e dinâmico, pois quando o número de exemplos de treinamento variava, o peso automaticamente era recalculado.

3.2.3 Características complementares

Características complementares foram usadas como dados do movimento dos olhos, *Action Units* e posição da cabeça dos participantes. Tais características podem ser classificadas em dois grupos: as características descobertas e as arquitetadas (tradução livre de *engineered features*). As características descobertas são aquelas que não possuem intervenção humana para sua criação. Elas são descobertas nas imagens através do processo de treinamento de uma rede convolucional. As características arquitetadas possuem seu conceito baseado em alguma teoria, e normalmente existem técnicas e algoritmos de visão computacional para extraí-las das imagens.

Nos modelos desenvolvidos foram utilizadas as faces alinhadas para a extração de características a serem descobertas, assim como as seguintes características arquitetadas:

- *Action units (AUs)* - Sistema de codificação de características faciais que representam relaxamentos e contrações de diferentes músculos do rosto. Representadas como um valor binário indicando presença ou ausência de cada uma das AUs.
- *Eyetracking* - Direção na tela para onde os olhos estão fixos, representados por coordenadas cartesianas. Representados por valores entre 0 e 1 para coordenadas X, Y e Z e para cada olho. Este limite indica os extremos opostos do campo visual capturado pela câmera. Este ponto no espaço indica para onde cada olho está direcionado.
- *Head pose* - Direção para onde a cabeça está virada. Representado por seis valores: Rotação X, Y, Z, e Translação X, Y e Z. A rotação em cada coordenada indica a rotação da cabeça, e a translação indica a posição no espaço.

Para a extração das características mencionadas, foi optada pela utilização da ferramenta OpenFace pelo seu notável desempenho obtido na extração de características faciais (Baltrusaitis et al., 2018). A ferramenta implementa uma série de algoritmos de visão computacional que demonstraram ótimo desempenho na extração das diversas características.

Além das características acima mencionadas, visando obter características faciais não correlacionadas com as acima mencionadas, características espaço-temporais foram extraídas diretamente dos vídeos das faces alinhadas. Características espaço-temporais são aquelas obtidas das

informações espaciais ao longo do tempo, ou seja, informações que possuem componentes bi ou tridimensionais, como imagens, e também se deslocam ao longo do tempo, como vídeos. O modelo da rede neural que realizou esta extração será explicado na Seção 3.3. Estas características possuem relação com informações abstratas extraídas diretamente dos píxeis que compõe as imagens, e da transição destas imagens que formam os vídeos.

3.3 Arquitetura Genérica do Modelo Desenvolvido

Em posse das características extraídas e dos trechos de vídeos selecionados, foi criado um classificador capaz de realizar o reconhecimento de uma das seguintes emoções de aprendizagem: **engajamento**, **confusão**, **frustração** ou **tédio**. Este reconhecimento se deu através de uma fusão de redes neurais implementadas utilizando algoritmos de aprendizado supervisionado de máquina. As redes neurais profundas têm se mostrado bastante eficientes na tarefa de detecção de padrões, obtendo resultados na tarefa de classificação até mesmo próximos aos resultados obtidos por humanos em certos casos (Goodfellow, Bengio, & Courville, 2016). Outro argumento a favor deste tipo de técnica é que, normalmente, para tarefas de classificação de emoções em faces, os trabalhos que costumam obter os melhores desempenhos empregam *Deep Learning* (Li & Deng, 2018). Por este motivo, as redes do modelo foram implementadas utilizando esta técnica, em diferentes tipos de configurações.

Um modelo de rede neural é normalmente composto de um extrator de características e de um classificador. O extrator de características possui o propósito de receber informações do mundo real, e através do algoritmo de propagação de pesos encontrar relações abstratas entre as informações fornecidas. Estas informações abstratas são chamadas características e, embora sejam de difícil compreensão para os humanos, elas são essenciais para a tarefa de classificação. Quanto mais complexa e profunda a rede neural, maior o nível de abstração das características que ela conseguirá obter.

O algoritmo classificador é responsável por receber as características inferidas pelo extrator e realizar a classificação das mesmas nos rótulos de saída utilizando o conjunto de treinamento para isso. Através da inferência de padrões relacionando valores das características de entrada com os rótulos esperados do conjunto de treinamento, espera-se que o classificador consiga obter regras com um potencial de generalização que sejam úteis na classificação de informações que não estejam presentes no conjunto de treinamento.

Um dos elementos que os modelos desenvolvidos neste trabalho possuem em comum é a ideia da utilização da fusão de diversas redes a fim de obter um modelo mais robusto. Para este fim, duas classes de modelos de *Deep Learning* foram utilizados. Cada uma destas classes de modelos pode ser implementada separadamente, como de fato aconteceu nas etapas preliminares deste trabalho. Porém, para a criação de uma fusão de redes, é necessário eliminar as camadas do topo de cada rede, responsáveis pela classificação, e unir ambas em uma camada de concatenação. Após, cria-se outro topo para a nova rede para realizar a associação das características concatenadas obtidas de ambas redes com os rótulos.

3.3.1 Rede de características complementares

O **primeiro tipo de rede** concatenada no mesmo modelo recebe três categorias de características previamente extraídas das faces alinhadas contidas nos vídeos: *action units*, *eye gaze* e *head pose*. O emprego do movimento dos olhos (*eye gaze*) é uma característica comumente associada ao aprendizado (Lai et al., 2013), enquanto o modelo FACS (Ekman, 1992), que descreve todas as movimentações de músculos faciais, ou Action Units, é um dos modelos de expressões faciais mais difundidos. Tais características, portanto, carregam informações valiosas sobre as emoções demonstradas por estudantes durante situações de aprendizagem, tornando assim características importantes para complementação do classificador do trabalho proposto.

Este primeiro tipo de rede neural é chamado de rede de características arquitetadas, ou complementares. Ele consiste em uma topologia específica de rede neural (detalhados na Seção 3.2.3), que recebe como entrada as informações extraídas do *OpenFace* e, através da utilização de camadas temporais (descritas em detalhes em cada modelo específico), fornece para o classificador características temporais extraídas a partir destes dados.

Um classificador temporal possui o papel de identificar os padrões presentes ao longo do tempo para cada uma das características inseridas como entrada e como resultado identifica as emoções de aprendizagem. A utilização de informações temporais na classificação se dá pelo fato de que imagens estáticas carregam menos informações sobre um estado afetivo de um indivíduo que sua análise temporal (Liu et al., 2018). Além disso, se tratando de emoções não básicas, sua detecção em um ambiente pouco propício a demonstrações de emoção, como durante a utilização de um ambiente virtual de aprendizagem, se torna ainda mais difícil. Para tal, a informação temporal visa amplificar a capacidade de reconhecimento de qualquer sinal que indique uma demonstração de emoção.

3.3.2 Rede de características espaciais

O **segundo tipo de rede** concatenada no mesmo modelo implementado utiliza a extração de características diretamente das faces alinhadas presentes nos vídeos através de redes convolucionais profundas. Independente da arquitetura e topologia específica, o processo de treinamento deste tipo de rede requer a entrada dos valores dos píxeis das imagens em estruturas convolucionais, variando em suas especificidades de acordo o modelo em questão. A ideia por trás de uma rede convolucional, ou de características espaciais, é que cada camada convolucional, através dos filtros, capture as informações mais importantes em blocos específicos da imagem, e as camadas *pooling* realizem a redução da dimensionalidade destas características. Por fim, é apresentado como resultado um vetor unidimensional das características encontradas na imagem. O processo de treinamento, além de requerer a engenharia da arquitetura, requer recursos computacionais significativos, segundo a largura de cada camada, a profundidade da rede e o tipo de operação realizada pelos nós da rede.

Para reduzir o problema da alta de demanda de recursos, pode-se empregar a transferência de aprendizado. Neste tipo de abordagem, utiliza-se uma rede originalmente treinada em uma quantidade grande de dados (normalmente de imagens) para classificar um conjunto específico de rótulos. No processo da transferência de aprendizado, removem-se as últimas camadas desta rede (chamadas de topo), responsáveis pela classificação, isto é, a associação das características

Tabela 2: Comparativo entre as redes convolucionais utilizadas em cada geração de modelos desenvolvidos.

Ger.	**Rede(s)**	**Pré-treinam.**	**Descrição**
1ª	VGG-Face	Sim	Classificador GRU
1ª	Conv3D	Não	Camadas sequenciais.
2ª	Conv3D	Não	Arquitetura com *skip connections*.
2ª	BLSTM + Conv2D	Não	Fusão BSLTM.
3ª	InceptionResnetV3 + ConvLSTM2D	Sim	Redes em sequência.
4ª	InceptionResnetV3 + ConvLSTM2D + TCN	Sim	Histórico de emoções incorporado com rede TCN.

abstratas descobertas nas camadas do fundo da rede com os rótulos apresentados. Em seguida, adiciona-se um novo conjunto de camadas de topo (que não receberam treinamento) bem como novos rótulos. A expectativa é que durante o processo de treinamento a rede utilize as características do fundo da rede no treinamento do novo topo e na classificação dos novos rótulos.

Na topologia desenvolvida neste trabalho, deseja-se realizar o *fine tuning* com vídeos em uma rede originalmente treinada para receber imagens, portanto, deve-se encapsular a rede pré-treinada com camadas recorrentes, fazendo com que o treinamento e classificação se dê também na dimensão do tempo. A vantagem do **fine-tuning** sobre da implementação fim-a-fim da rede convolucional profunda é o aproveitamento das informações aprendidas na rede complexa e treinada em uma base de dados extensa, o que demanda consideravelmente menos recursos computacionais e menos tempo.

Para esta segunda rede, foram utilizadas ambas abordagens, implementando redes baseadas em arquiteturas de redes que comprovadamente obtiveram bons desempenhos em trabalhos passados para tarefas relacionadas. Destas, podem-se citar, por exemplo, ResNet (He et al., 2016), DenseNet (G. Huang, Liu, Van Der Maaten, & Weinberger, 2017) e VGG (Simonyan & Zisserman, 2014). Estas redes implementadas desde seu princípio tiveram lugar nos modelos iniciais. Além do treinamento completo das redes, também foram treinados modelos baseados em redes pré-treinadas do tipo Inception-ResNet (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017), Inception (Szegedy et al., 2016), ResNet e Vgg. A Tabela 2 demonstra o comparativo entre os modelos convolucionais utilizados nos treinamentos. A descrição detalhada de cada geração de modelos implementados também pode ser vista nas Seções 3.5, 3.6, 3.7 e 3.8.

Para tratar da temporalidade das emoções, isto é, a hipótese de que as emoções não são isoladas umas das outras, e que experiências anteriores colaboram para a manifestação de emoções futuras, foram utilizados nos modelos mais avançados uma camada temporal superior em todas as redes. Isto se deu através da utilização de diversos tipos de redes recorrentes para modelar uma quarta dimensão nas entradas espaciais, e uma terceira dimensão nas entradas das características arquitetadas.

Para o treinamento de todos os modelos desenvolvidos durante o trabalho, foram utilizados rótulos para a classe que representa a emoção engajamento dos estudantes. A comparação do desempenho e escolhas quanto à evolução do desenvolvimento dos modelos foram todas tomadas tendo como base as respostas obtidas por modelos treinados para classificar o engajamento.

Posteriormente, os modelos que obtiveram melhores resultados na classificação do engajamento foram também treinados para classificar as classes de frustração, tédio e confusão. Esta decisão foi tomada para dar prioridade no treinamento dos modelos que atingiam melhor desempenho, e agilizar a evolução do desenvolvimento dos modelos seguintes. Embora não tenham sido realizados estudos a respeito, espera-se que modelos que mostrem bom desempenho na classificação de uma das emoções de aprendizagem também desempenhem bem quando utilizados para treinar outras emoções de aprendizagem, como visto em (Gupta, D'Cunha, et al., 2016; Gupta, Jaiswal, Adhikari, & Balasubramanian, 2016), por exemplo.

3.3.3 Fusão de modelos

Em vários dos modelos desenvolvidos, por haver mais de uma rede que compõe o modelo, existe a necessidade da implementação de um método de fusão das redes. Durante as etapas iniciais do trabalho, as redes de convolucionais e de características complementares foram treinadas individualmente. Para compor o modelo de fusão, foi desenvolvida uma terceira rede que recebia como entrada os rótulos de saída das duas redes previamente treinadas.

O segundo tipo de fusão de modelos foi implementado através da concatenação das saídas das redes convolucionais e de características complementares. Nesta etapa, o classificador da rede era disposto após a fusão, e as saídas de cada rede eram as características abstratas encontradas durante o respectivo processo de treinamento. Deste modo, o classificador recebe dados muito mais complexos para realizar seu treinamento, que agora é feito de maneira integrada ao modelo como um todo.

O método de fusão implementado nos últimos modelos segue o princípio do treinamento em conjunto com as redes principais, com o diferencial que além das características descobertas nos modelos principais, este método de fusão também recebe como entrada as emoções anteriores experienciadas. Desta forma, neste estágio, o treinamento se dá pela concatenação de quatro conjuntos de características não relacionadas.

3.4 Treinamento

O treinamento das redes foi realizado utilizando a validação cruzada de k partes (*k-fold crossvalidation*, do inglês). Os exemplos de treinamento e validação foram unificados e então divididos em k partes, ondes as primeiras foram usadas para treinamento e uma para validação. O conjunto original de testes das bases usadas não foi alterado para posterior comparação dos resultados do trabalho proposto com modelos do estado da arte. Os valores 3, 5 e 10 de k foram testados, pois o conjunto de treinamento deve ser maior que o de validação e teste, porém não tão grande que deixe estes dois outros conjuntos com poucos exemplos de cada categoria. Após, o treinamento foi realizado novamente permutando os grupos, de maneira que se obteve k resultados distintos. O resultado final do modelo é a média dos desempenhos dos k treinamentos. Este tipo de técnica é utilizada para minimizar as chances de uma seleção dos grupos de treinamento e validação que faça com que o modelo obtenha um desempenho enviesado. As seleções das partes foram realizadas respeitando os critérios de independência dos indivíduos, onde todas as amostras de vídeos de um determinado participante estão contidas na mesma parte (*split*) da seleção dos dados. Além disso, foi tomado o cuidado para que cada parte possuísse uma quantidade balanceada de cada

classe, pois como a distribuições das classes nestas bases de dados é desbalanceada, a seleção aleatória poderia fazer com que algum dos conjuntos de treinamento ficasse sub-representado em alguma classe.

Para o presente trabalho, foram desenvolvidos diversos modelos, explicados na Seção 3.3, e disponíveis em (Werlang, 2022). Os conjuntos de modelos foram divididos em quatro gerações, reunidas por características em comum, descritas a seguir:

- **Primeira geração** - Modelos compostos por redes treinadas independentemente, e um método de fusão que usa um modelo treinado com as saídas das redes anteriores.
- **Segunda geração** - Modelos construídos com inspiração nas conexões de redes residuais (*skip connections*). Fusão realizada através da concatenação das características extraídas das redes anteriores.
- **Terceira geração** - Uso de rede pré-treinada InceptionResNetV3 + ConvLSTM e TCN na construção dos modelos convolucionais e de características complementares. Uso da métrica F1 para comparação entre modelos.
- **Quarta geração** - Implementação do histórico de emoções nos modelos.

3.5 Primeira geração de modelos

Os modelos desenvolvidos durante a fase inicial de implementações, ou primeira geração, são modelos caracterizados pelo trabalho independente de cada rede e pelo método de fusão desconectado do treinamento das redes. Os modelos da primeira geração também possuem o objetivo de testarem os desempenhos de cada tipo de arquitetura específica isoladamente, antes da fusão. A Figura 2 demonstra uma visão geral da arquitetura dos modelos de primeira geração implementados neste trabalho.

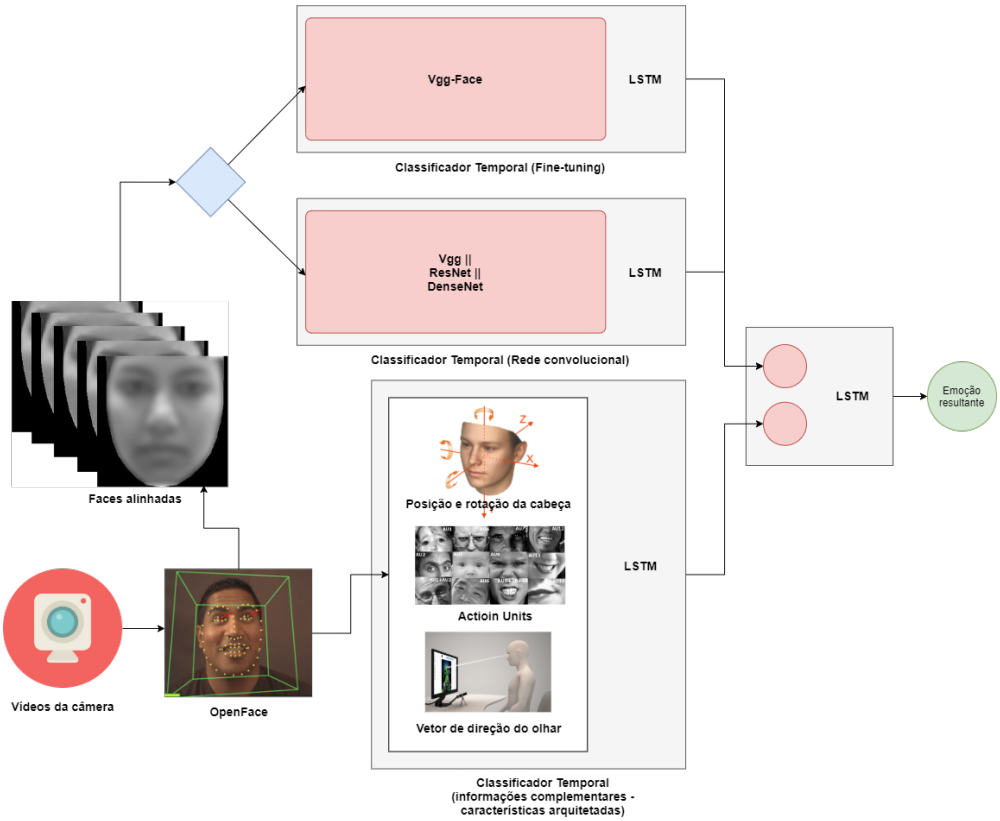
As entradas de todos os modelos de todas as gerações são obtidas dos vídeos das bases de dados. Os modelos convolucionais usam faces alinhadas obtidas das imagens extraídas dos quadros dos vídeos. Já os modelos de características complementares usam as informações das *Action Units*, vetores de posição dos olhos e posição e rotação da cabeça, vindos das mesmas imagens dos quadros dos vídeos.

Os modelos de características complementares de primeira geração usam redes recorrentes do tipo LSTM para extrair padrões temporais das características complementares das imagens. Através do treinamento com uso de memória, característico das redes recorrentes, as informações das imagens são consideradas ao longo do tempo, obtendo assim uma rede temporal. Foram empregadas camadas LSTM sequenciais para este fim.

Já os modelos convolucionais da primeira geração podem ser divididos em dois tipos: aqueles em que foi realizado o ajuste fino de modelos pré-treinados e os modelos em que a rede foi construída e treinada desde seu princípio. Destas duas propostas concorrentes, somente o melhor modelo seria o escolhido para a etapa seguinte.

Na primeira etapa, a principal métrica utilizada para determinação do melhor modelo foi a acurácia, que é a métrica presente na maioria dos trabalhos correlatos analisados.

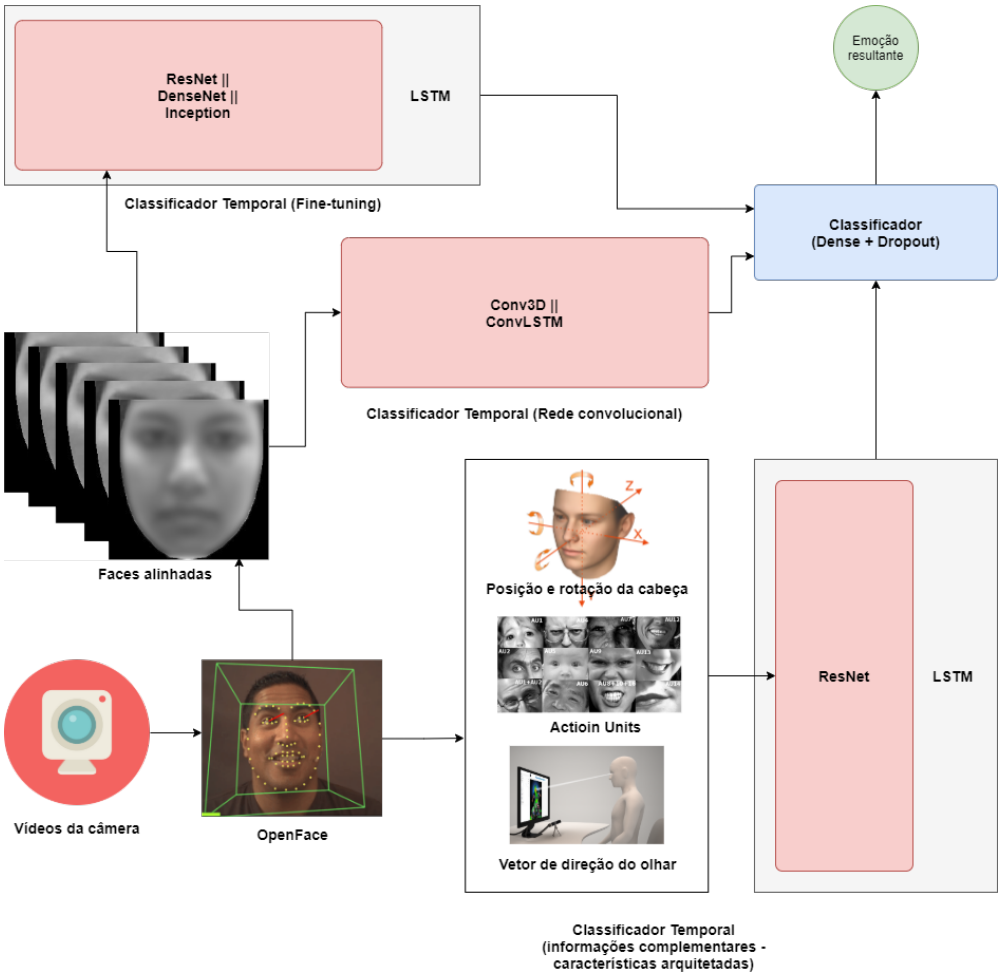
Figura 2: Arquitetura dos modelos de primeira geração: treinamento de redes convolucionais temporais e rede temporal de características arquitetadas. Fusão usando saídas dos modelos escolhidos.



Na etapa de fusão, a saída do melhor modelo convolucional e do melhor modelo complementar foram usadas como entrada de um novo modelo temporal. Ao utilizar a saída de modelos temporais (imagens temporais, ou vídeos) como entrada de outro modelo temporal, o resultado esperado era a consideração do histórico de emoções detectadas nos vídeos de um mesmo indivíduo na criação de um perfil abstrato deste indivíduo. O objetivo do modelo recorrente era conseguir detectar padrões nas saídas dos conjuntos de vídeos do indivíduo para criação destes perfis.

Dos modelos de primeira geração, poucos obtiveram resultados apresentáveis, pois na maioria dos modelos, as curvas de perda e acurácia não convergiam. Destaque pode ser dado para um modelo que realiza o ajuste fino de uma rede pré-treinada na base VGG-Face, e utiliza a arquitetura da rede VGG16. Este modelo utilizou os pesos da rede pré-treinada e ligou-as a camadas do tipo GRU e ao classificador. O modelo obteve 63,62% de acurácia. Outro modelo implementado nesta etapa utilizou camadas do tipo Conv3D (Tran et al., 2015) sequenciais. Este tipo de rede realiza convoluções em entradas de dados de três dimensões ao invés de duas, como as camadas convolucionais tradicionais. Na rede em questão, a terceira dimensão utilizada foi a sequência de imagens, caracterizando quadros do vídeo. Esta rede obteve 74,92% de acurácia na classificação do engajamento dos estudantes. Embora pareça um bom resultado, este modelo, como todos os modelos que antecedem a terceira geração, não aprendiam. Isto é, não apresentavam evolução significativa, independente da quantidade de etapas de treinamento que eram submetidos.

Figura 3: Arquitetura dos modelos de segunda geração: redes inspiradas em arquitetura ResNet e fusão de modelos integrada no treinamento.



3.6 Segunda geração de modelos

Os modelos de segunda geração foram implementados em decorrência do baixo desempenho dos modelos anteriores, denominados de modelos de primeira geração, na tarefa de detecção de emoções presentes na aprendizagem. Os modelos de segunda geração possuem como principais características a implementação de redes mais complexas, mais demandantes de recursos computacionais e com inspiração nas ligações dos modelos de rede residual (ResNet), além de uma rede de fusão integrada ao treinamento das redes que compõem o modelo, conforme ilustra a Figura 3.

Os modelos de ajuste fino de segunda geração usam redes pré-treinadas do tipo ResNet, DenseNet e Inception. Diferente da rede anterior, uma rede VGG16, estas três redes não foram treinadas na base VGGFace e sim na ImageNet (Deng et al., 2009), a qual é uma base de dados de imagens genéricas. Embora tenha sido usada uma base de dados menos específica para o problema de reconhecimento de emoções em faces, era esperado que estas redes performassem melhor que as redes de ajuste fino da primeira geração por que a arquitetura destas redes é mais complexa e com mais parâmetros treináveis, o que geralmente leva a melhores resultados quando treinando padrões complexos. Porém, dada a complexidade da tarefa de reconhecimento de emoções de aprendizagem, e a base de dados altamente desbalanceada, estas redes também não obtiveram

convergência nos resultados.

A segunda rede implementada para compor o modelo de segunda geração foi uma rede convolucional implementada e treinada sem utilização de pesos de redes pré-treinadas. Foi realizada uma nova implementação com redes do tipo Conv3D, que obteve 73,71% de acurácia. Desta vez, ao invés de camadas sequenciais Conv3D, foi utilizada uma arquitetura inspirada na ResNet, onde camadas convolucionais convencionais foram substituídas pelas Conv3D, resultando em uma rede que recebe vídeos como entrada, e realiza convoluções ao mesmo tempo que realiza o treinamento temporal. Esta abordagem difere da rede Conv+LSTM convencional porque na primeira, características temporais são aprendidas a partir da convolução em cada instante do vídeo, enquanto que na rede Conv+LSTM o treinamento temporal é realizado sobre os pesos das camadas convolucionais convertidos em unidimensionais, no processo chamado achatamento, o que resulta na rede aprender características temporais do conjunto de imagens.

Para o desenvolvimento da rede de características complementares de segunda geração também foi utilizado o conceito de atalhos nas conexões (*skip connections*). Camadas LSTM bidirecionais (BLSTM) foram agrupadas em blocos e conectadas de maneira paralela, de maneira semelhante à realizada na rede Conv3D. Esta rede recebeu como entrada as características do OpenFace, e obteve 85,03% de acurácia.

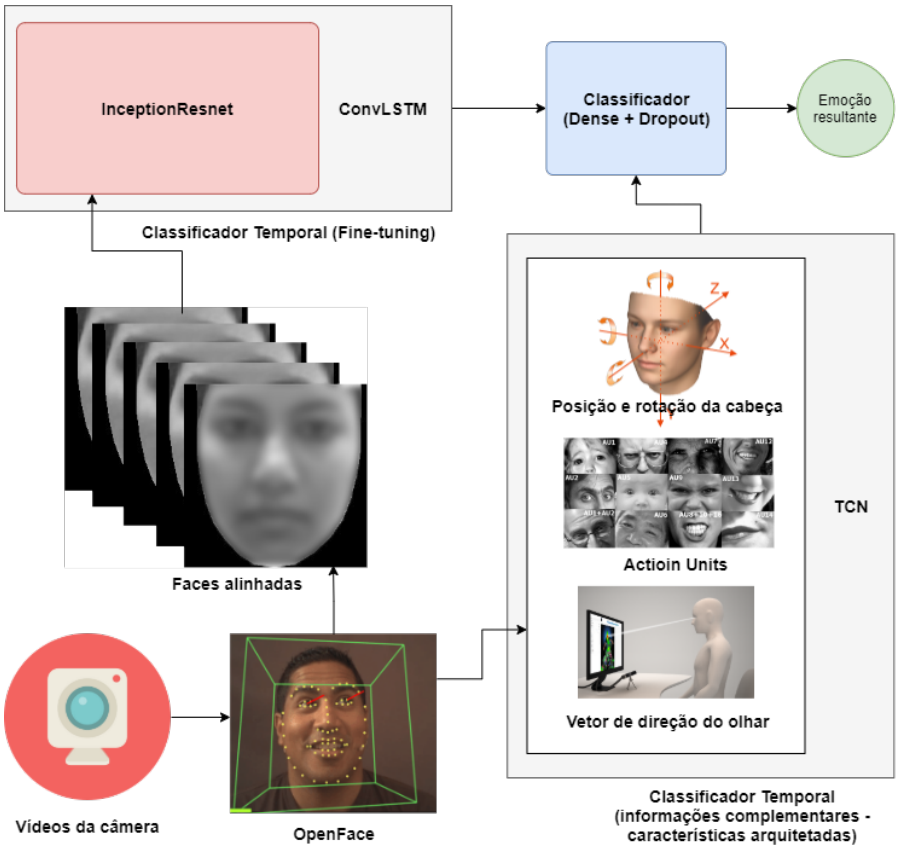
O próximo grupo de modelos desenvolvidos visou aprimorar o método de fusão previamente desenvolvido. Tomando por preceito que o classificador pode encontrar padrões em características não correlacionadas, foram desenvolvidos modelos que unem redes que trabalham com vídeos e redes de características complementares. Cada uma das duas redes teve o seu classificador removido, isto é, o topo da rede, responsável por encontrar os padrões nas características descobertas pelas camadas anteriores. As características descobertas pelas duas redes então foram unidas em uma camada de concatenação, e um único classificador foi usado para este novo modelo. Desta forma, ambas redes foram treinadas como uma só. Este modelo obteve 67,33% de acurácia. Embora este não tenha sido o modelo que melhor desempenhou segundo a métrica de acurácia, dois pontos devem ser considerados a respeito: (*i*) A acurácia, embora tenha sido usada como métrica dos modelos de primeira e segunda geração, não se mostra adequada para bases de dados que possuem um desbalanceamento muito grande. Modelos desenvolvidos posteriormente pelo presente trabalho solucionam esta limitação. (*ii*) A curva de acurácia foi progredindo ao longo do tempo e o modelo foi melhorando a cada época de treinamento, demonstrando que houve um aprendizado neste processo, diferente dos modelos anteriores que estagnavam nas primeiras épocas, demonstrando a falta de aptidão das redes preliminares de aprender efetivamente.

3.7 Terceira geração de modelos

Os próximos modelos desenvolvidos são caracterizados pelo uso de tipos diferentes de redes (comparado aos implementados nas gerações anteriores) e pelo uso da métrica F1 para comparação dos resultados. A métrica F1, por considerar tanto a precisão quanto o *recall*, é mais adequada para avaliar resultados em bases de dados altamente desbalanceadas. A Figura 4 demonstra uma visão geral dos modelos de terceira geração.

Uma das ideias principais desenvolvidas nesta etapa foi a de, ao invés de construir dois modelos convolucionais, utilizar os pesos da rede pré-treinada durante o treinamento de uma rede convolucional completa. A rede pré-treinada escolhida nesta etapa foi a InceptionResNetV3

Figura 4: Arquitetura dos modelos de terceira geração: ajuste fino de redes mais complexas, treinamento temporal integrado usando camadas ConvLSTM e rede TCN para características complementares.



(Szegedy et al., 2017), escolhida por conta do seu desempenhos nos testes de *benchmark* realizados (Bianco, Cadene, Celona, & Napoletano, 2018). Para o tratamento dos vídeos, foram escolhidas camadas ConvLSTM2D (Hu et al., 2020), que seguem uma lógica semelhante à rede Conv3D, com a diferença que na rede ConvLSTM2D a descoberta das características temporais é realizada usando a mesma estratégia de uma rede recorrente do tipo LSTM (com memória e esquecimento de informação), ao invés de por convoluções, como na camada Conv3D.

A rede de características complementares desta etapa foi desenvolvida substituindo o modelo de redes recorrentes por um modelo do tipo *Temporal Convolutional Network* (TCN) (Lea et al., 2017). As redes do tipo TCN são semelhantes às convolucionais unidimensionais. Este tipo de rede consegue realizar extração de características espaciais de baixo nível, como uma rede CNN convencional, enquanto descobre características de mais alto nível como dependências temporais, como uma RNN. Estas redes são causais, significando que a propagação das convoluções acontece somente com os elementos anteriores, permitindo o tratamento da temporalidade. Elas também são dilatadas, indicando que o tamanho da entrada é igual ao da saída, permitindo a concatenação de mais camadas.

Conforme observado na Figura 4, o modelo treinado nesta etapa utiliza como entrada do classificador a concatenação das características extraídas das redes TCN (características complementares) e Inception + ConvLSTM (vídeos), descritas anteriormente. Este modelo obteve $94,17\%$ de acurácia e $F1 = 0,5122$ quando treinado usando a base de dados DAiSEE. Quando

Tabela 3: Desempenho dos modelos de terceira geração utilizando diferentes conjuntos de dados para treinamento e teste.

Base de dados	**Acurácia**	**F1**
DAiSEE	0,9417	0,5122
PAT2Math	0,5972	0,5859
DaiSEE + PAT2Math	0,9573	0,6882

o mesmo modelo foi treinado usando a base de dados PAT2Math, $59,72\%$ de acurácia e $F1 = 0,5859$ foram obtidos, demonstrando uma perceptível diferença devido à base PAT2Math ser consideravelmente menos desbalanceada.

Ao realizar novamente o treinamento do mesmo modelo, mas usando um novo conjunto de treinamento que mescla DAiSEE + PAT2Math, obteve-se uma rede com maior capacidade de generalização: $95,73\%$ de acurácia e $F1 = 0,6882$ no DAiSEE e $68,52\%$ de acurácia e $F1 = 0,5850$ na PAT2Math. Este resultado demonstra o poder de generalização que se obtém ao aumentar os dados disponíveis para treinamento de uma rede. A Tabela 3 mostra a relação de desempenho entre os treinamentos dos mesmos modelos nas diferentes bases de dados.

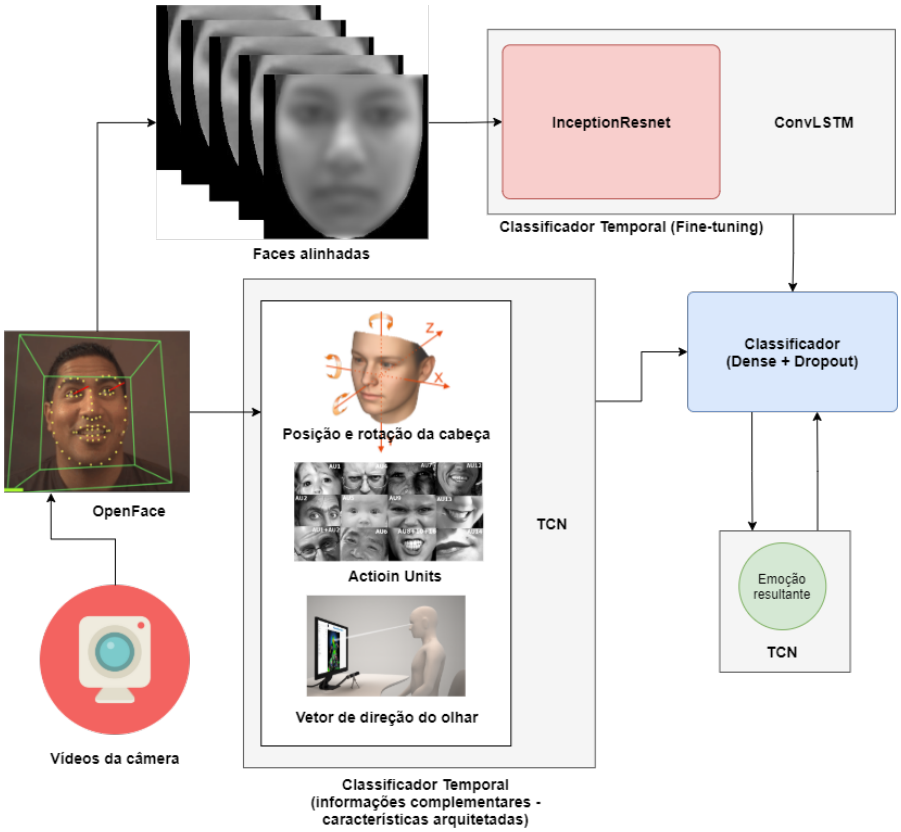
3.8 Quarta geração de modelos

Nesta etapa do desenvolvimento, muitos modelos já haviam sido desenvolvidos. A maioria, por conta da dificuldade de treinamento na base de dados altamente desbalanceada, não obteve sucesso em aprender os padrões necessários durante as épocas de treinamento. Dos resultados obtidos na etapa anterior, definiu-se que o modelo descrito na Seção 3.7 apresentou os melhores resultados vistos até então. Por conta disso, a nova etapa começou a ser desenvolvida, onde seriam integradas as informações relativas ao histórico das emoções. Este foi o grande diferencial dos modelos de quarta geração.

A ideia por trás da incorporação do histórico de emoções ao aprendizado vêm da intuição de que uma pessoa que já tenha experienciado uma determinada emoção num passado próximo, durante a exposição a um determinado conteúdo de aprendizado, tenderá a manifestar a mesma emoção em períodos subsequentes. Para verificar tal hipótese, foi desenvolvida uma rede que integra o histórico das emoções anteriores experienciadas pelo mesmo indivíduo em uma rede do tipo TCN. A saída desta rede foi concatenada às saídas das outras redes e integradas ao classificador, conforme mostra a Figura 5. Esta rede foi treinada na base de dados DAiSEE, e obteve um desempenho consideravelmente melhor que a rede anterior, atingindo $85,41\%$ de acurácia e $F1 = 0,6468$, enquanto a anterior obteve $F1 = 0,5122$, conforme visto na Seção 3.7. Este resultado demonstra a vantagem na utilização de informações históricas na detecção de emoções.

Cada subsequente geração de modelos trouxe uma nova melhoria comparativamente aos modelos de gerações anteriores. A Tabela 4 demonstra o desempenho obtida pelo melhor modelo em cada geração com a característica específica que o diferencia dos demais. A partir desta comparação, é possível observar a evolução dos modelos ao longo de seu desenvolvimento, que, com as figuras que os descrevem (Figura 2, Figura 3, Figura 4, e Figura 5), demonstram seu funcionamento, nível de complexidade e desempenho.

Figura 5: Arquitetura dos modelos de quarta geração: utilização de rótulos de predições anteriores para consideração do histórico de emoções.



3.8.1 Modelos treinados para outras emoções

Diante dos resultados promissores encontrados durante o treinamento dos modelos de quarta geração para a emoção engajamento, modelos com a mesma arquitetura foram treinados para reconhecimento das emoções **confusão**, **frustração**, e **tédio**.

O modelo treinado para reconhecimento da confusão obteve $88,74\%$ de acurácia e $F1 = 0,6123$. O modelo que reconhece frustração obteve acurácia $95,46\%$ e $F1 = 0,4992$. Já o modelo treinado para reconhecer tédio obteve 78.07% de acurácia e $F1 = 0,5428$. Um dos motivos de alguns modelos terem obtido resultados melhores para o reconhecimento de algumas destas emoções, que para outras, se deve pelas diferenças de desbalanceamento das classes presente nos subconjuntos de cada uma das emoções. A Tabela 5 demonstra alguns dados obtidos dos modelos treinados para cada emoção. A *acc. média* representa a média obtida entre as acurácias de cada uma das classes, por exemplo, engajamento positivo ou engajamento negativo. *Proporção* representa o nível de desbalanceamento presente nos exemplos, ou seja, a proporção de exemplos da classe negativa com relação aos exemplos da classe positiva. Por exemplo, $1:5$ significa 5 exemplos da classe positiva para cada exemplo da classe negativa presente nos exemplos de treinamento.

Tabela 4: Desempenho dos melhores modelos obtidos em cada geração para a emoção engajamento.

	Acurácia	**F1**	**Descrição**
Geração 1	0,7492	-	Modelo Conv3D sequencial
Geração 2	0,6733	-	BLSTM (carac. compl.) ǀ Time Distributed Conv2D com skip connections. Fusão BLSTM. Primeiro modelo com fusão por concatenação, e aprendizado observável
Geração 3	0,9417	0,5122	TCN (carc. compl.) ǀ Fine-tune Inception + ConvLSTM. Fusão TCN.
Geração 4	0,8541	0,6468	Histórico de emoções anteriores concatenadas na fusão em uma rede TCN.

Tabela 5: Melhores modelos para cada emoção: desempenho e desbalanceamento.

	Acurácia	**Acc. média**	**F1**	**Proporção**
Engajamento	0,8541	0,6741	0,6468	1:9,69
Confusão	0,8874	0,5845	0,6123	7,97:1
Frustração	0,9546	0,5047	0,4991	20,32:1
Tédio	0,7807	0,5424	0,5428	3,01:1

4 Discussão

O presente trabalho realizou o desenvolvimento de modelos computacionais para detecção e reconhecimento de emoções manifestadas durante situações de aprendizagem, sendo elas engajamento, confusão, frustração e tédio. Além de considerar informações visuais, tais como pose da cabeça, esse trabalho também usa no treinamento das redes neurais classificadoras a temporalidade das emoções dos estudantes para melhorar a acurácia da detecção das emoções de aprendizagem. Para a realização da classificação das emoções não básicas presentes em situação de aprendizagem, foram utilizadas bases de dados contendo vídeos da face de alunos assistindo a conteúdos educacionais ou resolvendo problemas de aprendizagem.

O **primeiro objetivo** específico do trabalho era obter precisão do reconhecimento de emoções acadêmicas por algoritmos de aprendizagem profunda melhor que o estado da arte para o reconhecimento de emoções por face para estas emoções. O modelo de aprendizado profundo construído se utiliza de informações temporais obtidas através da utilização de algoritmos para extração de características arquitetadas relacionadas a expressões faciais de alunos, bem como redes neurais convolucionais para a extração de características espaciais. O modelo temporal também se utiliza de informações sobre a sequência de emoções demonstradas nos exemplos para aumentar a eficácia da predição, tendo uma nova camada de abstração sobre as informações que dizem respeito à temporalidade das emoções presenciadas. Os modelos desenvolvidos obtiveram os seguintes resultados de acurácia $85,41\%$; $88,74\%$; $95,46\%$, e $78,07\%$ e valores de F1 $0,6468$, $0,6123$, $0,4991$, e $0,5428$ para as emoções de engajamento, confusão, frustração e tédio, respectivamente. Esses resultados são melhores que o estado da arte para emoções de aprendizagem, conforme apresentados na Seção 2.

O **segundo objetivo** específico do trabalho era verificar a influência da temporalidade das emoções (sequencia que as emoções foram sentidas pelos estudantes) na precisão do reconhe-

cimento de emoções acadêmicas por algoritmos de aprendizagem profunda. Os resultados dos experimentos também sugerem um significativo ganho de desempenho, de cerca de $26,27\%$ (de $0,5122$ para $0,6468$ F1, conforme Tabela 4) na métrica F1 para o engajamento usando as bases de dados DAiSEE + PAT2Math, e ao se considerar a sequência de emoções que um mesmo indivíduo presenciou anteriormente para inferir emoções manifestadas no presente momento. Dessa forma, os resultados do trabalho confirmam a hipótese de que considerar a temporalidade das emoções, fornecendo a sequência que as emoções são experimentadas pelo aluno e usando rede neurais temporais, melhora a acurácia dos reconhecedores de emoções de aprendizagem e que essa melhora é significativa.

5 Conclusão e trabalhos futuros

Entende-se que ambientes inteligentes de aprendizagem estão cada vez mais presentes no cotidiano tanto das salas de aula quanto do ensino à distância, portanto, o interesse na utilização das emoções do aluno a favor de seu processo de aprendizagem é crescente. Tendo isto em vista, o presente trabalho demonstrou a possibilidade da melhoria dos algoritmos de detecção das emoções não básicas presentes em situações de aprendizagem, as quais são naturalmente expressas por indivíduos de maneira muito mais sutil que emoções básicas, e dessa forma, requerem abordagens específicas para sua detecção. No presente trabalho, a abordagem utilizada para este tipo de detecção foi a consideração da temporalidade da sequência das emoções do aluno. Essa escolha se justifica pelo fato das emoções de aprendizagem experimentadas por um estudante depende das emoções manifestadas anteriormente (D'Mello et al., 2014; Reis, Alvares, Jaques, & Isotani, 2018). Consequentemente, demonstrou-se que a utilização da emoção obtida como saída nos modelos desenvolvidos podem ser utilizadas como entrada em etapas subsequentes do treinamento, visando aprimorar o desempenho do modelo de detecção.

Embora os resultados do presente trabalho tenham avançado significativamente o estado da arte de classificação de emoções de aprendizagem, eles ainda poderiam ser melhorados com bases de dados com maior quantidade de amostras, principalmente nas classes menos representadas, uma vez que o desempenho de algoritmos de aprendizagem profunda depende fortemente do tamanho da base de treinamento. No trabalho realizado, as limitações do tamanho reduzido da base foram contornados usando *transfer learning*, ou seja, foi empregada uma rede pré-treinada que utilizou os pesos do treinamento prévio como ponto de partida para o treinamento do modelo do presente trabalho. A limitação do desbalanceamento das bases foi mitigada utilizando pesos de treinamento e implementando o histórico de emoções. Dessa forma, como trabalho futuro, espera-se ampliar a coleta de dados para a base PAT2Math a fim de aprimorar o treinamento dos modelos desenvolvidos, objetivando a obtenção de dados que mantenham a base balanceada e o aprimoramento da capacidade de generalização dos modelos desenvolvidos através do uso de faces culturalmente mais diversas, atendendo melhor às demandas locais.

Outro ponto de melhoria é a pesquisa e o desenvolvimento de modelos mais especializados para as outras três emoções: confusão, frustração e tédio. O processo de pesquisa de um modelo que melhor se adaptasse à emoção de engajamento foi longo e demandou inúmeras experimentações. Embora façam parte da mesma família de emoções presentes no aprendizado, a confusão, frustração e tédio possuem suas próprias particularidades. Por conta disso, talvez hajam melhorias

a serem obtidas ao se criar modelos específicos e especializados na classificação destas emoções.

Finalmente, as duas únicas bases de dados públicas existentes que tratam de emoções de aprendizagem (EmotiW e DAiSEE) enfrentam alguns problemas, como o nível extremo de desbalanceamento. A única base que reporta as emoções confusão, frustração e tédio é a DAiSEE. A EmotiW, por sua vez, além de apresentar rótulos somente para a emoção engajamento, apresenta somente um único valor para cada vídeo, tornando o treinamento extremamente difícil. A base PAT2Math apesar de não ser pública e possuir bem menos dados que as outras duas, apresenta anotações das quatro emoções, além de ser consideravelmente menos desbalanceada.

Agradecimentos

O presente trabalho foi realizado com apoio da FAPERGS (Processo 17/2551-0001203-8) e do CNPq (processo 306005/2020-4).

Referências

Ackermann, P., Kohlschein, C., Bitsch, J. A., Wehrle, K., & Jeschke, S. (2016). Eeg-based automatic emotion recognition: Feature extraction, selection and classification methods. In *Int. conf. on e-health networking, applications and services* (pp. 1–6). [GS Search]

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0. In *Ieee int. conf. on automatic face & gesture recognition* (pp. 59–66). [GS Search]

Bianco, S., Cadene, R., Celona, L., & Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. *IEEE access*, *6*, 64270–64277. [GS Search]

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. [GS Search]

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. of Artif. Intel. Research*, *16*, 321–357. [GS Search]

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE CVPR* (pp. 248–255). [GS Search]

Dewan, M. A. A., Lin, F., Wen, D., Murshed, M., & Uddin, Z. (2018). A deep learning approach to detecting engagement of online learners. In *IEEE SmartWorld* (pp. 1895–1902). [GS Search]

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011a). Acted facial expressions in the wild database. *Australian National University, Technical Report TR-CS-11*, *2*, 1. [GS Search]

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011b). Static facial expression analysis in tough conditions. In *IEEE ICCV Workshops* (pp. 2106–2112). [GS Search]

D'Mello, S., & Calvo, R. A. (2013). Beyond the basic emotions: what should affective computing compute? In *CHI'13 Extended Abstracts* (pp. 2287–2294). [GS Search]

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. [GS Search]

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170. [GS Search]

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200. [GS

Search]
Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, 45–60. [GS Search]
Elrahman, S. M. A., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, *1*(2013), 332–340. [GS Search]
Fredrickson, B. L. (1998). What good are positive emotions? *Review of general psychology*, *2*(3), 300–319. [GS Search]
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press. [GS Search]
Goodfellow, I., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... Lee, D.-H. (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117–124). [GS Search]
Goodman, L. A. (1961). Snowball sampling. *The annals of mathematical statistics*, 148–170. [GS Search]
Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*. [GS Search]
Gupta, A., Jaiswal, R., Adhikari, S., & Balasubramanian, V. N. (2016). Daisee: Dataset for affective states in e-learning environments. *arXiv*, 1–22. [GS Search]
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE CVPR* (pp. 770–778). [GS Search]
Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, *44*(2), 174–199. [GS Search]
Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554. [GS Search]
Hu, W.-S., Li, H.-C., Pan, L., Li, W., Tao, R., & Du, Q. (2020). Spatial–spectral feature extraction via deep convlstm neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(6), 4237–4250. [GS Search]
Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *IEEE CVPR* (pp. 4700–4708). [GS Search]
Huang, W., Song, G., Li, M., Hu, W., & Xie, K. (2013). Adaptive weight optimization for classification of imbalanced data. In *Int. Conf. on Intelligent Science and Big Data Engineering* (pp. 546–553). [GS Search]
Jaques, P. A., Seffrin, H., Rubi, G. L., de Morais, F., Ghilardi, C., Bittencourt, I. I., & Isotani, S. (2013). Rule-based expert systems to support step-by-step guidance in algebraic problem solving: The case of the tutor pat2math. *Expert Systems with Applications*, *40*(14), 5456–5465. doi: http://dx.doi.org/10.1016/j.eswa.2013.04.004 [GS Search]
Kaggle (2013). Challenges in representation learning: Facial expression recognition challenge. Retrieved from https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/overview [Acessado em 17 de Julho de 2019]
Kaur, A., Mustafa, A., Mehta, L., & Dhall, A. (2018). Prediction and localization of student engagement in the wild. In *2018 digital image computing: Techniques and applications (dicta)* (pp. 1–8). [GS Search]
King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*(Jul), 1755–1758. [GS Search]
Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. [GS Search]
Kreifelts, B., Wildgruber, D., & Ethofer, T. (2013). Audiovisual integration of emotional informa-

tion from voice and face. In *Integrating face and voice in person perception* (pp. 225–251). Springer. [GS Search]

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., ... Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational research review*, *10*, 90–115. [GS Search]

Lazarus, R. S. (1982). Thoughts on the relations between emotion and cognition. *American psychologist*, *37*(9), 1019. [GS Search]

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In *IEEE CVPR* (pp. 156–165). [GS Search]

Li, S., & Deng, W. (2018). Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*. [GS Search]

Liu, C., Tang, T., Lv, K., & Wang, M. (2018). Multi-feature based emotion recognition for video clips. In *Int. Conf. on Multimodal Interaction* (pp. 630–634). [GS Search]

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*. [GS Search]

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE CVPR-Workshops* (pp. 94–101). [GS Search]

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*. [GS Search]

Morais, F., & Jaques, P. A. (2022). Dinâmica de afetos em um sistema tutor inteligente de matemática no contexto brasileiro: uma análise da transição de emoções acadêmicas. *Revista Brasileira de Informática na Educação*, *30*, 519-541. Retrieved from https://doi.org/10.5753/rbie.2022.2577 doi: 10.5753/rbie.2022.2577 [GS Search]

Morais, F., Kautzmann, T. R., Bittencourt, I. I., & Jaques, P. (2019). Emap-ml: A protocol of emotions and behaviors annotation for machine learning labels. In *EC-TEL*. [GS Search]

Nardelli, M., Valenza, G., Greco, A., Lanata, A., & Scilingo, E. P. (2015). Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Transactions on Affective Computing*, *6*(4), 385–394. [GS Search]

Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2018). Automatic recognition of student engagement using deep learning and facial expression. *arXiv preprint arXiv:1808.02324*. [GS Search]

Ocumpaugh, J. (2015). Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences*, *60*. [GS Search]

Pekrun, R. (2011). Emotions as drivers of learning and cognitive development. In *New perspectives on affect and learning technologies* (pp. 23–39). Springer. [GS Search]

Reis, H., Alvares, D., Jaques, P., & Isotani, S. (2018). Analysis of permanence time in emotional states: A case study using educational software. In *ITS* (pp. 180–190). [GS Search]

Reis, H., Jaques, P., & Isotani, S. (2018, 03). Sistemas tutores inteligentes que detectam as emoções dos estudantes: um mapeamento sistemático. *Revista Brasileira de Informática na Educação*, *26*, 76-107. doi: 10.5753/rbie.2018.26.03.76 [GS Search]

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2014). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *37*(6), 1113–1133. [GS Search]

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. [GS Search]

Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., & Sebe, N. (2016). Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, *9*(2), 147–160. [GS Search]

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conf. on Artif. Intellig.* [GS Search]

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *IEEE CVPR* (pp. 2818–2826). [GS Search]

Thomas, C., Nair, N., & Jayagopi, D. B. (2018). Predicting engagement intensity in the wild using temporal convolutional network. In *Int. Conf. on Multimodal Interaction* (pp. 604–610). [GS Search]

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *IEEE CV* (pp. 4489–4497). [GS Search]

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets. In *IJCNN* (pp. 4368–4374). [GS Search]

Werlang, P. (2022). Github: werlang/emolearn-ml-model. Retrieved from https://github.com/werlang/emolearn-ml-model [Acessado em 3 de Outubro de 2019]

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98. [GS Search]

Yang, J., Wang, K., Peng, X., & Qiao, Y. (2018). Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In *Int. Conf. on Multimodal Interaction* (pp. 594–598). [GS Search]