UNISINOS

Programa de Pós-Graduação em

Computação Aplicada

Doutorado Acadêmico

Pablo Santos Werlang

RECONHECIMENTO DE EMOÇÕES ACADÊMICAS POR FACE ATRAVÉS DE APRENDIZAGEM PROFUNDA

Considerando a sequência de emoções e a personalidade do estudante

São Leopoldo, 2022

UNIVERSIDADE DO VALE DO RIO DOS SINOS — UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
NÍVEL DOUTORADO

PABLO SANTOS WERLANG

RECONHECIMENTO DE EMOÇÕES ACADÊMICAS POR FACE ATRAVÉS DE APRENDIZAGEM PROFUNDA

Considerando a sequência de emoções e a personalidade do estudante

SÃO LEOPOLDO
2022

Pablo Santos Werlang

RECONHECIMENTO DE EMOÇÕES ACADÊMICAS POR FACE ATRAVÉS DE APRENDIZAGEM PROFUNDA

Considerando a sequência de emoções e a personalidade do estudante

Tese apresentada como requisito parcial para a obtenção do título de Doutor pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Orientador:
Prof. Dra. Patrícia A. Jaques Maillard

São Leopoldo
2022

W489r Werlang, Pablo Santos.
Reconhecimento de emoções acadêmicas por face através de aprendizagem profunda: considerando a sequência de emoções e a personalidade do estudante / Pablo Santos Werlang – 2022.
108 f. : il. color. ; 30 cm.

Tese (doutorado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2022.
"Orientador: Prof. Dra. Patrícia A. Jaques Maillard."

1. Redes neurais (Computação). 2. Reconhecimento de emoções. 3. Emoções no aprendizado. 4. Reconhecimento multimodal. 5. Computação afetiva. I. Título.

CDU 004.8

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Bruna Sant'Anna – CRB 10/2360)

AGRADECIMENTOS

Agradeço a todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho. Em especial à minha orientadora, Patrícia, pela dedicação e trabalho exemplar desempenhado durante todo o processo que envolveu minha pesquisa. Sua atuação sem dúvida foi primordial para o sucesso deste trabalho.

Gostaria também de agradecer ao Instituto Federal Sul-Riograndense pelo apoio que me foi proporcionado através do programa de incentivo à qualificação.

Por fim, agradeço também, aos meus familiares e amigos pela convivência e apoio nos momentos em que mais precisei.

RESUMO

A computação afetiva busca melhorar a interação homem-máquina, desenvolvendo ferramentas e técnicas para tornar os processos de decisão dos sistemas mais adaptados aos estados afetivos humanos. O reconhecimento automático de emoções através da face é uma área relativamente recente e que possui o potencial de tornar a interação com um sistema de computador uma experiência cada vez 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. As emoções **engajamento**, **confusão**, **frustração** e **tédio**, comumente presentes em contexto de aprendizagem, são a chave para manutenção do engajamento do aluno e, por consequência, o sucesso de seu aprendizado. O presente trabalho desenvolveu um modelo capaz de reconhecer através de 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 *Action Units*. Considerando o modelo psicológico proposto por D'Mello de interação entre as emoções de aprendizagem, que considera que existe um fluxo de interação entre as emoções que determina a ordem em que essas se manifestam, o trabalho possui como principal contribuição a consideração do fluxo das emoções, bem como características de personalidade para detecção mais precisa das emoções. Diversas configurações de modelos de aprendizado profundo de máquina foram testadas, e suas eficiências comparadas aos modelos mais recentemente desenvolvidos. Os resultados trazem evidências que considerar a sequência de emoções de aprendizagem e a personalidade 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, e na rede treinada na base PAT2Math o ganho de desempenho foi de $1,48\%$ na métrica F1 (de $0,8741$ para $0,8871$) quando também incluído os traços de personalidade do indivíduo. Quando comparado ao estado-da-arte, o modelo obteve um desempenho $5,6\%$ superior utilizando a métrica F1, porém a acurácia teve uma perda de $4,7\%$.

Palavras-chave: reconhecimento de emoções, redes neurais profundas, emoções no aprendizado, reconhecimento multimodal, computação afetiva.

ABSTRACT

Affective computing aims to improve human-machine interaction by developing tools and techniques to enable the system's decision-making processes to adjust to human affective states. Automatic face recognition of emotions is a relatively recent area that has the potential of turning human-computer interaction into 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 work created 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 used deep neural networks to classify one of these emotions, extracting statistical, temporal, and spatial features from the videos provided for training, including eye movement and Action Units. Considering the psychological model of affect dynamics proposed by D'Mello, which states that in learning situations, each emotion's experience is tied to each other, and their presence is determined by the order in which they are shown, this work's main contribution is to take into account the flow of emotions as well as the learner's personality traits as a mean for increasing emotion detection accuracy. We tested several model configurations and their efficiency compared to recently developed models. Results show that considering the learning emotions sequence and the personality 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, while we achieved 1.48% F1 improvement on the model trained using the PAT2Math dataset (from 0.8741 to 0.8871) when including subject's personality traits. Compared to the state-of-the-art, the model achieved a superior 5.6% using the F1 metric. However, its accuracy was 4.7% lower.

Keywords: emotion recognition, deep learning, neural networks, learning emotions, multimodal recognition, affective computing.

SUMÁRIO

1 INTRODUÇÃO **8**
1.1 Questão de Pesquisa, Hipóteses e Objetivos 11
1.2 Organização da tese 13

2 ESTADOS AFETIVOS **14**
2.1 Emoção 14
2.1.1 Teorias e modelos 14
2.2 Emoções na Aprendizagem 17
2.3 Personalidade 18
2.3.1 Teorias e modelos 18

3 AMBIENTES DE APRENDIZAGEM INTELIGENTES E EMOÇÕES **20**
3.1 Computação Afetiva 20
3.2 Sistemas Tutores 22
3.2.1 Sistemas Tutores Afetivos 24

4 DETECÇÃO AUTOMÁTICA DE EMOÇÕES POR FACE **27**
4.1 Redes Neurais Artificiais 30
4.1.1 Métricas de treinamento 31
4.1.2 Tipos comuns de redes neurais 33
4.2 Trabalhos e aplicações 38

5 TRABALHOS RELACIONADOS **42**
5.1 Método de revisão 42
5.2 Detecção automática de emoções básicas por face 44
5.3 Detecção automática de emoções acadêmicas por face 49
5.4 Pesquisa atualizada sobre detecção de emoções acadêmicas 54
5.5 Comparação dos trabalhos relacionados com o trabalho realizado . . . 56

6 MÉTODO **59**
6.1 Bases de dados 60
6.2 Pré-processamento 63
6.2.1 Tratamento dos vídeos 63
6.2.2 Balanceamento de Classes 65
6.2.3 Características complementares 68
6.3 Arquitetura Genérica do Modelo Desenvolvido 69
6.3.1 Rede de características complementares 70
6.3.2 Rede de características espaciais 70
6.3.3 Fusão de modelos 73
6.4 Treinamento 73

7 RESULTADOS **75**
7.1 Primeira geração de modelos 75
7.2 Segunda geração de modelos 78
7.3 Terceira geração de modelos 82
7.4 Quarta geração de modelos 84
7.4.1 Modelos usando personalidade 87
7.4.2 Modelos treinados para outras emoções 88

7.4.3 Comparativo com o estado-da-arte 89

8 CONCLUSÃO E TRABALHOS FUTUROS 93
8.1 Limitações 95

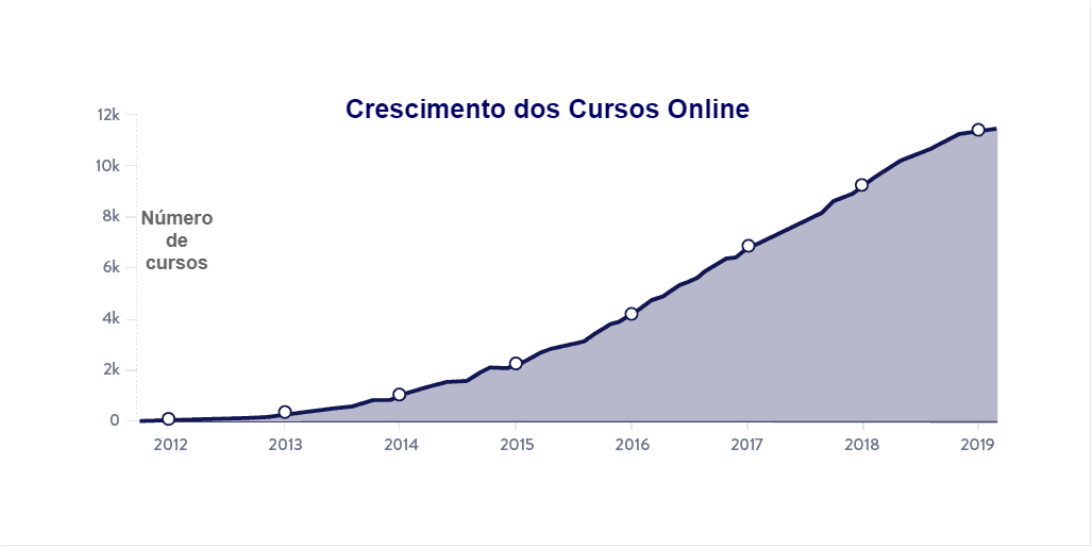
REFERÊNCIAS 97

1 INTRODUÇÃO

A maneira como as pessoas se relacionam com computadores está mudando. Antes vistos somente como uma ferramenta para execução de tarefas, hoje em dia tem cada vez mais ganhado espaço na vida pessoal da população (BROOKS, 2015). Antes o computador recebia ordens por comandos, hoje ele dá sugestões de compras, decide o melhor caminho e até mesmo faz o carro parar.

De todos os aspectos, um que tem ganhado bastante destaque nos últimos tempos é o aumento na disponibilidade e qualidade das plataformas que auxiliam o processo de aprendizagem. Estas plataformas vêm mudando a relação das pessoas com a educação, e têm encontrado crescente aumento de popularidade, seja no auxílio ao ensino tradicional ou aos cursos à distância, conforme mostra a Figura 1. Com essa revolução no ensino, os sistemas tutores inteligentes ganharam cada vez mais espaço, e hoje existem ótimas plataformas de ensino que permitem o aprendizado autônomo e assistido por parte do aluno, como, por exemplo, Codecademy[1], Khan Academy[2], dentre outros. Com essa crescente demanda, é natural que cada vez mais existam pesquisas a respeito de formas de melhorar esta relação de tutores inteligentes com alunos.

Figura 1: Crescimento do número de cursos online ao longo dos anos.



Fonte: Adaptado de `https://www.classcentral.com/report/mooc-stats-2018/`.

No passado, as emoções tiveram seu papel relegado ao contraponto do raciocínio humano, porém hoje 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

[1] https://www.codecademy.com/
[2] https://pt.khanacademy.org/

necessitam ter ciência das emoções que seus usuários experienciam, bem como saber como agir para instigar as emoções mais úteis no contexto da aprendizagem. A computação afetiva visa atender a essa necessidade inerente da interação humano-máquina (PICARD, 2000), visto que a compreensão dos estados afetivos humanos e o conhecimento sobre suas implicações são importantes para a melhoria nos processos de aprendizagem (PEKRUN; ELLIOT; MAIER, 2009; PEKRUN, 2011, 2014; FREDRICKSON, 1998; D'MELLO; PICARD; GRAESSER, 2007).

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. Muitos estudos relatam o reconhecimento automático de emoções através de diversas modalidades como, por exemplo, expressões faciais e movimento dos olhos (YANG et al., 2018), condutividade da pele (SUBRAMANIAN et al., 2016), batimentos cardíacos (NARDELLI et al., 2015), ondas cerebrais (ACKERMANN et al., 2016), áudio da voz (LIU et al., 2018), *logs* de uso do sistema (MORAIS et al., 2019), dentre outros.

Dentre todas as modalidades de detecção automática de emoções[3], uma das que mais se assemelham 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 (HUTMACHER, 2019). 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. A pesquisa em inteligência artificial cresceu em popularidade nos últimos tempos, em grande parte graças à evolução das redes neurais. E hoje estas redes conseguem aprender padrões complexos graças ao (1) aumento da capacidade computacional, que permitiu que redes mais complexas e com mais camadas sejam criadas, e (2) o aumento da quantidade de bases de dados para treinamento, pois as redes neurais necessitam de muita informação para aprender os padrões (MARCUS, 2018). Estes dois fatores, somado a

[3]Nessa tese, a expressão "detecção automática de emoções" se refere à detecção de emoções realizada por algum artefato computacional sem necessitar intervenção humana.

sucessos na resolução de problemas, como do gradiente de desaparecimento (Seção 4.1.2), 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, possibilitou a resolução de uma série de problemas para os quais não haviam sido encontradas soluções.

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 e 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).

Em contextos de aprendizagem, outra informação que deve ser considerada é a personalidade do estudante. Estudos realizados por Reis et al. (2018) demonstram que estudantes com característica de personalidade neuroticista (baixa estabilidade emocional), em geral, possuem menor tolerância à confusão, fazendo com que as emoções tédio e frustração se manifestem mais rapidamente. O oposto foi observado por estudantes com alta pontuação no traço de extroversão. Além disso, D'Mello e Graesser (2012) descreve um modelo que indica não só uma correlação entre as emoções engajamento, confusão, tédio e frustração e as situações de aprendizagem, mas também um provável fluxo de transição entre elas, indicando que, quando o estudante se encontra em um estado emocional específico, existe uma certa probabilidade de ele passar a outros específicos.

O presente trabalho realiza 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 (ver Seção 4.2), esses trabalhos são voltados ao reconhecimento das emoções básicas, incomuns em ambientes de aprendizagem. Conforme demonstrado por D'Mello e 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[4]. Tendo isso em vista, há a necessidade de uma abordagem diferenciada para realização da

[4]"Sutil"se refere a expressão facial ou corporal realizada de forma mais discreta, dificultando a identificação por outros agentes. Por exemplo, há menos amplitude da movimentação da boca e dos olhos.

mesma.

O presente trabalho aborda esta questão através da exploração da relação temporal entre as emoções, assim como por considerar características pessoais dos estudantes, como sua personalidade. De acordo com D'Mello et al. (2014), durante o aprendizado, um aluno transita entre as emoções de engajamento, confusão, frustração e tédio em um fluxo previsível, conforme descrito na Seção 2.2. Por exemplo, quando o aluno encontra-se confuso, ele tende a experienciar engajamento quando consegue resolver a causa dessa confusão; caso contrário, ele passa a sentir frustração e tédio, emoções negativamente correlacionadas à aprendizagem. Dessa forma, uma hipótese de pesquisa da tese é que considerar o histórico de emoções, na ordem em que elas são expressas pelo estudante, no treinamento de modelos classificadores pode melhorar a precisão da detecção de emoções futuras.

Além disso, a personalidade do estudante impacta igualmente na frequência das emoções experienciadas por ele e na duração dessas emoções. Conforme demonstrado nas pesquisas de (REIS et al., 2018; GOLDONI; REIS; JAQUES, 2022), alunos com alta pontuação em neuroticismo tendem a sustentar confusão por uma menor duração em situações complexas de aprendizagem, transitando rapidamente para emoções negativas mais prejudiciais à aprendizagem. Ao incluir estas características de personalidade no treinamento, é esperada uma melhora da precisão das previsões.

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. 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.

1.1 Questão de Pesquisa, Hipóteses e Objetivos

A **questão de pesquisa** que essa tese visa responder é: "*Como melhorar o reconhecimento facial de emoções acadêmicas (engajamento, confusão, frustração e tédio) expressas por estudantes através de vídeos capturados por uma webcam em ambientes digitais de aprendizagem?*"

Para responder a essa questão de pesquisa, as seguintes **hipóteses de pesquisa** foram definidas:

- H_1 : *Algoritmos e técnicas de aprendizagem profunda considerados o estado da arte em reconhecimento facial devem ser empregados no desenvolvimento dos classificadores de detecção emocional.* Para o desenvolvimento de um modelo classificador de emoções de

aprendizagem que apresente um desempenho compatível com o estado da arte, devem ser usados algoritmos considerados também o estado da arte na área, ou seja, que apresentem os melhores resultados para tarefas de reconhecimento facial. O benefício trazido pela utilização de dados extras úteis de personalidade e sequência de emoções (ver hipóteses H_2 e H_3) pode ser minimizado se forem empregados algoritmos e técnicas que não apresentem bom desempenho para o problema de reconhecimento facial de emoções.

- H_2 : *Rótulos de emoções obtidos sequencialmente devem ser usados no treinamento dos algoritmos de rede profunda*. O trabalho de D'Mello e Graesser (2012); D'Mello et al. (2014) mostrou que existe uma probabilidade de transição entre as emoções de aprendizagem, 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 informações vivenciadas pelo aprendiz.

- H_3 : *Dados de características de personalidade do aluno devem ser igualmente usados no treinamento dos algoritmos de aprendizagem profunda para melhorar a acurácia da detecção*. Como explicado anteriormente, a personalidade do estudante impacta igualmente na frequência das emoções experienciadas por ele e na duração dessas emoções (REIS et al., 2018; GOLDONI; REIS; JAQUES, 2022). Alunos com alta pontuação em neuroticismo tendem a sustentar confusão por uma menor duração em situações complexas de aprendizagem, transitando rapidamente para emoções negativas mais prejudiciais à aprendizagem. Consequentemente, os algoritmos de aprendizagem profunda, que trabalhem com dados temporais e que recebessem em seu treinamento rótulo de emoções dos estudantes na sequência em que elas foram vivenciadas, poderiam aprender padrões de associação entre a personalidade e a probabilidade do estudante sentir novamente uma emoção (estendendo ou encurtando a sua duração).

O trabalho proposto possui como **objetivo principal** 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 traços de personalidade e a temporalidade das emoções dos estudantes para melhorar a acurácia da detecção das emoções de aprendizagem (de acordo com as métricas F1 e acurácia), ou seja, melhorar o percentual de previsões corretas feitas pelo modelo em relação ao total de previsões (acurácia) e melhorar a média harmônica entre as métricas *precision* e *recall* do modelo (F1), o que é especialmente útil em problemas de classificação com classes desbalanceadas.

Em busca da realização deste objetivo central, podem-se elencar os seguintes **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.
- Verificar a influência da inclusão da informação de traço da personalidade do aluno na acurácia da detecção das emoções de aprendizagem quando essa informação é incluída no treinamento dos algoritmos de aprendizagem profunda de reconhecimento facial de emoções.

1.2 Organização da tese

O presente documento descreve o trabalho realizado. Sua organização está disposta da seguinte maneira. O capítulo 2 versa sobre estados afetivos, emoções e suas teorias, descrevendo igualmente sua relação com a aprendizagem. O capítulo 3 apresenta o histórico da área de computação afetiva e sua ligação com a detecção de emoções por diferentes modalidades, bem como sobre os sistemas tutores inteligentes e o uso de informações afetivas para benefício da aprendizagem. O capítulo 4 aborda o tema de detecção de emoções através da face, bem como os principais conceitos de redes neurais artificiais empregados neste tipo de detecção. O capítulo 5 apresenta diversos trabalhos relacionados ao projeto proposto, comparando suas características principais a fim de destacar a originalidade do trabalho proposto. O capítulo 6 detalha o presente trabalho, explicando desde seus conceitos gerais, as decisões de projeto, e características gerais das arquiteturas utilizadas. O capítulo 7 relata as especificidades de cada geração de modelos desenvolvidos, dando detalhes sobre sua implementação, bem como descreve os resultados obtidos e comparativos dos principais modelos que compõem cada geração. Por fim, o capítulo 8 apresenta as limitações do trabalho desenvolvido, conclusões e resultados esperados do trabalho.

2 ESTADOS AFETIVOS

Estados afetivos são mecanismos que informam sobre a natureza de eventos que o indivíduo experiencia (SCHWARZ, 1990). Há diversos estados afetivos, e Scherer et al. (2000) classificou-os em cinco tipos: emoção, humor, postura interpessoal, atitude e característica de personalidade. Cada estado afetivo pode ser identificado e classificado por características como intensidade, duração, impacto comportamental, dentre outros. Por exemplo, enquanto o humor possui como característica níveis de intensidade e duração medianos, característica de personalidade (outro estado afetivo) é reconhecido por possuir baixa intensidade e alta duração (SCHERER et al., 2000). A emoção, portanto, como um dos estados afetivos identificados, possui como característica uma curta duração, alta intensidade e é disparada por uma avaliação de um evento ou uma ação de pessoa (SCHERER et al., 2000). Neste capítulo serão apresentados os aspectos gerais sobre as teorias a respeito dos estados afetivos, focando nos estados afetivos emoções e personalidade, que serão empregados na tese.

2.1 Emoção

As emoções são um dos diversos estados afetivos observados nas pessoas. Embora não haja um consenso entre suas definições (KLEINGINNA; KLEINGINNA, 1981; SCHERER, 2005), Scherer (2005) as descreve como "mudanças nos estados de todos os subsistemas de um organismo em resposta à avaliação de um estímulo interno ou externo que representa uma inquietação do organismo" (p. 697). Apesar de ser popularmente aceite que emoções prejudicam a capacidade dos indivíduos de tomar decisões racionais, trabalhos científicos mostraram que as emoções estão envolvidas na tomada de decisão e em outros processos cognitivos (SCHERER et al., 2000; DALGLEISH; POWER, 2000). As emoções, estando intimamente ligadas à cognição, surgem de um processo de avaliação (*appraisal*) de atividades de interação com o ambiente em comparação com os objetivos do indivíduo (LANE; NADEL, 2002).

Scherer (2005) descreve que a emoção é o produto de cinco subsistemas, os componentes da emoção. A Tabela 2.1 mostra cada um destes componentes e os relaciona com suas funções, demonstrando a natureza multimodal das emoções. Através da análise dos componentes, Scherer (2005) discrimina emoções de outros fenômenos afetivos como atitudes, humor, preferências e disposições.

2.1.1 Teorias e modelos

De acordo com Scherer et al. (2000), uma das primeiras teorias sobre emoção vem de Platão, onde ele sugere que a alma é dividida entre cognição, emoção e motivação. Esta ideia foi reforçada até meados dos séculos XVIII e XIX, mesmo tendo Aristóteles, cinquenta anos depois de Platão, falado sobre a interação entre os diferentes níveis do funcionamento psicológico. Já

Tabela 1: Componentes emocionais

Componente Emocional	**Função**
Componente cognitivo (appraisal)	Avaliação de objetos e eventos
Componente neurofisiológico (sintomas corporais)	Sistema de regulação
Componente motivacional (tendências de ação)	Preparação e direção da ação
Componente de expressão motora (expressão vocal e facial)	Comunicação da intenção de reação e comportamento
Componente de sentimento subjetivo (experiência emocional)	Monitoramento de estado interno e interação organismo-ambiente

Fonte: (SCHERER, 2005).

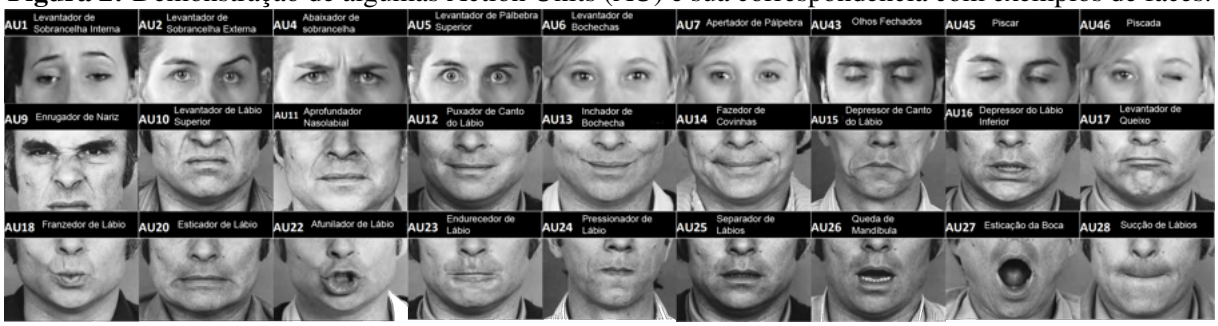
no século XVII, Descartes propôs que os processos psicológicos e fisiológicos são relacionados e sugere trabalhar com a mente e o corpo simultaneamente, mas foi Darwin e Prodger (1998) que propôs conceitos que tiveram influência decisiva no ramo da psicologia e que até hoje são defendidos e debatidos. Ele enfatizou que a expressão facial, corporal e da voz possui forte ligação com a emoção, e que a expressão facial é universal. De acordo com sua teoria, indivíduos de diferentes culturas demonstram as mesmas características indicativas de determinadas emoções com relação a outros indivíduos em contextos socioculturais diferentes. Suas observações serviram de base para a corrente atual da psicologia que defende a universalidade de um grande número de fenômenos emocionais.

Existem diversos modelos que buscam explicar e categorizar emoções e enquanto eles diferem no número de emoções que buscam explicar e no princípio para a diferenciação, eles podem ser classificados em quatro categorias: modelos dimensionais, modelos de emoções discretas, modelos orientados a significado e modelos componenciais. Modelos dimensionais buscam classificar emoções através de dimensões, onde os rótulos das emoções são definidos por estágios intermediários em escalas determinadas, por exemplo, valência da emoção ou excitação (BRADLEY; LANG, 1994). Modelos discretos baseiam-se na premissa que existe um conjunto de emoções básicas, que surgiram por mecanismos evolutivos, baseando-se no trabalho de Darwin e Prodger (1998), e destas emoções básicas derivam outras emoções através de uma mescla das principais (EKMAN; DAVIDSON, 1994). Os modelos orientados ao significado possuem em comum a crença que construtos sociais, como contexto cultural ou estrutura linguística, definem as emoções mais que contextos psicobiológicos. Modelos componenciais partem da premissa que as emoções são deduzidas a partir da avaliação cognitiva e são fruto das respostas dadas por diversos subsistemas (SCHERER et al., 2000).

Os trabalhos de Ekman (1997), utilizando a abordagem dos modelos discretos das emoções, foram responsáveis por disseminar a ideia que existem seis emoções básicas no ser humano, independente de características regionais ou culturais: alegria, tristeza, raiva, medo, surpresa e nojo. Baseado nas teorias das emoções básicas, Ekman (1992) desenvolveu o FACS (*Facial*

Action Coding System), um sistema que mapeia contrações musculares da face em uma codificação numérica, chamadas AUs (de *Action Units* em inglês). A combinação destas AUs possui o potencial de identificar todo tipo de expressão facial humana. Através destas, foi desenvolvido um trabalho que relaciona as AUs com cada uma das emoções básicas (Figura 2).

Figura 2: Demonstração de algumas Action Units (AU) e sua correspondência com exemplos de faces.



Fonte: Adaptado de (MARTINEZ et al., 2017)

A análise das expressões faciais pode ser embasada no julgamento de emoções que desencadeiam tais expressões ou em sinais relativos aos movimentos e deformações musculares ocorridos. Enquanto o primeiro método envolve a inferência da emoção através de conceitos culturais a respeito da expressão facial analisada, o segundo método utiliza as teorias de Ekman das FACS, expressando como resultado as AUs, não sendo suscetível à interpretação psicológica, pois trabalha com dados relativos a flexões musculares da face (FASEL; LUETTIN, 2003).

Outra importante teoria é o *appraisal*, que, de acordo com Scherer (1999), possui suas origens no trabalho de Arnold (1960), e se baseia na análise do componente cognitivo das emoções. O processo de *appraisal* é um processo inconsciente de avaliação de diversos critérios pessoais do indivíduo, que disparam uma emoção. Um mesmo estímulo pode ocasionar emoções distintas segundo os objetivos e metas pessoais do indivíduo. Pesquisadores da abordagem cognitivista acreditam que para um indivíduo experienciar uma emoção, um evento ou objeto precisa ser avaliado (*appraisal*) e este precisa afetar o indivíduo de alguma maneira, baseada em seus objetivos, experiências pessoais ou oportunidade de ação (CALVO; D'MELLO, 2010). Embora existam diversas teorias de *appraisal* com suas próprias abordagens, podem-se classificar quatro principais tipos, descritos a seguir:

- Critério - Abordagem clássica. Sugere um número de características e avalia o significado dos eventos para os objetivos do indivíduo.
- Atribuição - Foca no objeto agente do evento que ocasiona a emoção.
- Temática - Busca elencar relações entre a emoção e padrões relacionados a objetivos do indivíduo.

- Significado - Se preocupam com as operações lógicas que determinam a rotulação de um estado de sentimento com uma palavra de emoção específica.

Mais recentemente, Inzlicht, Bartholow e Hirsh (2015) defenderam que a relação entre as emoções e os processos cognitivos é ainda mais interligada. Em seu trabalho, alegam que os diversos processos de controle cognitivo podem ser compreendidos da mesma maneira que os processos que desencadeiam as emoções. Em seus experimentos demonstram que os mesmos conflitos que geram a dissonância cognitiva geram respostas afetivas que desencadeiam as emoções, e concluem que as emoções são o resultado de múltiplos processos que possuem origem nos processos cognitivos.

2.2 Emoções na Aprendizagem

As emoções estão intimamente ligadas aos processos cognitivos, em especial quando se tratando de processos relacionados à aprendizagem. Apesar de sua tamanha importância, as pesquisas que tratam das emoções em ambientes de aprendizagem estão em estágio inicial e recém começam a ser percebidas com a devida importância (PEKRUN, 2011).

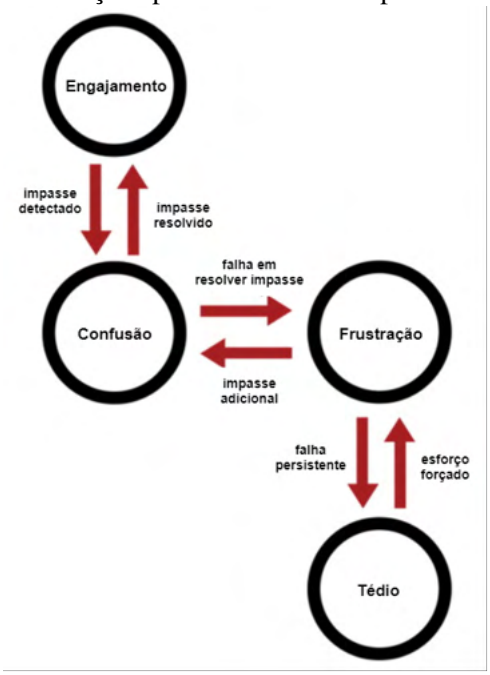
Conforme visto anteriormente, Ekman (1999) descreve sobre as emoções básicas e sua universalidade. No entanto, nem todas as emoções são experienciadas da mesma maneira em todos os ambientes. O trabalho de D'Mello e Calvo (2013) demonstra que em situações de aprendizagem, emoções não básicas como engajamento, confusão, frustração e tédio são encontradas com muito maior frequência que as emoções básicas. Essas emoções são chamadas emoções de aprendizagem ou acadêmicas. Por este motivo, trabalhos focando em detecção de emoções em ambientes de aprendizagem costumam se utilizar destas outras emoções para tentar inferir características que predizem o desempenho do aluno usuário destes sistemas.

Um estudo recente realizado por Bosch e D'Mello (2017) demonstra que as emoções acadêmicas, além de mais frequentes em ambientes de aprendizagem, também possuem correlação com o aprendizado do aluno. Dada a importância deste fato, é pertinente a análise do fluxo de uma emoção para outra como estratégia de melhoria nas taxas de sucesso dos alunos envolvidos (BOSCH; D'MELLO, 2017).

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 et al. (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) demonstram como as emoções positivas influenciam positivamente o repertório de ações e pensamento das pessoas. A Figura 3 demonstra esta dinâmica do fluxo das emoções durante o aprendizado e seus gatilhos.

Figura 3: Fluxo das emoções predominantes do aprendizado e seus gatilhos.



Fonte: Adaptado de (D'MELLO et al., 2014)

2.3 Personalidade

A personalidade dos indivíduos é uma característica que os torna únicos. Ela define seus gostos e modos de pensamentos para a maioria dos assuntos que tangem suas vidas. No entanto, definir o que é a personalidade humana não é uma tarefa fácil, e os pesquisadores da área não possuem um consenso sobre o assunto, existindo igualmente diversas teorias. Hall, Lindzey e Campbell (2000) e Nunes (2012) apresentam trabalhos transcorrendo sobre as teorias de personalidade.

2.3.1 Teorias e modelos

Dentre as inúmeras teorias sobre a personalidade, a teoria dos traços é a mais comumente utilizada quando se trata de modelagem computacional (JAQUES; NUNES, 2019). Segundo esta teoria, os indivíduos possuem diversos traços de personalidade em diferentes intensidades e estes traços formam a personalidade e são originários da experiência pessoal de cada indivíduo.

Derivada desta teoria, surgiu o modelo dos cinco grandes fatores (*Big Five*), que considera que todos os inúmeros traços de personalidade observados pelos teóricos da teoria dos traços

podem ser resumidos em somente cinco traços principais: abertura para experiências, estabilidade emocional, extroversão, amabilidade e consciência (GOLDBERG, 1990). A Tabela 2 relaciona os cinco grandes traços de personalidade com as principais características do indivíduo que apresenta o traço.

Tabela 2: Adjetivos característicos dos Cinco Grandes Fatores de Personalidade.

	Extroversão	**Amabilidade**	**Consciência**	**Estabilidade emocional**	**Abertura para experiências**
Polo do rótulo	Ativo	Altruísta	Confiável	Ansioso	Artístico
	Aventureiro	Amigável	Consciente	Apreensivo	Curioso
	Barulhento	Carinhoso	Eficiente	Emotivo	Engenhoso
	Energético	Confiante	Minucioso	Instável	Esperto
	Entusiástico	Cooperativo	Organizado	Nervoso	Imaginativo
	Exibido	Gentil	Prático	Preocupado	Inteligente
	Sociável	Sensível	Preciso	Temeroso	Original
	Tagarela	Simpático	Responsável	Tenso	Sofisticado
Polo oposto	Acanhado	Calmo	Antipático	Desatento	Comum
	Introvertido	Contido	Brigão	Descuidado	Simples
	Quieto	Estável	Bruto	Desorganizado	Superficial
	Reservado	Indiferente	Critico	Distraído	Tolo
	Silencioso	Sereno	Frio	Imprudente	Trivial
	Tímido	Tranquilo	Insensível	Irresponsável	Vulgar

Fonte: Adaptado de Nunes (2012).

Para ajudar na descoberta da personalidade de um indivíduo, foram criadas ferramentas, chamadas inventários, compostas de questionários e outros itens que podem ser aplicados diretamente aos indivíduos analisados através de perguntas, exames e análises. Porém, tais inventários, por exemplo, o *International Personality Item Pool* (GOLDBERG et al., 2006), possuem a grande desvantagem de serem muito intrusivos aos indivíduos submetidos à análise. Visando formas de realizar a prospecção da personalidade de indivíduos de forma que os usuários se sintam mais confortáveis, ferramentas computacionais são utilizadas para capturar dados específicos do usuário, como voz, expressões faciais, logs de uso do sistema, dentre outros. Tais informações são processadas por algoritmos de predição de modo a detectar automaticamente a personalidade do indivíduo, como no trabalho de Majumder et al. (2017).

3 AMBIENTES DE APRENDIZAGEM INTELIGENTES E EMOÇÕES

Ambientes de aprendizagem são sistemas criados para auxiliar o processo de ensino e aprendizagem das pessoas. Estes ambientes têm se tornado cada vez mais frequentes na vida cotidiana, seja pela popularização da internet ou pelo modelo prático de ensino que estas ferramentas proporcionam. Com a melhoria das capacidades de processamento dos computadores, das tecnologias de desenvolvimento de software e da evolução das técnicas de inteligência artificial, é natural que os ambientes de aprendizagem se beneficiem de tudo isso. Uma evolução natural para sistemas que abordam o aprendizado do usuário como atividade fim, é que sistemas inteligentes cumpram o papel de tutores que auxiliem o aprendizado do aluno enquanto interagindo com o sistema. Sendo que as emoções são parte importante do processo cognitivo humano, é importante que os sistemas tutores inteligentes consigam compreender e/ou expressar emoções. Este capítulo trata a respeito da computação afetiva, e quais suas implicações nestas relações do aluno com os sistemas tutores.

3.1 Computação Afetiva

Até o início dos anos 80 mantinha-se a teoria que as componentes da mente, emoção, cognição e motivação, eram independentes. Lazarus (1999) afirma em seu trabalho que tamanha distinção não existe naturalmente, mas que foi a partir do artigo de Zajonc (1980), em 1980 na revista *American Psychologist* que a comunidade da psicologia começou um debate que culminaria na teoria mais aceita hoje sobre o papel das emoções no processo cognitivo dos seres humanos. A partir disso, naturalmente o assunto começou a despertar um crescente interesse da comunidade científica na totalidade. Pioneira nesta área, Picard (2000) traz uma visão de que para tomar melhores decisões e interagir com humanos mais efetiva e naturalmente, sistemas computacionais precisam detectar, expressar ou sintetizar emoções. Destes estudos, nasce a computação afetiva, que se propõe a resolver problemas se utilizando de sistemas computacionais dotados de capacidade de detecção, expressão e/ou sintetização de emoções.

As pesquisas no ramo da computação afetiva têm se mostrado bastante promissoras e surgem a cada dia aplicações nas mais diversas áreas, como, por exemplo, realidade virtual, tutores de ensino, recomendações de produtos, etc. (TAO; TAN, 2005). A visão das emoções como importante parte do processo cognitivo humano tem se tornado cada vez mais expressiva, e por consequência também o uso das emoções como ferramenta computacional. O futuro da Inteligência Artificial, pelo menos quanto às aplicações que visam interação direta com pessoas ou processos relacionados aos desejos e necessidades humanas, está fortemente ligado às pesquisas em computação afetiva (CAMBRIA, 2016).

A detecção das emoções pode ser aplicada, por exemplo, em sistemas que se comunicam com o usuário, como *chatbots*. Em um sistema deste tipo, identificar a emoção o usuário está experienciando é importante para trilhar o rumo da conversação. Em sistemas tutores, a de-

tecção de emoções pode servir para descobrir se o aluno está engajado em uma atividade ou frustrado e, então, agir de acordo sugerindo-lhe novos conteúdos ou dicas. Tais sistemas eram tradicionalmente programados para realizar a detecção das emoções do usuário e classificá-las de acordo com as cinco emoções básicas, descritas na Seção 2. Com o tempo, os pesquisadores da área resolveram expandir seus trabalhos para a detecção de outras emoções não-básicas e até mesmo estados afetivos. Pode-se citar como um dos principais motivos para esta mudança de paradigma a percepção de que emoções básicas eram raramente notadas em determinados tipos de ambiente, como em um ambiente de aprendizagem, por exemplo, como demonstra D'Mello e Calvo (2013).

Na computação afetiva, existem diversas maneiras de um sistema realizar a detecção das emoções do usuário, seja pelo uso de algum hardware específico (sensores) ou pelos próprios padrões de uso do sistema. Enquanto interagindo com o sistema, o usuário fornece cliques do mouse, digitação de texto, logs de uso do sistema (cliques em links, tempo para execução de atividade, etc.) (ZIMMERMANN et al., 2003). Este tipo de interação fornece informações para que o sistema, por algoritmos específicos de extração de conhecimento e inferência, possam inferir as emoções do usuário. Caso o uso do sistema inclua a interação direta com um tutor virtual, a detecção de emoções poderá se dar através do reconhecimento do texto dado como entrada (STRAPPARAVA; MIHALCEA, 2008).

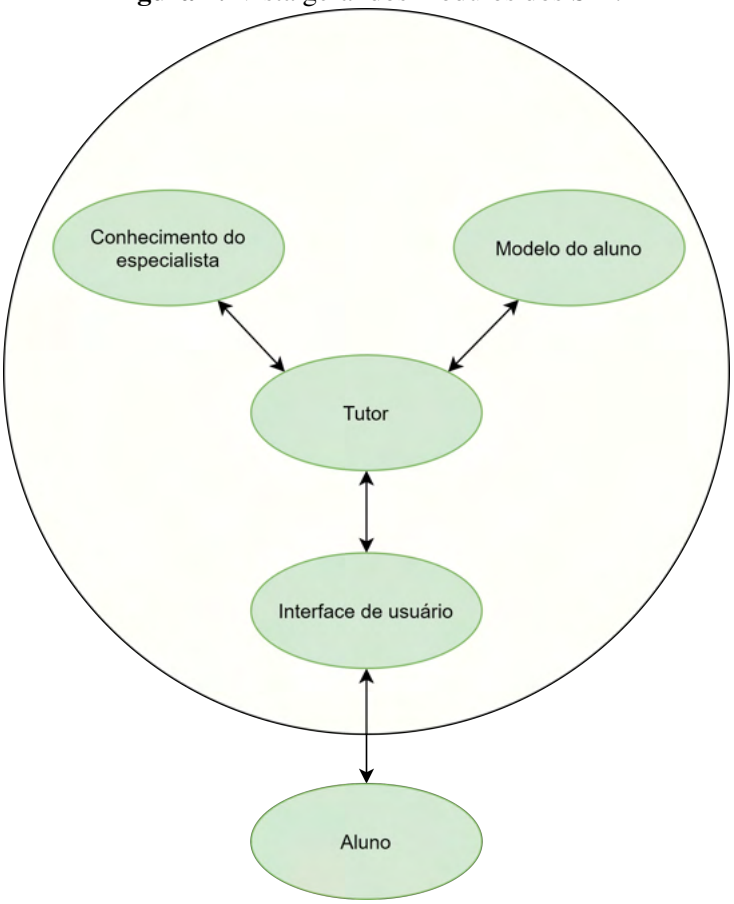
A detecção pode também se utilizar de hardware comumente disponível nos dispositivos, como, por exemplo, a detecção por voz, que utiliza o microfone ou a detecção de movimentos do corpo e expressão facial que utilizam câmeras. Dispositivos de hardware mais específicos também são utilizados para realizar estas detecções, tais como câmeras infravermelhas para captura do movimento dos olhos (*eyetracking*), e outros sensores fisiológicos, como detector de batimentos cardíacos ou ondas cerebrais (CALVO; D'MELLO, 2010).

De todos os métodos citados, um dos que atualmente recebe considerável atenção da comunidade de pesquisa em computação afetiva é o método de detecção de emoções através de expressões faciais. Isto se dá principalmente pela característica do ser humano de possuir expressões únicas para cada tipo de emoção vivenciada, pelos avanços na área de visão computacional e redes neurais, que permitem que padrões sejam encontrados em imagens de rostos de pessoas, o aumento da disponibilidade de bases de dados para treinamento de tais redes, a alta aceitação do uso do método por parte do usuário por não precisar instalar nenhum tipo de aparelho ao seu corpo, bem como da facilidade do acesso ao hardware necessário para a extração dos dados (pois qualquer notebook possui uma câmera integrada) (MOLLAHOSSEINI; CHAN; MAHOOR, 2016). A detecção das emoções pela face do aprendiz é o tema de interesse desse trabalho de pesquisa.

3.2 Sistemas Tutores

Um Sistema Tutor Inteligente (STI) é um agente de *software* que possui conhecimento sobre um domínio específico, consegue reconhecer a compreensão do usuário a respeito do mesmo domínio e utiliza algum tipo de interface para fornecer auxílio no processo de aprendizagem do usuário (NWANA, 1990). De maneira geral, STIs possuem em comum o emprego dos seguintes quatro módulos básicos: Módulo do conhecimento do especialista, Módulo do modelo do aluno, Módulo tutor e o Módulo de interface de usuário. A Figura 4 ilustra a arquitetura genérica de um STI composta por esses quatro módulos.

Figura 4: Vista geral dos módulos dos STI.



Fonte: Adaptado de (NWANA, 1990)

O **módulo especialista** é responsável por transformar o conhecimento sobre o domínio em estruturas que relacionam este conhecimento entre si. Além de possuir uma vasta gama de informações sobre o domínio, este módulo precisa possuir algum tipo de codificação que relacione estas informações com a maneira que pessoas trabalham com este conhecimento. Este tipo de trabalho possui relação com a maneira que sistemas especialistas são construídos, no sentido que o conhecimento de especialistas humanos é necessário para a construção de estruturas hierárquicas que dão sentido às informações e à maneira que pessoas aprendem sobre elas (ANDERSON, 1988).

O **modelo do aluno**, a partir das interações do usuário com a interface do STI, infere o conhecimento e habilidades que o usuário possui sobre o domínio, avaliando suas principais dificuldades e falhas (*misconceptions*) no conhecimento. O modelo do aluno é periodicamente consultado pelo sistema tutor de modo a buscar as habilidades que o usuário possui relacionadas aos conhecimentos relevantes ao domínio para saber que tipos de desafios ou dicas deve apresentar. Através da entrada do usuário, o sistema tutor atualiza o modelo do aluno refletindo o novo estado das habilidades (VANLEHN, 1987).

O **módulo tutor** é responsável por processar as informações relativas ao domínio e compará-las ao modelo do aluno, para prestar a assistência ao usuário. O tutor usa o conhecimento do modelo do aluno para determinar qual a tarefa é a mais recomendada ao usuário. Então, a intervenção é planejada segundo as especificações pedagógicas do sistema, podendo ser dicas sobre a alternativa correta para uma questão, alterar uma tarefa para atender às necessidades do usuário, explicar o conteúdo, etc (NWANA, 1990; WOOLF, 2008).

Há algumas tarefas que o tutor precisa desempenhar para realizar seu trabalho com eficiência; uma delas é realizar alterações no modelo do aluno, permitindo que ele avance em seus estudos. O tutor deve estar sempre monitorando as habilidades do aprendiz e conseguir decidir quando o estudante possui conhecimento suficiente sobre o assunto para ser apresentado às tarefas mais avançadas, ou até mesmo instruir o estudante a progredir para o próximo assunto. O tutor deve também fornecer assistência não solicitada quando detectar que o aprendiz se beneficiará das mesmas. Este tipo de intervenção ocorre quando o aluno está tendo dificuldade em solucionar um desafio. Através do módulo especialista, o tutor possui ciência deste fato, e conhecendo o conteúdo ou dica adequada para mostrar ao aluno, ele o faz de modo a ajudá-lo a avançar. A recuperação do conteúdo adequado do banco de dados é também uma das tarefas que o tutor precisa desempenhar; ele consegue realizar esta atribuição somente quando conhece o nível de dificuldade e/ou pré-requisitos de conhecimentos que o conteúdo sendo buscado possui, bem como as habilidades do usuário em questão (VANLEHN, 1987).

O **módulo de interface de usuário** cria e gerencia um ambiente onde o usuário pode interagir com o sistema. Quaisquer ações que o agente tutor desempenha podem ser vistas somente quando mostradas através da interface de usuário, bem como quaisquer ações por parte do usuário (por exemplo, responder a alguma questão ou ler um conteúdo) somente podem ser capturadas pelo agente através da interface, como mostrado na Figura 4.

De acordo com Vanlehn (2006), existem alguns conceitos básicos para o desenvolvimento de STIs, conforme observado na Tabela 3. Toda vez que uma tarefa é designada ao usuário e ele se engaja na mesma, o tutor coleta informações sobre esta interação e atualiza o modelo do aluno. Quando for necessário designar uma nova tarefa ao usuário, o tutor passa por um processo de decisão que visa determinar qual a melhor tarefa que deve ser fornecida para aquele momento. Este ciclo de monitoramento, decisão, desígnio é chamado **ciclo externo**. Uma tarefa pode ser dividida em partes menores que apresentam desafios por si só. Cada uma destas partes é chamada **passo**, e representa etapas na evolução do conhecimento específico àquele assunto por

parte do aluno. Outras tarefas que o tutor deve conseguir desempenhar são: a detecção de quais etapas (passos) o aluno está tendo maior dificuldade; e se existe uma falha no fornecimento de uma assistência específica para o passo sendo trabalhado pelo aluno. O **ciclo interno** se refere a este monitoramento que o tutor deve desempenhar para garantir que os passos internos a uma tarefa estão sendo desenvolvidos, e que o aluno está recebendo assistência adequada para desenvolvê-los (VANLEHN, 2006).

Tabela 3: Terminologia de STIs.

Termo	Breve definição
Tarefa	Atividade designada para o aluno pelo tutor.
Passo	Uma parte menor da tarefa que pode ser avaliada como certa ou errada.
Ciclo Externo	Parte do funcionamento do STI responsável por escolher a tarefa a ser designada.
Ciclo Interno	Parte do funcionamento do STI responsável por avaliar os passos dados pelo aluno para a tarefa e fornecer assistência (ex. dicas) para o estudante em relação ao passo.

Fonte: Adaptado de Vanlehn (2006).

Um exemplo de sistema tutor inteligente é o PAT2Math (JAQUES et al., 2013). Neste sistema, alunos são desafiados a resolver problemas de álgebra, onde equações estão divididas em planos de acordo com suas dificuldades. Dentro de um plano, o aluno pode resolver qualquer equação, mas só pode resolver as equações de um próximo plano quando tiver resolvido todas do anterior. Ao resolver uma equação, o aluno pode informar cada passo da resolução ou fornecer diretamente a resposta final. Após a entrada da etapa, o sistema informa se o passo está ou não correto, pontuando o sucesso do aluno, ou fornecendo uma dica caso esteja incorreto. Caso o aluno esteja com dificuldade para resolver uma equação, ele pode pedir dicas para o sistema. Neste sistema, o tutor possui o conhecimento do domínio (resolver equações algébricas) e é responsável por corrigir os passos do aluno, e modelar o conhecimento do aluno e demonstrar como resolve um determinado passo.

3.2.1 Sistemas Tutores Afetivos

Conforme visto na Seção 2.1, as emoções possuem papel importante nos processos cognitivos. Dado que o aprendizado depende destes processos, pode-se assumir que as emoções possuem forte relação nos processos de aprendizado, conforme demonstrado nos trabalho de Sidney et al. (2005). Como os Sistemas Tutores Inteligentes possuem por objetivo o auxílio neste processo de aprendizado, a união dos STI com a Computação Afetiva busca integrar o conhecimento das emoções dos alunos no processo de modelagem e tomada de decisões desempenhadas pelo tutor (AMMAR et al., 2010).

Trabalhos da área integrando Sistemas Tutores Inteligentes com detecção automática de emoções são frequentes, e geralmente possuem como característica o uso das informações das

emoções expressadas pelos alunos durante a interação com o sistema para aprimorar o conhecimento e a efetividade das ações do agente pedagógico. Um exemplo que pode ser citado como um trabalho deste tipo foi desenvolvido por Gordon et al. (2016). Seu objetivo é ajudar crianças no processo de aprendizado de um segundo idioma. O trabalho faz a integração detectando as emoções dos estudantes através de reconhecimento automático por vídeos e inserindo-as no modelo de aluno do agente pedagógico. Para a realização da detecção é utilizado o software comercial Affdex mobile SDK, da empresa Affectiva Inc (cuja co-fundadora é Rosalind Picard) (MCDUFF et al., 2016). O Affdex realiza detecção facial em tempo real através do algoritmo de Viola, Jones et al. (2001), e extrai características relativas a algumas AUs de Ekman: duas relacionadas aos lábios e duas relacionadas às sobrancelhas, utilizando o classificador *Support Vector Machines* (SENECHAL; MCDUFF; KALIOUBY, 2015). O Affdex também realiza a classificação da expressão facial em uma das emoções básicas utilizando as AUs detectadas, além dos estados afetivos valência e engajamento, embora em seu artigo não relatem qual o método é utilizado para a classificação. No trabalho, engajamento e valência são extraídos utilizando um dispositivo que com o sistema operacional Android que transmite os dados para o tutor através da rede local. Através do modelo, o tutor se adapta a cada estudante e em sua interação demonstra atitudes específicas personalizadas para cada aluno, como satisfação, desapontamento e surpresa. Os resultados de seus experimentos indicam que tutores pedagógicos afetivos conseguem auxiliar de maneira mais efetiva no aprendido dos alunos, e seus resultados indicam um maior número de palavras aprendidas pelos alunos submetidos ao tutor afetivo.

Assim como em outros ambientes, em ambientes de aprendizagem, a detecção de emoções pode se dar através de múltiplas fontes de informação. A detecção multimodal de emoções em ambientes de aprendizagem, porém, possui certas particularidades. O trabalho de Harley et al. (2015) realiza a detecção multimodal de emoções em um sistema tutor inteligente utilizando expressões faciais detectadas por vídeo, sinais biométricos de tensão elétrica da superfície da pele (ativação eletrodérmica, ou EDA do inglês) e um questionário auto-avaliativo. O sistema de detecção de expressões é realizado utilizando o software FaceReader, da VicarVision (DEN UYL; VAN KUILENBURG, 2005). O FaceReader realiza a classificação da expressão facial em três etapas. Primeiramente, a face é encontrada utilizando um método chamado de *Active Template Method*, que encontra o rosto comparando a imagem à um molde deformável de um rosto, em um método semelhante ao descrito por Sung e Poggio (1994). Após, é utilizado o método *Active Appearence Model* (COOTES; TAYLOR et al., 2004) para redução de dimensionalidade da imagem e detecção dos *landmarks* da face. Por fim, é realizada a classificação utilizando uma rede neural multi-camadas. O FaceReader pode fornecer como saída uma variedade de informações como emoções básicas e *Action Units*. O trabalho conclui que as emoções relatadas no questionário possuíam forte correlação com as emoções detectadas pelo algoritmo de detecção de expressões faciais, porém os sinais capturados do sensor EDA não possuem correlação com nenhum dos outros dois métodos. Uma provável causa, de acordo com os autores, é a baixa valência das emoções presenciadas durante o experimento, que não desencadeiam uma diferença

elétrica na pele que possa ser sentida pelo sensor. Esta conclusão demonstra que existem importantes diferenças entre a detecção e o tratamento de emoções básicas e de emoções não-básicas quando se tratando de situações de aprendizagem.

Uma das principais emoções presentes em ambientes de aprendizagem é o engajamento. É de extrema utilidade para o processo de aprendizagem do aluno que ele experiencie o engajamento pelo maior tempo possível enquanto interagindo com o sistema, pois o engajamento está intimamente ligado com aumento de produtividade e melhoria na leitura (DEWAN; MURSHED; LIN, 2019). Devido a esta importância, trabalhos que se propõe a detectar o engajamento em situações de aprendizagem têm se tornado bastante populares. Um exemplo a ser citado é o trabalho de Monkaresi et al. (2016) que realiza a detecção do engajamento de estudantes durante o desenvolvimento de uma tarefa em frente ao computador. O experimento se utilizou de um sensor de batimentos cardíacos e de câmeras do dispositivo *Microsoft Kinect*. Através do software de detecção facial das câmeras foram extraídas as características faciais dos vídeos, em ANUs (*Animation Units*) um sistema similar às *Action Units* de Ekman. Os autores utilizaram diversos classificadores presentes na ferramenta WEKA (HOLMES; DONKIN; WITTEN, 1994) e obtiveram resultador que superaram trabalhos estado-da-arte da época com seu modelo multimodal.

4 DETECÇÃO AUTOMÁTICA DE EMOÇÕES POR FACE

Quando pessoas detectam emoções de outras, utilizam métodos multimodais que obtém como entrada informações como tom de voz, expressão facial e linguagem corporal e geram como saída uma classificação da emoção. De acordo com Kreifelts, Wildgruber e Ethofer (2013), um dos sentidos do ser humano mais utilizado para o discernimento das emoções é a visão. Portanto, naturalmente as expressões faciais, perceptíveis através da visão, desempenham um grande papel na detecção da emoção. Como as emoções são importantes no processo cognitivo humano, naturalmente a detecção das emoções através das expressões faciais é um interessante e promissor ramo de estudo.

A criação de algoritmos que realizam este processo tem evoluído substancialmente desde trabalhos pioneiros como de Mase (1991). Com a melhoria de fatores como maior e mais barato poder de processamento das máquinas e melhores técnicas de reconhecimento facial, o campo de detecção de emoções pela face ganhou um aumento de popularidade nos últimos anos (FASEL; LUETTIN, 2003).

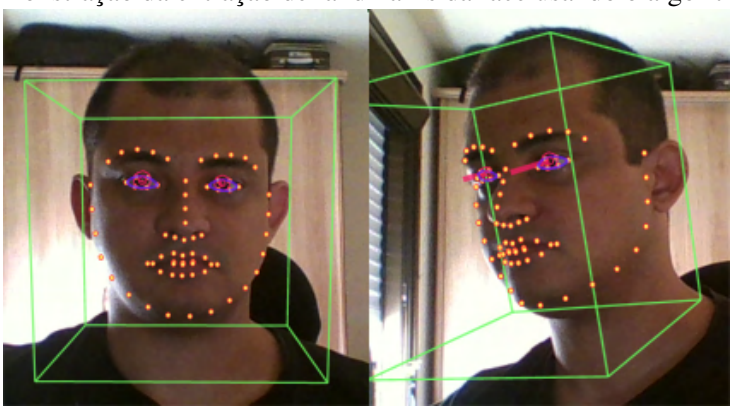
Para a realização da detecção automática de emoções pela face, primeiramente é necessária a análise da expressão facial do indivíduo. Vale notar que embora possam parecer equivalentes, ambas são disciplinas diferentes, pois a análise de emoções requer um julgamento psicológico do indivíduo, enquanto a expressão facial é um fenômeno observado fisicamente, que muitas vezes não condiz necessariamente a uma emoção (FASEL; LUETTIN, 2003). Ao realizar a detecção automática de emoções através de imagens de vídeo da face, normalmente precisa-se passar por três etapas (CASTILLO et al., 2018):

- Detecção da face na imagem
- Extração de características faciais
- Classificação da emoção

Para a detecção da face na imagem, alguns algoritmos comumente utilizados podem ser citados, como Jones e Viola (2003), HOG-SVM (DALAL; TRIGGS, 2005) ou MTCNN (ZHANG et al., 2016). Enquanto os dois primeiros se utilizam de técnicas de abordagens convencionais, o MTCNN se utiliza de redes neurais convolucionais profundas para realizar a detecção da face e *landmarks* e é atualmente o algoritmo que apresenta melhores resultados na detecção facial. *Facial landmarks* são pontos que compõe um padrão na geometria facial humana que correspondem ao contorno do rosto e de outros pontos-chave como nariz, boca e olhos (Figura 5). Algoritmos de detecção de expressões faciais analisam imagens ou sequências de imagens e buscam identificar os *landmarks*. Os principais desafios na detecção facial por imagem, de acordo com Fasel e Luettin (2003), são:

- Extrair o ruído da imagem da face: às vezes, o contraste dos objetos ao fundo dificulta a identificação do contorno do rosto.

Figura 5: Demonstração da extração de landmarks da face usando o algoritmo OpenFace.



Fonte: Elaborado pelo autor.

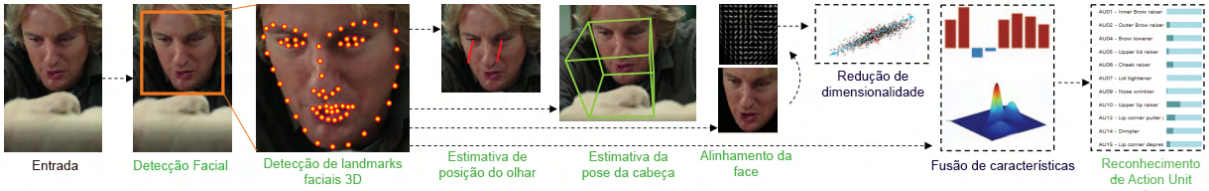
- Determinação da pose: como demonstrado da Figura 5, em grande parte das vezes, o rosto não está em posição frontal. Nestes casos, o algoritmo precisa realizar uma compensação e estimar os *landmarks* faltantes.

- Iluminação inadequada: interfere no contraste entre os objetos da cena e dos rostos, dificultando a detecção.

A detecção facial é uma área de pesquisa amplamente explorada, e hoje em dia existem diversos trabalhos que realizam a tarefa com precisão, como de Ramanan e Zhu (2012). Outros notáveis exemplos são as bibliotecas de visão computacional OpenCV (BRADSKI, 2000) e OpenFace (BALTRUSAITIS et al., 2018). Por serem projetos abertos e bem documentados, proporcionam que futuros pesquisadores da área se beneficiem de seus resultados. O OpenCV é uma ferramenta de visão computacional de uso geral, que possui mais de 2500 algoritmos tanto clássicos quanto mais recentes que realizam tarefas que abrangem detecção e reconhecimento facial, rastreamento e extração de modelos 3D de objetos, tratamento de imagens, dentre outros. O OpenFace é uma ferramenta específica para comunidade computação afetiva e para o desenvolvimento de aplicações baseadas em análise de comportamento facial. Conforme mostrado na Figura 6, ele implementa detecção facial, detecção de *landmarks*, movimento dos olhos, posicionamento da cabeça, e extração de características como reconhecimento de *Action Units*. Além disso, o OpenFace obteve ótimo desempenho nestas tarefas, conseguindo realizar seu processamento em tempo real e obtendo desempenho compatível aos melhores resultados observados em outras implementações (BALTRUSAITIS et al., 2018).

A extração de características faciais (*facial features*, do inglês) é a segunda etapa para a detecção automática de emoções através da face. De acordo com Kumari, Rajesh e Pooja (2015), as características faciais são representações abstratas de informações importantes sobre a descrição de regiões da face. Algumas das características tradicionalmente utilizadas no reconhecimento facial através de imagens são LBP-TOP (ZHAO; PIETIKAINEN, 2007), HOG (DÉNIZ et al., 2011), e as Action Units (EKMAN, 1992).

Figura 6: Visão geral das etapas e funcionalidades do OpenFace.

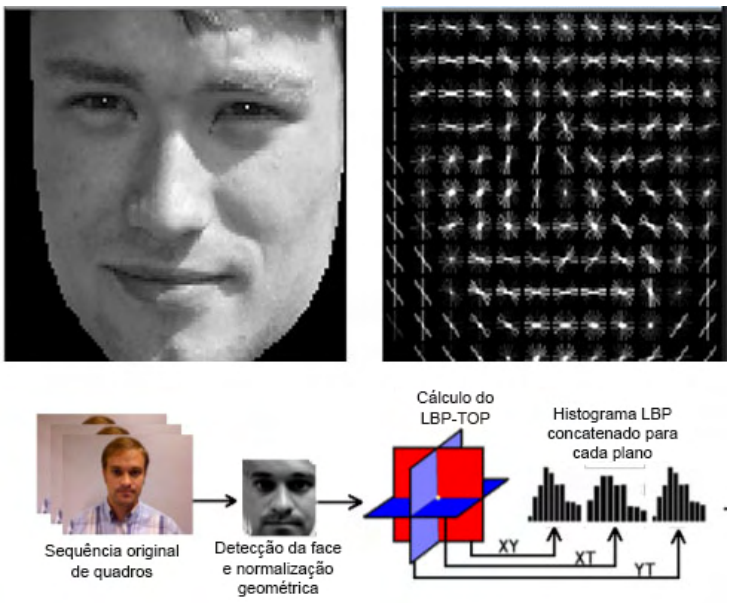


Fonte: Adaptado de Baltrusaitis et al. (2018).

LBP-TOP utiliza o conceito dos padrões binários locais (*Local Binary Patterns — LBT*, do inglês) das imagens bidimensionais e os estende para vídeos (AHONEN; HADID; PIETIKÄINEN, 2004). O princípio do LBP é separar a imagem em matrizes quadradas de um tamanho específico e, baseando-se nos valores dos píxeis, atribuir valores binários a esta matriz, que quando convertidos em decimais são organizados em um histograma. O LBP-TOP utiliza histogramas montados a partir das matrizes LBP dos planos XY, XT e YT e os concatena.

HOG, ou histograma de gradientes orientados (*Histogram of Oriented Gradientes*, do inglês), é uma técnica de extração de características da imagem, que calcula o gradiente das regiões da imagem, isto é, o quão abrupta a mudança de valores dos píxeis é de uma região para outra. Estes gradientes são úteis para a detecção de bordas de objetos. Através destes gradientes que possuem uma intensidade e direção, representados na imagem por vetores, é realizado um cálculo para representá-los em um histograma que define as características dos objetos presentes nas imagens. A Figura 7 demonstra um exemplo de extração de características de imagens utilizando as técnicas LBP-TOP e HOG.

Figura 7: (a)*Local Binary Patterns on Three Orthogonal Planes* (LBP-TOP) e (b) *Histogram of Oriented Gradients* (HOG).



Fonte: Adaptado de (a) Déniz et al. (2011) e (b) Freitas Pereira et al. (2012).

Por último, na etapa de classificação da emoção, as características faciais encontradas na segunda etapa são fornecidas como entrada em um algoritmo de aprendizado de máquina, que visa, por exemplos fornecidos, encontrar padrões que generalizam estes exemplos e possam ser aplicados a outros dados de entrada até então desconhecidos. Os exemplos conhecidos, neste contexto, são sequências de imagens extraídas de vídeos com rótulos textuais que indicam o estado afetivo demonstrado no exemplo. Este processo se chama aprendizado supervisionado. Após o aprendizado (chamado treinamento) concluído, novos dados cujos rótulos não foram informados no treinamento são fornecidos e o algoritmo é executado novamente com estes dados, porém, desta vez, os rótulos são estimados através do processo chamado classificação. A partir dos resultados da classificação, consegue-se estimar a taxa de acerto que o algoritmo obterá ao ser exposto a dados desconhecidos. Para classificação de emoções, os algoritmos tradicionalmente usados são as árvores de decisão, redes neurais multicamadas e máquinas de vetores de suporte (HUANG; LECUN, 2006; KUMARI; RAJESH; POOJA, 2015).

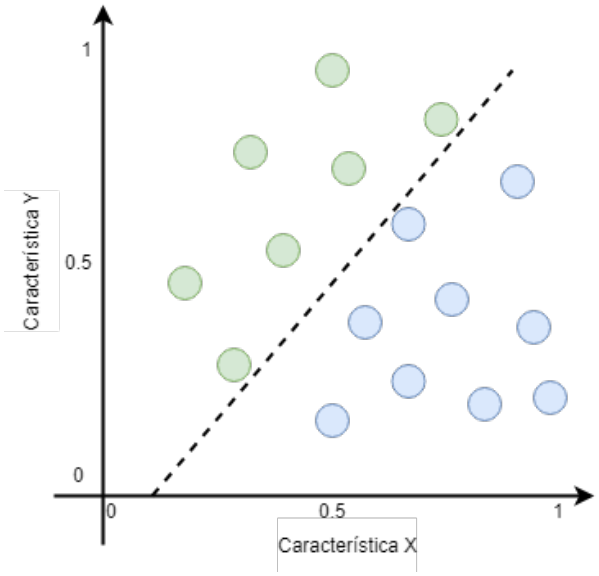
4.1 Redes Neurais Artificiais

Para realizar a classificação de padrões, uma das técnicas mais utilizadas na atualidade são redes neurais artificiais. Rede neural é uma técnica de aprendizado de máquina que possui inspiração nas redes neurais naturais (SIQUEIRA-BATISTA et al., 2014). Em uma rede neural, há nodos que se conectam e são organizados em camadas. Um nodo de uma camada pode ativar outro nodo da próxima camada se ele passar um sinal (um valor numérico) forte o suficiente. Uma vez que um nodo receba um sinal, ele o multiplica por um peso e usa uma função de ativação (como limiar, ou sigmoide) para determinar se ele ativará ou não (MITCHELL et al., 1990).

Redes neurais primitivas (chamadas perceptron) aprendiam um determinado padrão somente caso ele fosse linearmente separável, isto é, diante de um determinado espaço dimensional, em que os exemplos rotulados podem ser divididos em diferentes conjuntos através de uma linha. A Figura 8 demonstra um exemplo onde um classificador simples separa amostras de dados em duas classes. Este era o tipo de classificação que podia ser feita pelos perceptrons, e esta grande limitação foi superada através da criação de redes multicamadas, capazes de aprender padrões mais complexos.

Há três tipos de camadas em uma rede neural: entrada, saída e camadas escondidas. Camada de entrada é onde a informação é inserida, neste caso as imagens para detecção facial (codificadas em uma matriz que relaciona as posições de seus píxeis com os canais RGB). A camada de saída é onde o resultado esperado é informado, neste caso os rótulos das emoções. As camadas escondidas recebem valores automaticamente através da execução do algoritmo *backpropagation*, em que o desvio entre o rótulo da saída observada e esperada é usado em uma fórmula sofisticada de correção de peso, onde as mudanças são aplicadas na direção inversa (OSÓRIO; BITTENCOURT, 2000).

Figura 8: Separação linear de dois padrões através de um perceptron.



Fonte: Elaborado pelo autor.

Desde então, as redes neurais evoluíram muito, e hoje elas conseguem detectar padrões muito complexos, beirando ou ultrapassando a habilidade humana em alguns casos (HE et al., 2015). Assim como em outros algoritmos de aprendizado de máquina, o aprendizado nas redes neurais pode ser supervisionado ou não supervisionado. No primeiro caso, devem-se fornecer informações para que a rede possa aprender os padrões, bem como rótulos indicando como ela deve classificar cada informação. O objetivo deste processo é que com uma abundância de informações de entrada, o algoritmo de ajuste dos pesos deverá encontrar padrões de semelhança entre as informações de entrada, e assimilá-las aos rótulos fornecidos, assim como bebês aprendem novas informações.

O processo de aprendizado utilizando uma rede neural para treinamento de reconhecimento de emoções por face é supervisionado. No processo de treinamento, a informação é fornecida e propagada através dos nós das camadas de maneira que a cada nova camada transposta, características que compõe os padrões da informação fornecida são aprendidos e repassados às camadas seguintes. No caso da detecção de emoções através da imagem da face, a rede acaba conhecendo informações espaciais que determinam como um rosto deve parecer para que determinada emoção esteja presente. Após este processo, a rede está pronta para receber como entrada imagens não rotuladas. A rede então poderá fazer uma predição das emoções na camada de saída.

4.1.1 Métricas de treinamento

As métricas de treinamento de redes neurais são usadas para medir o desempenho do modelo durante o treinamento. Elas ajudam a avaliar se o modelo está aprendendo corretamente a partir

dos dados e se ele está fazendo previsões precisas. Além disso, as métricas de treinamento ajudam a determinar se o modelo está tendo *overfitting* (quando ele se ajusta muito bem aos dados de treinamento, mas não consegue generalizar bem para novos dados) ou *underfitting* (quando ele não se ajusta bem nem aos dados de treinamento, nem a novos dados). Monitorar as métricas de treinamento é importante para determinar quando parar o treinamento e para avaliar a qualidade do modelo resultante. Algumas das métricas mais comuns são resumidas abaixo (GOODFELLOW; BENGIO; COURVILLE, 2016; POWERS, 2011; FAWCETT, 2006):

- **Acurácia**: A porcentagem de previsões corretas feitas pelo modelo em relação ao número total de previsões.
- **Erro Quadrático Médio (MSE)**: A média da soma dos quadrados das diferenças entre as previsões do modelo e os valores alvos.
- **Binary cross-entropy**: Mede a distância entre a distribuição de probabilidade prevista pelo modelo e a distribuição alvo para problemas de classificação binária.
- **Categorical cross-entropy**: Mede a distância entre a distribuição de probabilidade prevista pelo modelo e a distribuição alvo para problemas de classificação multiclasse.
- **F1 Score**: É a média harmônica entre a **Precisão** e o **Recall**, e é uma medida mais robusta que a acurácia quando há um desbalanceamento entre as classes.
- **Área sob a Curva ROC**: É uma métrica calculada a partir da curva ROC, que representa a relação entre a taxa de verdadeiros positivos (recall) e a taxa de falsos positivos. A AUC-ROC fornece uma medida de quanto bem o modelo é capaz de distinguir entre duas classes, e quanto maior a AUC-ROC, melhor o modelo é capaz de fazer previsões corretas.

A Acurácia é uma métrica simples e amplamente utilizada que mede a porcentagem de previsões corretas em relação ao total de previsões efetuadas pelo modelo. É uma boa opção quando o desbalanceamento de classe não é um problema importante e quando é importante obter uma medida geral da precisão do modelo. A Equação 4.1 descreve o método de cálculo da acurácia:

$$Acc = \frac{TP + TN}{Total} \tag{4.1}$$

Onde TN são as previsões corretas para a classe positiva, TN são as previsões corretas para a classe negativa e $Total$ é o número total de previsões feitas pelo modelo.

O F1 Score, por outro lado, é uma métrica mais robusta que considera tanto a precisão quanto o recall. Ele é especialmente útil quando o desbalanceamento de classe é um problema importante, pois penaliza modelos que têm alta precisão, mas baixo recall, ou baixa precisão,

mas alto recall. A Equação 4.3 demonstra o cálculo do F1, enquanto a Equação 4.3 e Equação 4.4 demonstra os cálculos da precisão e recall, utilizados no cálculo do F1.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

Onde TP são as previsões corretas para a classe positiva, FP são as previsões erradas para a classe positiva, FN são as previsões erradas para a classe negativa.

O presente trabalho utilizou as métricas de avaliação Acurácia e F1. A acurácia, embora não tão recomendada quando trabalhando com dados desbalanceados, é a métrica mais comumente vista em trabalhos de reconhecimento de emoções por face, assim sendo, é uma métrica importante para realizar comparações com trabalhos relacionados. A métrica F1, por sua vez, foi utilizada com o intuito de realizar um comparativo de desempenho entre os próprios modelos desenvolvidos, considerando o alto desbalanceamento das classes observado das bases de dados utilizadas, conforme abordado em na seção 6.2.2.

4.1.2 Tipos comuns de redes neurais

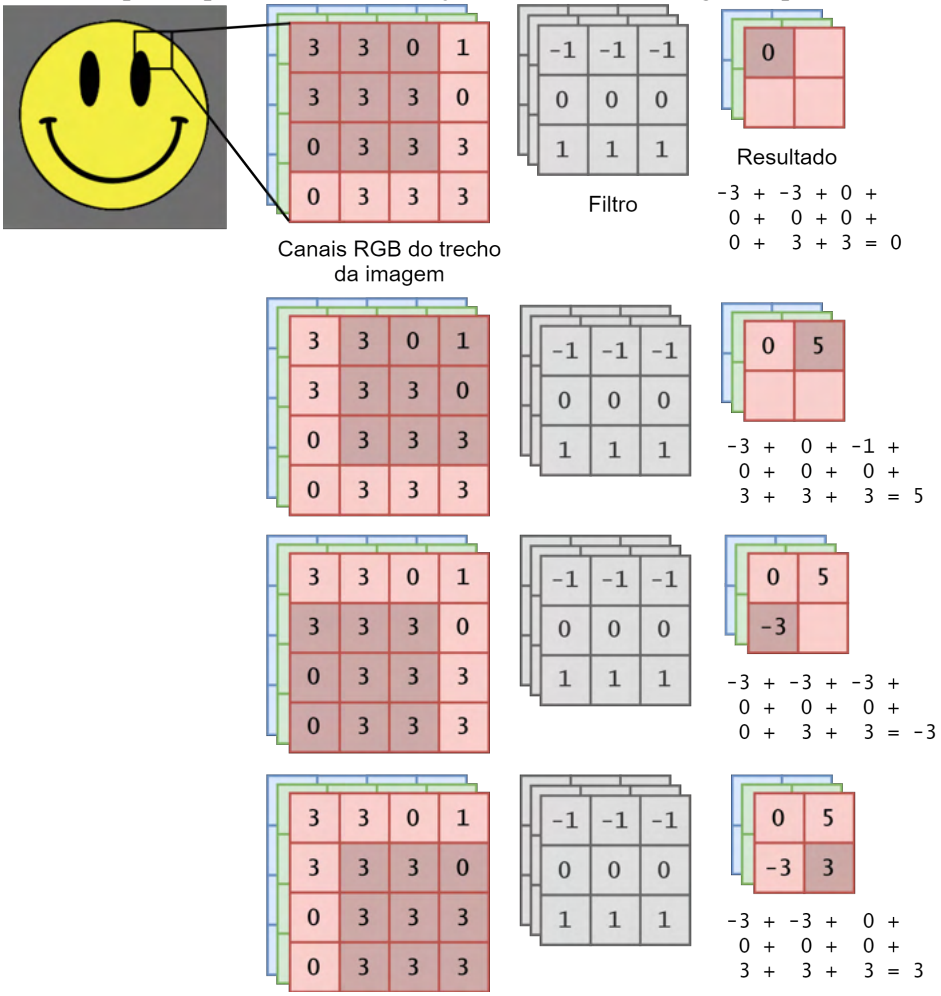
O tipo de rede tipicamente utilizada na análise de imagens é chamada rede neural convolucional (*Convolutional Neural Network*, CNN do inglês). Em contraponto às tradicionais redes neurais com nós totalmente conectados (aqueles em que sua saída é conectada à entrada de todos os nós da próxima camada), as CNNs possuem conexões esparsas, de maneira que um nó é conectado somente a outros nós que descrevem uma determinada região espacial. Este tipo de arquitetura foi inspirada no córtex visual animal, onde determinadas regiões são ativadas somente por estímulos em determinadas áreas do campo visual do indivíduo. Uma rede neural convolucional realiza operações de convolução com as informações recebidas, resultando em uma redução da dimensionalidade dos dados, mantendo os aspectos principais sobre a informação.

Na Figura 9 pode-se observar um exemplo do processo de convolução, onde o *smiley* representa uma imagem qualquer que passará pelo processo. A imagem é decomposta em matrizes de *píxeis* dos componentes RGB (*Red, Green, blue*). Trechos da imagem são multiplicados pela matriz filtro e seus resultados, após somados, compõe uma nova matriz de dimensão reduzida, conforme demonstra a Figura 9. O processo de convolução faz com que se mantenham características importantes que definem a imagem, como bordas de objetos, por exemplo, enquanto a dimensão das informações que representam tais características é reduzida.

As redes convolucionais são aplicadas atualmente em uma enorme gama de trabalhos que

vão desde detecção de pessoas e objetos em imagens, reconhecimento de caracteres, reconhecimento de expressões faciais e emoções, dentre outros.

Figura 9: Exemplo do processo de convolução de uma CNN. A figura representa uma imagem.

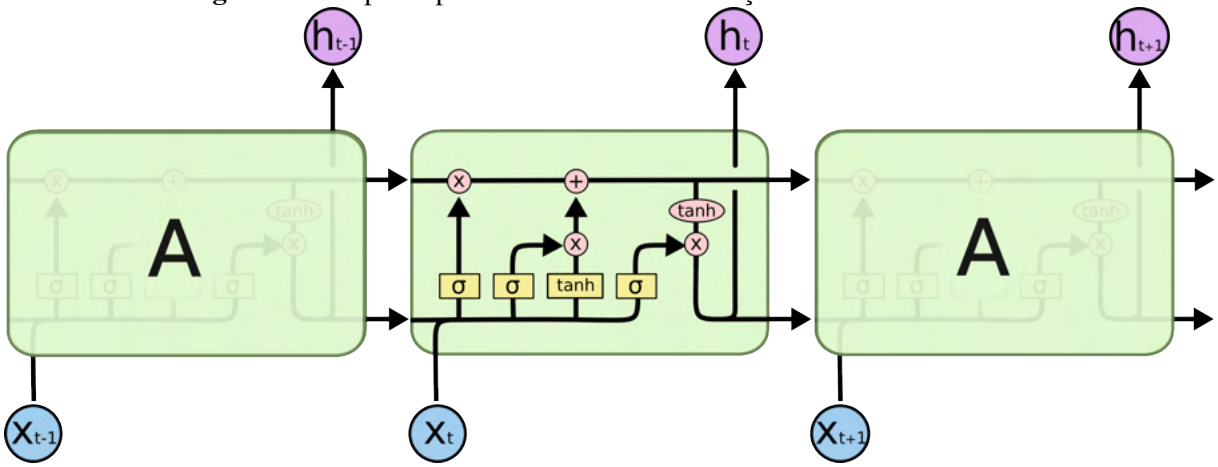


Fonte: Elaborado pelo autor.

Outro tipo de rede neural que têm ganhado bastante destaque são as Redes Neurais Recorrentes (*Recurrent Neural Networks*, ou RNN). As redes recorrentes são semelhantes às redes tradicionais, com a particularidade que elas conseguem trabalhar com informações sequenciais. RNNs são especialmente úteis quando trabalhando com dados temporais, onde uma dada informação depende da informação de estados anteriores (MIKOLOV et al., 2010). Elas podem guardar informações alimentando a camada de entrada com informações obtidas na saída, criando um *loop*. Um fato conhecido sobre as RNNs é que elas são difíceis de treinar adequadamente devido às dependências de longo-prazo, conforme explicado em Pascanu, Mikolov e Bengio (2013). Para resolver este problema, as redes *Long Short-Term Memory* (LSTM) foram introduzidas (HOCHREITER; SCHMIDHUBER, 1997), e mais recentemente as *Gated Recurrent Unit* (GRU) (CHO et al., 2014). Elas possuem a capacidade de esquecer informação, pois possuem um chaveamento que permite informar o quanto das informações antigas serão mantidas e o quanto das informações serão propagadas para as próximas camadas.

Na Figura 10, pode-se observar o diagrama de uma rede LSTM, onde cada bloco representa a informação em um dado instante de tempo. A linha superior, chamada estado da célula, é por onde a informação temporal flui. **X** é a entrada em dado instante de tempo (como um quadro de um vídeo, por exemplo). **H** são as saídas em cada instante de tempo. Em uma LSTM, existem três chaves principais que controlam o fluxo da informação de um instante de tempo para o outro. A primeira chave indica o quanto das informações antigas serão mantidas. A segunda chave representa o quanto da informação nova é usada para a geração da nova informação. E a terceira chave indica o quanto desta nova informação será usada para a geração da saída da etapa em questão.

Figura 10: Etapa do processamento da informação em uma rede LSTM.

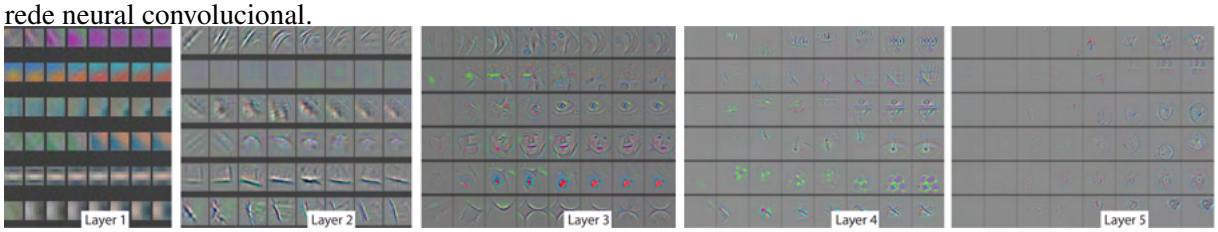


Fonte: (OLAH, 2015).

Embora seja teoricamente possível aprender funções complexas dada uma rede com largura suficiente (como visto em (ANDONI et al., 2014)), uma rede mais larga (com mais nós em cada camada) possui melhor capacidade de memorização de padrões, o que em certas situações pode levar ao *overfitting*, que é quando a rede consegue aprender muito bem diante dos exemplos fornecidos, mas possui baixa capacidade de generalização para novos exemplos. A adição de camadas em uma rede faz com que cada camada posterior aprenda características mais abstratas, mais especializadas e relacionadas com a classe da imagem em si e menos com a representação espacial da mesma. A Figura 11 mostra uma representação espacial da ativação das camadas de uma rede do trabalho de Zeiler e Fergus (2014).

As redes neurais ganharam atenção na atualidade devido a avanços em áreas de classificação de imagens, reconhecimento de fala, texto, dentre outros. Tais façanhas puderam ser desempenhadas devido às redes neurais profundas, que deram origem a uma área de pesquisa cunhada aprendizagem profunda, ou *Deep Learning* em inglês. De acordo com LeCun, Bengio e Hinton (2015), as redes neurais profundas são redes neurais com múltiplas camadas, capazes de aprender por representações incrementalmente mais abstratas de conceitos, de maneira que a composição de tais transformações formam funções complexas. Tais redes são geralmente combinadas com redes convolucionais para classificadores de imagens e vídeos ou redes recorrentes

Figura 11: Representação espacial das ativações dos mapas de características em cada camada de uma rede neural convolucional.



Fonte: (ZEILER; FERGUS, 2014).

para dados sequenciais (GOODFELLOW; BENGIO; COURVILLE, 2016).

Na tarefa da classificação de emoções, redes neurais profundas podem ser aplicadas tanto no reconhecimento facial, extração das faces alinhadas, extração de características faciais, ou treinamento das redes. No treinamento, podem ser utilizadas para reconhecer emoções recebendo como entrada imagens ou vídeos em que tais emoções estejam presentes utilizando camadas convolucionais juntamente com camadas recorrentes, como em Ebrahimi Kahou et al. (2015), ou na tarefa da classificação através das AUs do FACS, utilizando camadas recorrentes, como em Niu et al. (2018).

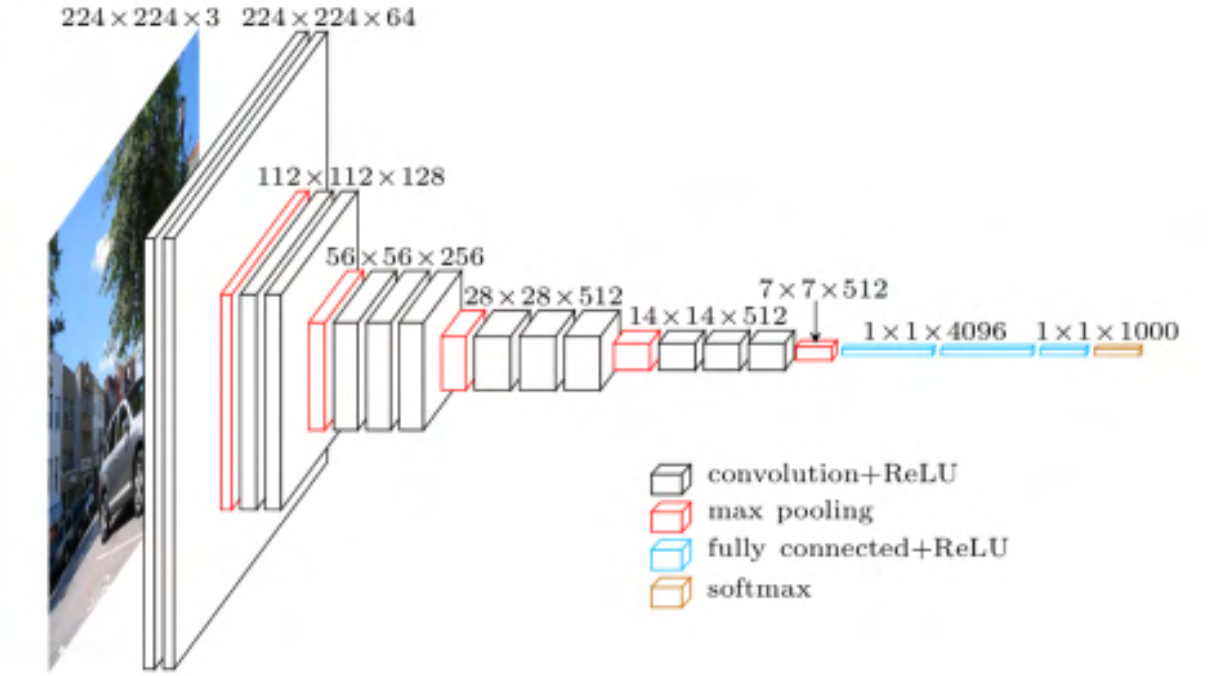
Existem diversas arquiteturas de redes neurais profundas que têm obtido grande sucesso na tarefa de classificação de classes de objetos. Um exemplo recente é a rede VGG16 (onde o 16 representa número de camadas treináveis), que conforme pode ser observado na Figura 12, possui blocos compostos por duas ou três camadas convolucionais seguida por uma camada *pooling*. A rede é composta por cinco destes blocos, e três camadas totalmente conectadas ao fim. Esta rede foi treinada na base de dados ImageNet (RUSSAKOVSKY et al., 2015) para o reconhecimento de 1000 classes diferentes de objetos.

Um grande problema em adicionar muitas camadas ao modelo neural é o chamado problema do gradiente de desaparecimento (*vanishing gradient*, do inglês). Este problema ocorre quando o erro de uma rede muito longa é propagado, fazendo com que multiplicações sucessivas aproximem seu valor de zero, tornando a rede incapaz de continuar aprendendo. Por este motivo, redes como a VGG16 eram consideradas muito profundas até poucos anos atrás.

A rede ResNet, proposta por He et al. (2016), visa solucionar este problema. O diferencial de uma ResNet está em seu módulo residual (Figura 13). Este módulo cria um atalho da informação que chega da entrada até a saída do bloco. A saída normalmente computada é somada a esta informação que percorreu o atalho, efetivamente fazendo com que o erro da rede profunda nunca seja maior que o de suas versões mais rasas. Uma rede completa do tipo ResNet possui diversos módulos residuais, ligados um após o outro.

Uma evolução recente da rede ResNet, é a rede DenseNet, do trabalho de Huang et al. (2017). A diferença conceitual entre as duas é que na DenseNet, as saídas de cada camada são propagadas para todas as próximas camadas, ao invés de somente para a posterior, como num módulo residual da ResNet, conforme demonstrado na Figura 14. A vantagem deste tipo de arquitetura, relatam os autores, é que por cada camada possuir como entrada uma conexão

Figura 12: Ilustração das camadas de uma rede VGG16.

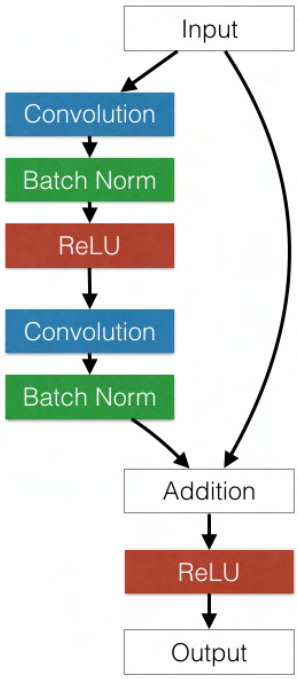


Fonte: (Frossard, Davi, 2019), criada a partir da descrição textual de Simonyan e Zisserman (2014).

mais direta com as características de camadas anteriores, o problema do gradiente de desaparecimento é diminuído, incentivando a propagação de características e possibilitando a inclusão de mais camadas no modelo sem que isto se torne um problema.

Uma alternativa ao processo de treinamento completo de uma rede neural, o que pode levar muito tempo e demandar um hardware custoso e especializado, é a transferência de aprendizado (*Transfer Learning*, do inglês). Nesta modalidade utiliza-se uma rede onde são conhecidos a topologia e os pesos atribuídos aos nós da rede durante seu treinamento. Em posse destas informações, é possível fazer com que ela aprenda de maneira muito mais rápida se comparada ao treinamento completo. Sabendo-se que as camadas iniciais de uma rede convolucional são responsáveis por guardar informações das características mais gerais dos exemplos fornecidos ao modelo, e as camadas finais da pelas características mais específicas e de classificação, pode-se utilizar *transfer learning* para obter o vetor de características simplesmente removendo as camadas finais da rede. Este tipo de processo é útil quando se deseja alimentar outro algoritmo de aprendizado de máquina com as características descobertas pela CNN. Outro caminho é o ajuste fino (*fine tuning*, do inglês), que é o processo de retreinar a rede com os exemplos desejados, porém impedindo que as camadas iniciais modifiquem seus pesos. Desta maneira, somente as camadas finais aprendem as informações importantes sobre a classificação dos exemplos fornecidos, enquanto a rede mantém as informações gerais das características originalmente aprendidas durante o seu treinamento inicial. Uma das grandes vantagens da utilização do *transfer learning*, além da economia de recursos computacionais, é o uso de redes comprovadamente eficientes no processo. No entanto, para as características iniciais serem de algum valor para

Figura 13: Módulo residual da ResNet.



Fonte: (Gross, Sam and Wilber, Michael, 2016).

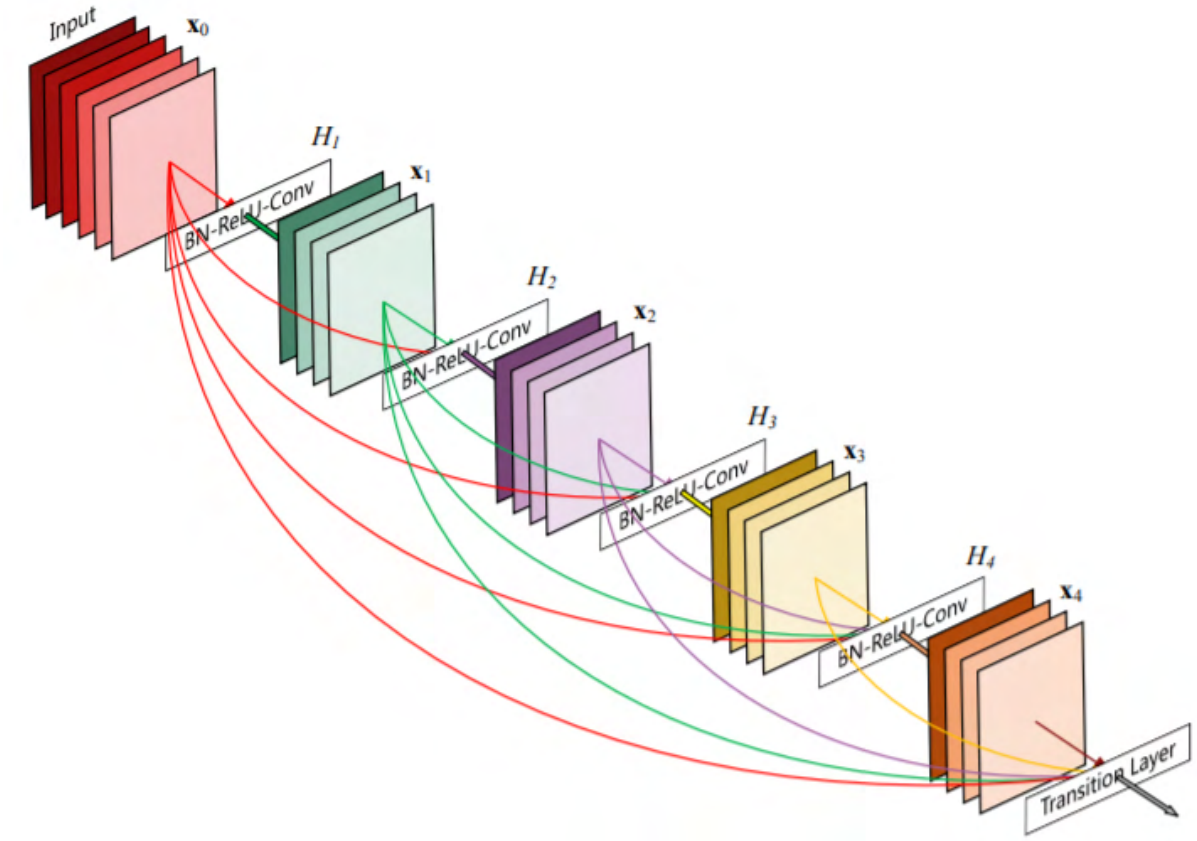
o problema em questão, a rede deve ter sido originalmente treinada em uma base de dados semelhante à base que deseja-se realizar o *fine tuning*.

4.2 Trabalhos e aplicações

Para realizar a detecção automática de emoções pela face é necessária a detecção de face na imagem, extração de características faciais e classificação da emoção. O trabalho de Jyoti, Sharma e Dhall (2018), por exemplo, integra estas três etapas, criando um detector de emoções a partir da face utilizando três redes neurais hierárquicas distintas. A primeira rede detecta a face do usuário descobrindo as *face landmarks*. A segunda é uma rede neural convolucional e é utilizada para detecção das características, especificamente as *Action Units* do sistema FACS. A terceira rede é composta por camadas de neurônios ligados densamente e é responsável pela classificação das AUs em uma emoção. Todas as redes são utilizadas simultaneamente, de maneira que a saída da primeira alimenta a entrada da segunda e a segunda alimenta a terceira. Como as informações são constantemente recebidas dos *frames* do vídeo, todas as redes ficam trabalhando paralelamente.

O trabalho de Jyoti, Sharma e Dhall (2018) também afirma que o uso das características faciais do sistema FACS para detecção de emoções é um método que foi mostrado benéfico pela literatura, embora necessite de uma detecção precisa dos pontos-chave faciais para se manter eficiente (DAHMANE; MEUNIER, 2011). Embora hajam discordâncias quanto ao método

Figura 14: Arquitetura de uma rede DenseNet.



Fonte: (HUANG et al., 2017).

utilizado para a extração de características, Khorrami, Paine e Huang (2015) realizou experimentos utilizando redes neurais não supervisionadas para extração de características faciais e concluiu que redes convolucionais baseadas em aparência descobrem características faciais muito similares às AUs em suas etapas de classificação. Zhang et al. (2015) corrobora com esta teoria quando extraiu automaticamente de uma rede não supervisionada características faciais associadas às AUs, que, por sua vez, são utilizadas com sucesso para detecção de emoções básicas.

Quanto às técnicas de classificação, Schmidhuber (2015) fala em seu trabalho que as redes neurais recorrentes (RNN), mais especificamente as redes de memória de longo prazo (LSTM e GRU) são amplamente utilizadas hoje em dia, alertando sobre a utilidade de um pré-treinamento não supervisionado para extração de características somado ao treinamento supervisionado para classificação. Na mesma linha, Kollias et al. (2017) defende que o uso de redes neurais profundas é o método mais em voga no momento para a detecção de emoções, e em seu trabalho ele utiliza uma rede convolucional para extração de características e uma rede recorrente na classificação das emoções, utilizando um banco de dados de vídeos capturados espontaneamente, em condições não controladas, o que é chamado de *in the wild*.

Além dos trabalhos citados, existem diversos outros trabalhos recentes que se propõem a

detectar emoções utilizando imagens e vídeo. Com a recente popularidade do uso de redes neurais profundas para classificação de emoções, surgiram também iniciativas comerciais que propõe soluções para este fim. Microsoft, Google e Amazon são exemplos de empresas que lançaram seus produtos para atender a esta demanda. A NeuroData Lab, uma empresa que desenvolveu uma destas soluções (KONSTANTINOVA; KAZIMIROVA; PEREPELKINA, 2018), e organiza o *International Workshop on Social & Emotion AI for Industry* (SEAIxI), realizou um comparativo entre sua solução e as soluções Microsoft Azure Face, Amazon Rekognition, Affectiva Affdex (Figura 15) (NeuroData Lab, 2019).

Figura 15: Comparativo entre soluções comerciais de detecção de emoções. O sinal de adição indica que o algoritmo acertou a emoção do fragmento de vídeo na maioria dos vídeos fornecidos, e as células destacadas indicam que o algoritmo obteve o melhor resultado de todos os comparados para a emoção analisada. S, R e A indicam os *datasets* utilizados.

	NEURODATA LAB			AFFECTIVA			MICROSOFT			AMAZON		
Dataset	S	A	R	S	A	R	S	A	R	S	A	R
Neutral	–	–	+	–	–	–	+	+	+	–	+	+
Anger	+	+	+	–	–	–	+	+	–	+	+	+
Disgust	–	–	+	+	–	+	–	–	+	–	–	+
Fear	+	+	+	–	–	+	–	–	–	–	–	–
Happiness	+	+	+	+	–	+	+	+	+	+	+	+
Sadness	+	+	+	+	–	–	+	+	+	+	+	+
Surprise	+	+	+	+	–	+	+	+	+	+	+	+

Fonte: (NeuroData Lab, 2019).

Com os bons resultados obtidos em trabalhos, e o crescente interesse da comunidade pelo tema, surgem competições que se propõe a eleger os melhores algoritmos classificadores automáticos de emoções. Um exemplo notável é a EmotiW (*Emotion Recognition in the Wild Challenge*), que possui o objetivo de definir uma plataforma para avaliação de métodos de reconhecimento de emoções em condições reais (DHALL et al., 2018). Com sua primeira edição em 2013, a EmotiW possui três categorias: *Group-level emotion recognition*, onde o algoritmo deve classificar como positivo, negativo ou neutro as emoções percebidas de imagens de grupos de pessoas em eventos sociais; *Audio-video Sub-challenge*, onde pequenos trechos de vídeo com áudio são fornecidos e o algoritmo deve classificar o vídeo como uma das seis emoções básicas; e *Engagement in the Wild*, que consiste em detectar o engajamento de indivíduos assistindo a vídeos educacionais. Como principais trabalhos da Seção de detecção de engajamento, podem-se citar (YANG et al., 2018) e (NIU et al., 2018), que obtiveram a primeira e segunda colocações na competição, respectivamente. O trabalho de Yang et al. (2018) utiliza a ferramenta Openface e Openpose para realizar a detecção analisando expressões faciais (através das AUs), movimento dos olhos e posição da cabeça. Além disso, outro *dataset* próprio também é utilizado além do *dataset* de treinamento fornecido para a competição. Eles se utilizam de redes

LSTM para obter os resultados de cada modalidade. O trabalho de Niu et al. (2018) também utiliza a Openface para extrair AUs, movimentos dos olhos e posição da cabeça. Eles implementam uma rede do tipo GRU recebendo como entrada um vetor de 117 dimensões com todas as características extraídas dos segmentos do vídeo. A saída da GRU é conectada em uma rede densa que obtém como resposta o nível de engajamento do segmento.

5 TRABALHOS RELACIONADOS

O presente trabalho realiza a detecção de emoções relacionadas à aprendizagem (engajamento, frustração, tédio e confusão), se utilizando de técnicas automáticas de extração de características faciais e classificação em emoções através da aplicação de redes neurais profundas. Neste contexto, o estudo de trabalhos anteriores que realizam a tarefa de detecção de emoções através da face permite identificar métodos e algoritmos aplicados para essa tarefa e salientar como o trabalho proposto avança o estado da arte.

Dessa forma, foi realizado um levantamento dos principais trabalhos na área, ressaltando suas características mais importantes de modo a identificar o estado da arte em detecção de emoções por face, elaborar um comparativo e salientar o diferencial do trabalho realizado. Inicialmente, foi empregada a técnica de revisão *Snowball* a partir de trabalhos selecionados por sua relevância no tema. Em seguida, foi realizada uma pesquisa sistemática no Google Acadêmico e nas bases de dados ACM e IEEE. Os trabalhos levantados foram classificados em duas classes, segundo o tipo de emoção que reconhecem: as emoções básicas e acadêmicas (ou de aprendizagem). A Seção 5.2 apresenta o estado da arte em relação aos trabalhos que *detectam emoções básicas* pela face e a Seção 5.3 as pesquisas que vêm trabalhado na *detecção de emoções relacionadas à aprendizagem*. Mesmo que o trabalho proposto objetive a detecção de emoções acadêmicas, optou-se por igualmente apresentar os trabalhos envolvendo o reconhecimento de emoções básicas, cuja pesquisa está mais avançada, pois envolvem o mesmo tipo de algoritmos e técnicas. Antes disso, na Seção 5.1 é apresentado o método aplicado para seleção dos trabalhos.

5.1 Método de revisão

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 básicas e detecção de emoções presentes no aprendizado, obtidos através da pesquisa que constituiu a busca inicial, foi possível elaborar uma *string* de busca com os termos mais relevantes para o trabalho proposto. Foram realizadas pesquisas nas bibliotecas digitais Google Scholar[1], ACM Digital Library[2] e IEEE Xplore Digital Library[3], reconhecidas por serem bases de pesquisa

[1] https://scholar.google.com
[2] https://dl.acm.org
[3] https://ieeexplore.ieee.org

acadêmica de publicações científicas. Para a seleção dos trabalhos relacionados foram feitas duas pesquisas, uma identificando os trabalhos sobre detecção de emoções básicas e outra sobre as emoções não básicas presentes na aprendizagem. O resultado das buscas bem como seus parâmetros podem ser analisados nas tabelas 4 e 5.

Tabela 4: Resultados obtidos pela pesquisa nas bibliotecas digitais utilizando a *string* de busca relativa às emoções básicas entre os anos de 2015 e 2019.

Fonte	**String de busca**	**Ocorrências**
Google Scholar	(facial expression OR emotion) AND (detection OR recognition) AND (deep AND (learning OR neural network)) AND ((happy AND sad AND angry AND (fear OR scared) AND disgust AND surprise) OR basic emotions)	16 300 resultados
ACM Digital Library		141 resultados
IEEE Xplore Digital Library		49 resultados

Fonte: Elaborado pelo autor.

Tabela 5: Resultados obtidos pela pesquisa nas bibliotecas digitais utilizando a *string* de busca relativa às emoções acadêmicas entre os anos de 2015 e 2019.

Fonte	**String de busca**	**Ocorrências**
Google Scholar	(engagement OR confusion OR frustration OR boredom) AND (detection OR recognition OR prediction OR classification) AND deep learning	103 000 resultados
ACM Digital Library		62 resultados
IEEE Xplore Digital Library		116 resultados

Fonte: Elaborado pelo autor.

No total da busca sobre emoções básicas foram analisados 240 artigos através da pesquisa usando a *string* de busca da tabela 4, sendo que foram excluídos da busca os artigos publicados antes de 2015. Somente os primeiros 50 retornados pela busca do Google Scholar foram analisados. Foram incluídos na seleção somente os trabalhos relacionados ao tema de detecção de emoções básicas, reconhecimento de expressões faciais, e *action units*. Destes, foram selecionados 146 artigos pelos seus títulos, e após a leitura de seus resumos, 20 foram selecionados para uma leitura completa. Seis trabalhos deste grupo foram considerados mais relevantes à presente proposta, descritos na Seção 5.2.

Na busca por emoções de aprendizagem, foram selecionados 221 trabalho usando a *string* de busca da tabela 5, sendo que, como realizado com a pesquisa sobre emoções básicas, foram excluídos os artigos publicados antes de 2015. Dada a ordem decrescente de relevância nos artigos retornados pelo algoritmo do Google, foram analisados os artigos retornados até que os resultados retornados deixassem de ser relevantes à pesquisa. Foram incluídos na seleção somente os trabalhos relacionados ao tema de detecção de emoções, aprendizado e engajamento. Destes, 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 5.3.

Para as duas buscas supracitadas, 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.

Durante o período da pesquisa do presente trabalho, diversas outras publicações que se enquadram com o tema de pesquisa foram publicados. Para contemplar estes novos trabalhos, uma nova pesquisa foi realizada ao fim do desenvolvimento dos modelos propostos neste trabalho, em novembro de 2022. Esta nova pesquisa visou complementar a revisão de trabalhos relacionados e atualizar os comparativos do trabalho desenvolvido com os trabalhos recém lançados, a fim de manter em perspectiva o desempenho do modelo desenvolvido perante o estado-da-arte.

Seguindo os mesmos critérios da primeira revisão, durante esta nova etapa, 38 trabalhos foram selecionados pelos seus títulos. Destes, 25 foram escolhidos pela leitura de seus resumos, e 17 para leitura completa. Ao fim, cinco trabalhos foram considerados mais relevantes à presente proposta, descritos na Seção 5.4.

5.2 Detecção automática de emoções básicas por face

O emprego das redes neurais profundas em tarefas de classificação trouxe ganhos significativos em diversas tarefas, como reconhecimento de imagens, áudio, e texto (LECUN; BENGIO; HINTON, 2015). Um exemplo de seu uso, relevante ao presente trabalho, é a tarefa de detecção de emoções através de vídeos ou imagens. As emoções básicas, por se embasarem em uma teoria amplamente reconhecida (EKMAN, 1992) e por serem úteis em uma grande gama de aplicações, foram as primeiras a serem estudadas com maior profundidade e, consequentemente, bases de dados que as relacionam também foram as primeiras a serem criadas.

Antes do uso de redes neurais profundas, uma das técnicas que despontava maior sucesso na tarefa de reconhecimento de emoções eram as *Support Vector Machines* (SVM) (HUANG; LECUN, 2006). O trabalho de Li e Lam (2015) utiliza uma rede neural profunda totalmente conectada para realizar a classificação de emoções básicas. Os autores comparam seus resultados aos trabalhos anteriormente tidos como estado-da-arte e obtém um ganho de 3,7% de acurácia sobre o uso de SVMs. O modelo do trabalho obteve uma acurácia de 91,7% e foi treinado utilizando a base de dados Extended Cohn-Kanade (CK+) (LUCEY et al., 2010), uma das mais conhecidas bases que contém imagens de pessoas demonstrando diferentes expressões faciais (Figura 16).

Assim como o CK+, existem várias outras bases de dados para treinamento de algoritmos de detecção de emoções básicas. No trabalho de Mollahosseini, Chan e Mahoor (2016) por exemplo, os autores treinam uma rede neural convolucional profunda com rótulos e imagens de diversas bases de dados e relatam a acurácia que seu modelo apresentou em cada uma delas, bem como a acurácia obtida ao treinar seu modelo em todas as bases. As bases utilizadas para

Figura 16: Exemplo de imagens demonstrando as seis emoções básicas da base de dados Extended Cohn-Kanade (CK+).

Fonte: Adaptado de Lucey et al. (2010).

os treinamentos do modelo estão relacionadas na Tabela 6, e representam as principais bases de dados anotadas de emoções básicas existentes. A estratégia utilizada no trabalho foi a utilização de camadas convolucionais seguidas de três módulos *inception*. Estes módulos são compostos de redes convolucionais treinadas em paralelo de modo que cada uma se especializa no aprendizado de características de regiões específicas da imagem. Para integrar ao modelo, as saídas das redes do módulo *inception* são concatenadas. A saída do último módulo é ligada a duas camadas totalmente conectadas, que geram a resposta da classificação. O trabalho obteve acurácia 94,7%, 77,9%, 55%, 76,7%, 47,7%, 93,2% e 66,4% nas bases de dados MultiPie, MMI, DISFA, FERA, SFEW, CK+ e FER2013 respectivamente, enquanto treinado individualmente em cada uma das bases. O treinamento realizado em todas as bases simultaneamente teve a finalidade de fornecer uma capacidade superior de generalização da rede (evitando *overfitting*), porém os resultados de acurácia neste tipo de treinamento foram bem abaixo dos resultados relatados nos treinamentos individuais, tendo sido: 47,7%, 55,6%, 37,7%, 39,4%, 39,8%, 64,2%, 34,0%.

A maioria dos trabalhos de redes neurais profundas aplicadas ao reconhecimento de expressões faciais se utilizam de redes convolucionais para a extração de características faciais e classificação. Todavia, existem trabalhos, como o de Zeng et al. (2018), que utilizam outros métodos de extração de características faciais, que são fornecidos como entrada para o classificador neural. O trabalho em questão usa os padrões extraídos pelas técnicas LBP (AHONEN; HADID; PIETIKÄINEN, 2004) e HOG (DÉNIZ et al., 2011) como características que descrevem as faces presentes nas imagens da base CK+. O modelo realiza a classificação através de um *deep sparse autoencoder* (DSAE), uma rede profunda que implementa o conceito de camadas sobrepostas de *sparse autoencoders* (SAE) (NG et al., 2011). SAE é uma técnica de aprendizado de máquina não supervisionada que, a partir de dados de entrada, mapeia correlações entre características desta entrada e reconstrói os dados na saída de forma que eles se assemelhem aos dados de entrada. O processo de codificação-decodificação do *autoencoder* força a rede a eliminar redundâncias na informação, mantendo a representação dos dados cor-

Tabela 6: Características principais de bases de dados sobre emoções básicas.

Base	Rótulos	Participantes	Observações
CMU MultiPIE (GROSS et al., 2010)	Neutro, sorriso, surpresa, olhos cerrados, nojo, grito	750000 imagens de 337 participantes	Múltiplos ângulos, diferentes níveis de iluminação
MMI (PANTIC et al., 2005)	AUs	2900 vídeos de 75 participantes	
Extended Cohn-Kanade (CK+) (LUCEY et al., 2010)	30 AUs, 6 emoções básicas e contentamento	486 sequências de 97 participantes	Sequências de imagens em ambiente controlado. Participantes são requisitados a demonstrar expressões.
Denver Intensity of Spontaneous Facial Actions (DISFA) (MAVADATI et al., 2013)	Intensidades entre 0-5 de 12 AUs	Vídeos de 4 minutos de 27 participantes	
FERA (BÄNZIGER; SCHERER, 2010)	Raiva, medo, felicidade, alívio, tristeza	189 sequências de 10 participantes	
Static Facial Expressions in the Wild (SFEW) (DHALL et al., 2011a)	6 emoções básicas	663 imagens de 95 participantes	Criado a partir de quadros estáticos da base AFEW. *in the wild*.
Facial Expression Recognition 2013 (FER2013) (Kaggle, 2013; GOODFELLOW et al., 2013)	6 emoções básicas	35887 imagens	Criado a partir de imagens de busca do Google, *in the wild*.

Fonte: Elaborado pelo autor.

relacionados. A rede DSAE aplicada no trabalho é treinada em um primeiro momento de forma não supervisionada de modo a encontrar os pesos da rede que mapeiam as características fornecidas na entrada. Após, é realizado um ajuste fino supervisionado com os rótulos de saída encontrados na base CK+. O trabalho conseguiu obter acurácia 95,79% no reconhecimento de sete classes de emoções utilizando o HOG como característica extraída da face, superando os algoritmos de estado-da-arte comparados.

Um tipo de rede neural que tem recebido considerável atenção em tarefas de classificação de vídeo são as redes neurais recorrentes (4.1.2), mais especificamente sua variante LSTM (*Long Short-Term Memory*). O trabalho de Kim et al. (2017), por exemplo, realiza a classificação das emoções básicas através de uma rede neural que se utiliza de informações espaciais e temporais. O modelo realiza a extração das características utilizando camadas convolucionais da rede. Para realizar a classificação, camadas LSTM são adicionadas, fazendo com que o treinamento considere a sequência de imagens como um processo contínuo, de modo a encontrar os padrões

temporais das características extraídas das sequências de imagens. A rede foi treinada nas bases MMI (PANTIC et al., 2005) e CASME II (YAN et al., 2014), e obteve acurácia 78,61% na base MMI.

Na mesma ideia de utilizar informações temporais para a realizar a classificação de vídeos, os trabalhos de Lu et al. (2018) e de Liu et al. (2018) implementam seus modelos classificadores utilizando redes recorrentes. Ambos são participantes da categoria *Audio-Video Sub-challenge* da EmotiW 2018 (DHALL et al., 2018). A EmotiW (*Emotion Recognition in the Wild Challenge*) é uma competição que teve sua primeira edição em 2013 e possui o objetivo de definir uma plataforma para avaliação de métodos de reconhecimento de emoções em condições reais. A EmotiW possui três categorias: (*i*) *Group-level emotion recognition*, em que o algoritmo deve classificar as emoções percebidas de imagens de grupos de pessoas em eventos sociais como positivo, negativo ou neutro; (*ii*) *Audio-video Sub-challenge*, em que pequenos trechos de vídeo com áudio são fornecidos e o algoritmo deve classificar o vídeo como uma das seis emoções básicas; e (*iii*) *Engagement in the Wild*, que consiste em detectar o engajamento de indivíduos assistindo a vídeos educacionais.

Outra competição de grande impacto na área das redes neurais profundas é a *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) (RUSSAKOVSKY et al., 2015). Os competidores da ILSVRC precisam utilizar a base de dados *ImageNet* (DENG et al., 2009) para realizar o treinamento de redes que classificam diversas categorias de objetos. Embora não seja o objetivo da competição realizar o treinamento de emoções, redes convolucionais bastante complexas, que demandaram altos recursos computacionais, foram treinadas para a competição. Estas redes podem ser utilizadas para classificação de outros padrões em imagens através do processo chamado ajuste fino (*fine tuning*, do inglês), onde a maioria dos pesos da rede pré-treinada é mantida (normalmente as camadas convolucionais), e somente as camadas finais (normalmente do tipo totalmente conectadas) passam pelo processo de treinamento. Desta maneira, através do ajuste fino, uma rede mantém a capacidade de generalização de características espaciais aprendidas pelas camadas convolucionais, enquanto é treinada para classificação em um conjunto de dados próprio. Algumas das redes mais famosas que podem ser citadas são AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), VGG (SIMONYAN; ZISSERMAN, 2014), GoogLeNet (SZEGEDY et al., 2015) e ResNet (HE et al., 2016).

Alguns trabalhos utilizam o ajuste fino destas redes para auxiliar na classificação de emoções, como é o caso do trabalho de Lu et al. (2018), que utiliza uma fusão de quatro classificadores com respectivos pesos para estimar a emoção básica presente nos trechos de vídeos da base AFEW (*Acted facial expressions in the wild* (DHALL et al., 2011b). O primeiro classificador utiliza a rede convolucional pré-treinada VGG por um ajuste fino para a extração de características, e um classificador LSTM bidirecional para as seis emoções básicas. A diferença de uma rede LSTM do tipo bidirecional para uma rede LSTM convencional é que enquanto a convencional utiliza informações temporais do passado para a classificação, a bidirecional considera também o futuro. No contexto em questão, a LSTM convencional usa as características

espaciais da expressão facial demonstrada no vídeo até o momento para determinar que emoção o indivíduo está presenciando. Já a LSTM bidirecional utiliza as mesmas características de quadros anteriores do vídeo, além de também utilizar as informações dos quadros seguintes do vídeo gravado previamente. Para o segundo classificador, o trabalho utiliza o mesmo tipo de rede LSTM bidirecional, porém as características faciais são extraídas do ajuste fino da rede convolucional ResNet. O terceiro classificador é construído utilizando uma rede do tipo C3D (TRAN et al., 2015), que são redes convolucionais tridimensionais. Este tipo de rede, além de modelar informações espaciais das redes convolucionais tradicionais, também possui a capacidade de realizar a modelagem temporal utilizando a terceira dimensão. Segundo os autores, a vantagem de sua utilização sobre as redes do tipo convolucional tradicionais encapsuladas em redes temporais recorrentes (como as LSTM) é que nas redes C3D as convoluções são realizadas simultaneamente com a passagem do tempo, em oposição ao modelo Conv+LSTM, em que as características temporais são extraídas das características espaciais estáticas. O quarto classificador do trabalho utiliza informações de áudio, portanto, não é relevante para o presente trabalho. Através da fusão dos quatro classificadores, o modelo obteve 60,64% de acurácia, e foi classificado como segundo colocado da EmotiW 2018. É importante ressaltar que o modelo base da competição obteve 41,07% de acurácia, e o trabalho de Mollahosseini, Chan e Mahoor (2016) obteve 47,7% na base SFEW, que é uma base que utiliza imagens extraídas dos vídeos da AFEW.

A utilização da fusão de classificadores é uma estratégia que também foi implementada no trabalho de Liu et al. (2018), vencedor da categoria *Audio-video Sub-challenge* da EmotiW 2018. Os autores criaram o modelo com quatro classificadores. Para o primeiro classificador, as *landmarks* presentes nas faces em cada quadro dos vídeos foram extraídas, e então foram calculadas as distâncias euclidianas entre 34 *landmarks*. Destas 34 distâncias foram calculadas a média, o máximo e a variância, resultando em 102 características faciais. Para a classificação deste método foi utilizada uma rede do tipo SVM, que obteve 39,95% de acurácia na base AFEW. Para o segundo classificador, foram capturadas as faces alinhadas com o algoritmo MTCNN e após foi realizado o treinamento de quatro tipos de rede convolucionais: Inception-v3 (SZEGEDY et al., 2016), DenseNet-121, DenseNet-161 e DenseNet-201 (HUANG et al., 2017). A rede Inception-v3, que é a terceira versão da rede GoogLeNet (SZEGEDY et al., 2015), implementa o conceito explicado no trabalho de Mollahosseini, Chan e Mahoor (2016). As redes DenseNet são uma evolução do conceito da ResNet (HE et al., 2016), onde cada camada obtém entrada de todas as camadas antecessoras e fornece para todas as camadas subsequentes o resultado de seu mapeamento de características. As diferentes versões da DenseNet, utilizadas no trabalho, indicam topologias distintas do mesmo tipo de rede. Para o treinamento destas redes, as bases FER2013 (Kaggle, 2013; GOODFELLOW et al., 2013) e RAF (LI; DENG; DU, 2017) foram utilizadas, e o ajuste fino foi realizado com dados da base AFEW. Com as características extraídas das quatro redes convolucionais, foi implementada uma rede SVM para realizar a classificação que obteve acurácia de 51,44%. O terceiro classificador foi implementado com

uma rede VGG para extração de características faciais espaciais e uma camada LSTM para as características temporais, obtendo 46,21% de acurácia. A rede VGG foi treinada na base de dados FER2013 e o ajuste fino foi realizado com a base AFEW. O quarto classificador utiliza informações de áudio. Cada um dos quatro classificadores realiza a predição de uma das emoções básicas, e os autores realizam a fusão dos classificadores aplicando pesos em cada uma das quatro saídas conforme o desempenho obtido no conjunto de validação da base, obtendo uma acurácia de 61,87% no total.

A Tabela 7 relaciona cada um dos trabalhos acima mencionados explicitando suas características principais e resultados obtidos.

Tabela 7: Trabalhos relacionados que classificam emoções básicas.

Trabalho	**Base de dados**	**Pré-proc.**	**Caract.**	**Classificador**	**Resultados**
(LI; LAM, 2015)	CK+	Viola-Jones	Gabor, kpca	4x FC 200	ACC: 91,7%
(MOLLAHOSSEINI; CHAN; MAHOOR, 2016)	MultiPie, MMI, CK+, DISFA, FERA, SFEW, FER2013	AAM, SDM	CNN	CNN	ACC: MultiPie: 94,7% MMI: 77,9% DISFA: 55,0% FERA: 76,7% SFEW: 47,7% CK+: 93,2% FER2013: 66,4%
(ZENG et al., 2018)	CK+	AAM	HOG, LBP	4 DSAE 100	ACC 95,79%
(KIM et al., 2017)	MMI, CAS-MEII	?	CNN	LSTM	ACC 78,61%
(LU et al., 2018)	AFEW	MTCNN	1: VGG 2: ResNet 3: C3D	1: BLSTM 2: BLSTM 3: C3D	ACC 60,64%
(LIU et al., 2018)	AFEW	MTCNN	1: LMED 2: 4xCNN 3: VGG	1: SVM 2: SVM 3: LSTM	ACC 61,87%

Fonte: Elaborado pelo autor.

5.3 Detecção automática de emoções acadêmicas por face

Os trabalhos descritos na Seção 5.2 realizam a classificação das características faciais em uma das emoções básicas ou classificam as Action Units (movimentos faciais básicos) das faces. Conforme explicado na Seção 2.2, D'Mello e 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. A Tabela 8 relaciona os trabalhos da área analisados, descrevendo suas características principais, métodos de detecção, e tipos de redes utilizadas.

Tabela 8: Trabalhos relacionados que classificam emoções não básicas presentes em situações de aprendizagem.

Trabalho	Emoção detectada	Tipo de classificação	Base de dados	Pré-proc.	Características	Classificação	Resultados
(GUPTA et al., 2016a)	Engajamento Confusão Frustração Tédio	Classificação categórica	DAiSEE	Viola-Jones (FaceAlign)	VGG HOG LBP-TOP	CONV + LSTM	ACC: Eng: 57,9% Ted: 53,7% Con: 72,3% Fru: 73,5%
(DEWAN et al., 2018)	Engajamento	Classificação categórica	DAiSEE	Viola-Jones	DBN	DBN 50 DBN 50 DBN 100	ACC (3lvl) 87,25% ACC (2lvl) 93,23%
(NEZAMI et al., 2018)	Engajamento	Classificação categórica; Classificação binária	FER2013, outro próprio	CNN, Dlib-ml	8 CONV 2 FC	Ajuste fino da rede de características	ACC 72,38%
(KAUR et al., 2018)	Engajamento	Regressão	EmotiW	OpenFace	LBP-TOP Mov. dos olhos P. da cabeça	LSTM 32 3 FC 128,128,100	MSE 0.10
(THOMAS; NAIR; JAYAGOPI, 2018)	Engajamento	Regressão	EmotiW	OpenFace	Mov. dos olhos P. da cabeça *Action Units*	3 TCN 24	MSE 0.0792
(YANG et al., 2018)	Engajamento	Regressão	EmotiW	MTCNN (OpenFace)	Mov. dos olhos P. da cabeça P. do corpo LBP-TOP C3D	2 LSTM 64; 3 Dense 1024, 512, 128	MSE 0.0626

Os trabalhos de Gupta et al. (2016a,b) 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 et al., 2016), treinamento de uma rede C3D (TRAN et al., 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 et al. (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 das redes, porém em nenhum momento mencionaram respeitar 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 et al. (2016a), 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 de 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 de Kaur et al. (2018); Thomas, Nair e 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 et al., 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 de Thomas, Nair e Jayagopi (2018), que obteve a segunda colocação. Ele utiliza uma rede convolucional unidimensional temporal Lea et al. (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.

5.4 Pesquisa atualizada sobre detecção de emoções acadêmicas

A segunda etapa da pesquisa de trabalhos relacionados foi realizada com dois objetivos principais: (*i*) Encontrar trabalhos que realizem a detecção de emoções acadêmicas utilizando vídeos da face em um período posterior ao período de realização da pesquisa inicial; (*ii*) Selecionar principalmente trabalhos que possam ser comparados com o trabalho desenvolvido. Para garantir este segundo objetivo o trabalho precisa se adequar em alguns critérios, descritos a seguir:

- Base de dados comum: A etapa de teste do trabalho precisar ter sido realizada sobre o conjunto de teste da base DAiSEE, sem que o conjunto de teste em questão tenha sido alterado com relação ao conjunto original.

- Classificador: O algoritmo desenvolvido no trabalho precisa ser um classificador. Algoritmos regressores não são comparáveis.

- Métricas compatíveis: O trabalho precisa reportar seus resultados utilizando F1 e/ou acurácia como métricas de desempenho.

- Classificação individual das emoções: Cada emoção precisa ter sido treinada individualmente. Trabalhos que agregam todos os rótulos da base de dados em uma única emoção não são comparáveis.

- Dois níveis da emoção: A sensibilidade do modelo precisa ser de dois níveis da emoção treinada, ou apresentar uma matriz de confusão que permita agrupar os resultados em dois níveis.

De acordo com estes critérios, cinco trabalhos foram selecionados e suas principais características podem ser observadas na Tabela 9. Um comparativo de desempenho com o trabalho desenvolvido pode ser observado na Seção 7.4.3.

Tabela 9: Trabalhos relacionados selecionados na pesquisa posterior à implementação dos modelos.

Trabalho	**Emoção**	**Modelo**	**Resultados**
(LEONG, 2020)	Tédio e Frustração	LSTM 50	**TED**: F1=0,696; ACC=58,78% **FRU**: F1=0,419; ACC=73,09%
(ABEDI; KHAN, 2021)	Engajamento	ResNet + TCN	F1=0,558; ACC=92,21%
(ZHANG et al., 2019)	Engajamento	I3D + LSTM	ACC=98,82%
(ZHENG et al., 2021)	Engajamento	CNN (?) + LSTM	F1=0,902; ACC=91,99%

Fonte: Elaborado pelo autor.

O trabalho de Leong (2020) usa a base de dados DAiSEE para realizar a detecção das emoções frustração e tédio. Os autores usaram os vídeos com sua duração original de 10 segundos, e cada emoção foi rotulada em dois níveis (presença e ausência da emoção). Seus modelos utilizam uma rede pré-treinada FaceNet, que é uma rede do tipo Inception-ResNet treinada originalmente na base VGGFace2. Eles também utilizam uma rede que calcula a distância euclideana entre os *landmarks* detectados das faces. As características extraídas foram dadas como entrada em uma camada do tipo LSTM, e os modelos foram treinados durante 50 *epochs*, obtendo $F1 = 0,696$ e acurácia 58,78% para a emoção tédio e $F1 = 0,419$ e acurácia 73,09% para a emoção frustração.

O trabalho de Abedi e Khan (2021) implementa modelos de redes convolucionais com camadas recorrentes como estratégia para realizar o reconhecimento do engajamento em quatro níveis utilizando a base de dados DAiSEE. Os autores utilizam modelos dos tipos C3D e ResNet para realizar a extração das características espaciais dos quadros dos vídeos, e camadas LSTM e TCN para contabilizar os vídeos como conjunto de imagens. Para contrabalancear o desbalanceamento da base de dados, os autores utilizaram pesos no treinamento inversamente proporcionais à quantidade de exemplos em cada classe, bem como redistribuíram os exemplos de treinamento de maneira que todos os *batches* possuíssem a mesma proporção de exemplos de cada classe. Embora o trabalho se propôs a realizar a classificação em quatro níveis para a emoção engajamento, os autores apresentaram a matriz de confusão de seus modelos, o que permitiu agregar seus resultados e calcular as métricas F1 e acurácia para dois níveis, a fim de comparar com o presente trabalho. Para dois níveis, os autores obtiveram $F1 = 0,558$ e acurácia 92,21%.

O trabalho de Zhang et al. (2019) realiza o reconhecimento da emoção engajamento em qua-

tro e dois níveis utilizando a base de dados DAiSEE. O modelo criado pelos autores utiliza uma rede do tipo I3D (CARREIRA; ZISSERMAN, 2017), onde eles realizam o ajuste fino da mesma concatenada a camadas do tipo LSTM para a extração das características espaço-temporais dos vídeos. Durante o treinamento, eles adicionaram pesos inversamente proporcionais à representatividade das classes. Eles obtiveram 98,82% de acurácia para dois níveis de engajamento.

O trabalho de Zheng et al. (2021) se propõe a realizar a detecção do engajamento nos estudantes utilizando duas bases de dados distintas: DAiSEE e uma base própria, menor. Eles realizam a extração das características faciais com o OpenFace. Os autores não relatam com detalhes como usam as características nem descrevem a arquitetura de rede convolucional que usam na construção de seu modelo. Eles utilizam camadas LSTM para extrair as relações temporais dos vídeos. Eles realizam o treinamento do modelo na base DAiSEE, depois aplicam um método de *transfer learning* para realizar o ajuste fino da rede treinada na base própria. Os autores unificam os exemplos das classes de rótulo 0 e 1 em seus experimentos, porém, como reportam a matriz de confusão, pôde-se obter $F1 = 0.902$ e acurácia 91,99% para dois níveis. Embora os resultados sejam excelentes, em seu texto, muita informação é desconhecida a respeito da implementação do modelo, bem como alguns existem alguns erros metodológicos a respeito da implementação do conceito de *transfer learning*.

5.5 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). É importante ressaltar que estas quatro emoções estão intimamente ligadas no contexto de aprendizagem, conforme visto na Seção 2.2.

O trabalho realizado possui como principal diferencial o modelo de reconhecimento de emoções, que realiza a detecção do engajamento, confusão, frustração e tédio em ambientes de aprendizagem. Acredita-se que 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, conforme demonstrado no trabalho de D'Mello e Graesser (2012), bem como a relação do tempo de permanência de um

estudante nestas emoções com a sua personalidade, conforme descrito no trabalho de Reis et al. (2018), é justificável a investigação sobre a manifestação destas emoções em uma escala temporal, considerando a personalidade do aprendiz. Dessa foram, outro diferencial do trabalho envolve o uso das informações de transição destas emoções ao longo do tempo e da personalidade do estudante para melhorar a capacidade de detecção e previsão do modelo construído.

A Tabela 10 demonstra o comparativo entre as principais características dos trabalhos relacionados e o do trabalho realizado, evidenciando suas diferenças e ressaltando a justificativa da presente tese. Na coluna Emoções analisadas, é listado o escopo das emoções que o trabalho se propõe a detectar. Em informação temporal (vídeo), é dito se o trabalho considera imagens individuais, ou sequências de imagens (vídeos) para realizar a detecção. Na coluna Informação temporal (sequência de emoções) é descrito se o trabalho realiza a detecção considerando o histórico de emoções já presenciadas até o momento. Em Características da rede convolucional / rede auxiliar, é relatado se o modelo do trabalho extrai ou não características de uma rede convolucional, bem como quantas características auxiliares o modelo utiliza no treinamento. Em bases de dados utilizadas, são citadas as base de dados utilizadas para o treinamento dos modelos dos respectivos trabalhos.

[4]O trabalho utiliza as outras emoções presentes na base de dados somente para estimar o engajamento.

[5]O trabalho não utiliza uma rede convolucional. Ele treina seu modelo utilizando informações temporais extraídas de características arquitetadas.

Tabela 10: Comparativo entre as principais características dos trabalhos relacionados que tratam de emoções presentes no aprendizado e a presente proposta.

Trabalho	Emoções analisadas	Inf. temporal (vídeo)	Inf. temporal (seq. de emoções)	Características rede Conv. / rede auxiliar	Bases de dados utilizadas
(GUPTA et al., 2016a)	Engajamento, Confusão, Frustração, Tédio	Sim	Não	Sim / -	DAiSEE
(DEWAN et al., 2018)	Engajamento[4]	Não	Não	Sim / -	DAiSEE
(NEZAMI et al., 2018)	Engajamento	Não	Não	Sim / -	FER-2013, base própria
(KAUR et al., 2018)	Engajamento	Sim	Não	Não[5] / 3	EmotiW
(THOMAS; NAIR; JAYAGOPI, 2018)	Engajamento	Sim	Não	Não[5] / 3	EmotiW
(YANG et al., 2018)	Engajamento	Sim	Não	Sim / 4	EmotiW
(LEONG, 2020)	Tédio e Frustração	Sim	Não	Sim / 2	DAiSEE
(ABEDI; KHAN, 2021)	Engajamento	Sim	Não	Sim / -	DAiSEE
(ZHANG et al., 2019)	Engajamento	Sim	Não	Sim / -	DAiSEE
(ZHENG et al., 2021)	Engajamento	Sim	Não	Sim / 3	DAiSEE, base própria
Modelo desenvolvido	Engajamento, Confusão, Frustração, Tédio	Sim	Sim	Sim / 3	DAiSEE, PAT2Math

6 MÉTODO

O presente trabalho realiza 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 et al., 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. Foi optado pela utilização somente de informações de vídeos, ao invés de outras modalidades como voz, sinais biométricos ou texto por dois principais motivos. O principal motivo é a facilidade para utilização do método, um sistema tutor que pode facilmente integrar um algoritmo que executa o método proposto em computadores portáteis, uma vez que todos possuem câmeras. O segundo motivo é a falta de disponibilidade de informações multimodais para treinamento de modelos, pois nenhuma base de dados de emoções de aprendizagem, até o presente momento, apresenta informações além dos vídeos da face dos alunos. Tendo isso em vista, 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 apresentadas. Este classificador realiza seu trabalho utilizando-se de 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.

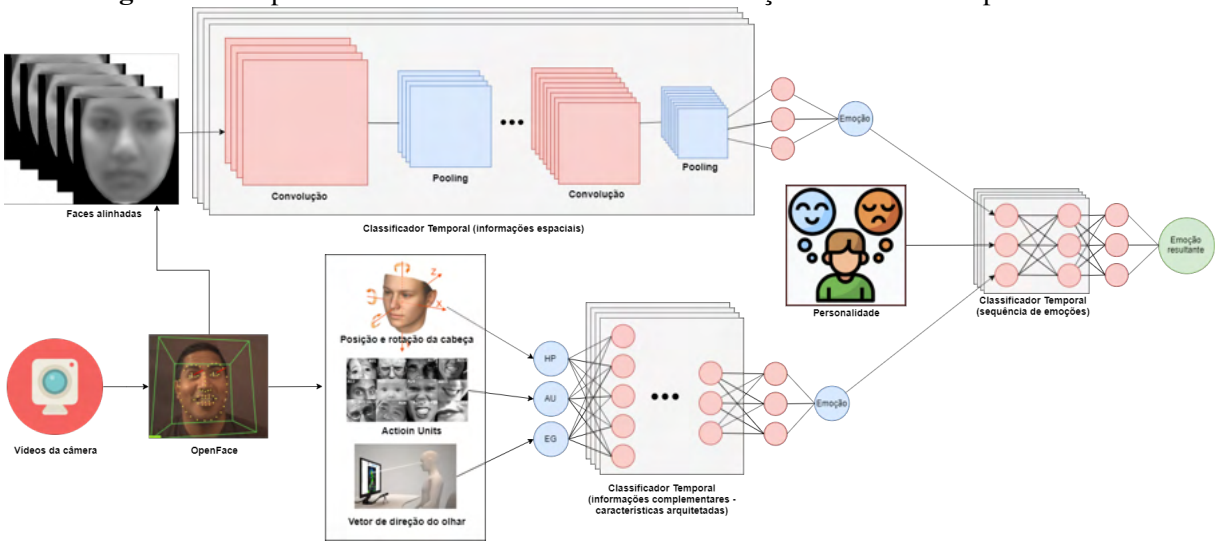
A rede de características extraídas dos vídeos (i) recebe como entrada os píxeis dos quadros dos vídeos contidos nas bases de dados e, por camadas convolucionais e recorrentes, obtém como saída durante o processo de treinamento características abstratasutilizadas pelo classificador para determinar se 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 17 demonstra uma visão geral do modelo desenvolvido. Um detalhamento maior

de cada um dos elementos que o compõe podem ser observados na Seção 6.3.

Figura 17: Etapas do módulo de reconhecimento de emoções do modelo implementado.



Fonte: Elaborado pelo autor.

6.1 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) (descrita na Seção 5.2). Outros, na tentativa de melhorar a generalidade de seus modelos, propuseram bases em ambientes não controlados, como a SFEW (DHALL et al., 2011a), composta de imagens estáticas de vídeos da base AFEW (DHALL et al., 2011b), 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. (2011b) 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. A Tabela 6 relaciona as principais bases de dados de emoções básicas e *Action Units*. Como o presente trabalho realiza 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, o trabalho usa 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 redes três bases de dados encontradas. A primeira, chamada *Dataset for Affective States in E-Environments* - DAiSEE (GUPTA et al., 2016a), 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* de 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. A Figura 18 demonstra a captura de quadros de alguns vídeos que compõe a base de dados. 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 resolvendo equações de primeiro grau passo a passo (explicado na Seção 3.2). O trabalho de Morais et al. (2019), por exemplo, realizou a gravação de vídeos e anotação das emoções de 34 alunos durante a interação de uso do sistema

Figura 18: Captura de alguns quadros de vídeos presentes na base de dados EmotiW Kaur et al. (2018)

Fonte: Kaur et al. (2018)

PAT2Math (JAQUES et al., 2013). Os dados foram coletados de alunos de uma escola particular da cidade de São Leopoldo, Rio Grande do Sul. Alunos de duas turmas da sétima série usaram o PAT2Math durante algumas de suas aulas de matemática. O professor era o mesmo para as duas turmas e estava sempre presente enquanto os alunos utilizavam o sistema. O tempo gasto pelos alunos no uso do ITS (*Intelligent Tutor System*) foi o mesmo para ambas as turmas. No total, participaram da coleta de dados 55 alunos (29 meninas e 26 meninos), com idades entre 12 e 13 anos. Os alunos usaram o ITS por 10 sessões uma vez por semana, entre maio e outubro de 2018. Cada sessão teve duração aproximada de 40 minutos e ocorreram no turno matinal. Os alunos usaram o ITS por aproximadamente 360 horas. Isso resultou na geração de 21 horas de vídeo de gravação dos alunos usando o ITS PAT2Math. Destas, aproximadamente 2,5 horas divididas em 30 vídeos de cinco minutos cada possuem anotações das emoções, e puderam ser

utilizadas no treinamento. Além disso, durante esta coleta de dados, também aplicamos aos alunos o questionário de personalidade, para identificar a personalidade, usada nesse trabalho. Foi disponibilizado aos alunos e seus pais um Termo de Consentimento Livre e Esclarecido para informar os alunos e seus pais sobre os procedimentos de coleta de dados, assim como o tratamento dos dados. Como os alunos eram menores de idade, foi solicitado que os pais dos alunos assinassem o termo. Embora a participação era opcional, no presente estudo, todos os alunos participantes entregaram o termo de consentimento livre e esclarecido assinado por pelo menos um responsável.

6.2 Pré-processamento

O pré-processamento é uma importante fase do treinamento de algoritmos de aprendizado de máquina. Nesta fase, as informações que serão usadas como entrada dos modelos devem ser tratadas para que o modelo posso aprender adequadamente com as mesmas. 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 em meio às tabelas de dados de entrada, adequação de dimensões das informações, dentre outros.

6.2.1 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 algumas razões principais:

- Vídeos de mais curta duração possuem menos requisito computacional para realização do treinamento.
- De acordo com o protocolo de anotações EmAP-ML (MORAIS et al., 2019), durante o período de cinco segundos os estudantes chegam a experienciar de uma a duas emoções diferentes. Por este fato, a escolha da granularidade de dois segundos também visa capturar estes momentos isoladamente.
- 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 um tamanho determinado para serem utilizados nos modelos implementados. Dos modelos de primeira e segunda geração este tamanho foi de 112x112 píxeis, e para os modelos de terceira e quarta geração foram utilizados 224x224 píxeis. A decisão pelo tamanho das imagens é baseada no tamanho mínimo necessário para realizar o ajuste fino de uma rede da topologia VGG (112x112 píxeis) e ResNet (HE et al., 2016) (224x224 píxeis). Todos as faces alinhadas foram extraídas em escala de cinza, o que fez com que todos os três canais de cor tivessem a mesma informação. O OpenFace utiliza o algoritmo de conversão para escala de cinza do OpenCV, que calcula cada píxel resultante segundo a soma ponderada de todos os três canais de cor, conforme a Equação 6.1.

$$Grayscale = 0.2989 \cdot Red + 0.5870 \cdot Green + 0.1140 \cdot Blue \tag{6.1}$$

Onde Red, $Green$ e $Blue$ são os valores dos canais vermelho, verde e azul, respectivamente, da imagem.

Conforme visto na Seção 6.1, 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 11 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.

Tabela 11: 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

Fonte: Elaborado pelo autor.

Em resumo, os vídeos passaram pelo seguinte tratamento até a obtenção das faces alinhadas:

- Isolamento dos quadros dos vídeos, utilização de metade destes quadros intercaladamente (redução de 30 para 15 quadros por segundo).
- Detecção dos *face landmarks*. A partir dos mesmos, recorte da área que possui o resto detectado.
- Remoção do fundo da imagem (área externa aos *landmarks*).
- Redimensionamento da imagem para o tamanho 224x224 ou 112x112 píxeis.
- Conversão dos três canais de cor da imagem em escala de cinza, de acordo com a Equação 6.1.

6.2.2 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 classe 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 *under-*

sampling, 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, conforme Elrahman e Abraham (2013), este tipo de abordagem frequentemente conduz ao *overfitting* da rede (explicados na Seção 4.1.2), pois a mesma aprende a identificar os padrões específicos dos exemplos fornecidos. 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. De acordo com Huang et al. (2013), 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. Em seu trabalho, 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.

O presente trabalho utilizou todas as três técnicas em seus modelos, embora nem todas as técnicas foram usadas neles 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 (processo explicado na Seção 6.2), 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 et al., 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 efetivo e dinâmico, pois quando o número de exemplos de treinamento variava, o peso automaticamente era recalculado. O cálculo da proporcionalidade é feito através de ponderação simples, conforme a Equação 6.2 abaixo:

$$w_i = \frac{\sum_{j=1}^{c} n_j}{n_i} \tag{6.2}$$

Onde w_i é o peso da classe i, n_i é a quantidade de exemplos de treinamento para a classe i, e c é o número total de classes.

6.2.3 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.

O modelo desenvolvido utiliza 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. Podem ser representadas como um valor binário indicando presença ou ausência de cada uma das AUs, ou valores contínuos que representam a intensidade do movimento muscular relacionado à AU.

- *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, em radianos, e a translação indica a posição no espaço, expresso em milímetros, onde $X = Y = 0$ é o centro da câmera e Z é a distância da câmera.

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 6.3. Estas características possuem relação com informações abstratas extraídas diretamente dos píxeis que compõe as imagens, e da transição destas imagens que formam os vídeos.

6.3 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**. Foi optado pelo tratamento do problema como classificação, ao invés de regressão, pelos seguintes motivos:

- A principal base de dados utilizadas DAiSEE apesar de reportar em seus rótulos as emoções em intensidades entre 1 e 4, o faz através de números discretos. Os autores implementam seus modelos, usados por diversos outros trabalhos como *benchmark*, como um problema de classificação.
- A segunda base utilizada, PAT2Math, responsável pela agregação de informações de personalidade dos estudantes nos modelos desenvolvidos, reporta as emoções em seus rótulos como presença ou ausência de uma ou mais emoções em cada trecho dos vídeos.
- A facilitação da comparação com outros trabalhos relacionados, que tenham utilizado a mesma base de dados na implementação de seus modelos.

Este reconhecimento se deu através de uma fusão de redes neurais implementadas utilizando algoritmos de aprendizado supervisionado de máquina. Conforme visto na Seção 4.1.2, 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, conforme demonstrado no trabalho de Li e Deng (2018), é que normalmente para tarefas de classificação de emoções em faces, os trabalhos que costumam obter os melhores desempenhos empregam *Deep Learning*. 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 to 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.

6.3.1 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 é uma característica comumente associada ao aprendizado, como mostra o trabalho de Lai et al. (2013), enquanto o modelo FACS (EKMAN, 1992) 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, e consiste em uma topologia específica de rede neural (detalhados na seção 6.2.3) 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 identificar 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.

6.3.2 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, conforme foi descrito na Seção 4.1. A ideia por trás de uma rede convolucional, ou de características espaciais, é que cada camada convolucional, através dos filtros (conforme demonstrado na Figura 9), 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, conforme visto na Seção 4.1.2. 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 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, o presente trabalho se utilizou de 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 (HUANG et al., 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 et al., 2017), Inception (SZEGEDY et al., 2016), ResNet e Vgg. A Tabela 12 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 7.1, 7.2, 7.3 e 7.4.

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 emo-

Tabela 12: Comparativo entre as redes convolucionais utilizadas em cada geração de modelos desenvolvidos. Todos estes modelos podem ser acessados em Werlang (2022).

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 / personalidade incorporado com rede TCN.

Fonte: Elaborado pelo autor.

ções futuras, os modelos mais avançados utilizaram 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.

Ainda no tema da temporalidade das emoções, a personalidade do indivíduo foi uma das características utilizadas para melhorar a compreensão do modelo a respeito das emoções no aprendizado. Durante a última fase das implementações, a personalidade foi uma das características utilizadas para melhorar a detecção das emoções nos vídeos. Porém, como só foi possível obter os rótulos de emoções dos vídeos da base de dados PAT2Math, o modelo que utilizou a personalidade só pode ser comparado com outros modelos treinados usando somente o PAT2Math como base de treinamento.

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 desenvolvimentos dos modelos foram todas realizadas de acordo com os resultados obtidos 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 et al. (2016a,b), por exemplo.

6.3.3 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 e características de personalidade. Desta forma, neste estágio, o treinamento se dá pela concatenação de quatro conjuntos de características não relacionadas.

6.4 Treinamento

Para a realização do treinamento, foi utilizada a plataforma *Google Colaboratory*[1]. A máquina contratada varia conforme a disponibilidade, e o plano contratado foi o *Colab Pro+*, o nível que garante a máquina com a maior quantidade de recursos disponível. Por questão de demanda de recursos, o treinamento dos modelos desenvolvidos utilizando o plano gratuito do Colab, ou uma máquina local é inviável.

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 to-

[1] https://colab.research.google.com/

das 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 proporção similar de cada classe, pois como a distribuição 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.

Todos modelos desenvolvidos utilizaram método de classificação binária para uma única emoção, onde os rótulos representavam a presença ou ausência da emoção treinada. Durante o decorrer deste trabalho, os modelos desenvolvidos foram treinados para reconhecer a emoção engajamento. Após os últimos modelos terem sido desenvolvidos gerando resultados satisfatórios para o reconhecimento do engajamento, réplicas dos mesmos foram criadas e treinadas com os rótulos de cada uma das outras emoções de aprendizagem (confusão, frustração e tédio). Foi utilizada esta abordagem por dois principais motivos: (*i*) Modelos intermediários foram treinados para reconhecimento das quatro emoções simultaneamente e os mesmos apresentaram maior dificuldade em convergir a curva de perda durante o treinamento. (*ii*) Na grande maioria dos trabalhos relacionados os modelos foram treinados somente para o reconhecimento do engajamento, portanto treinamentos individuais foram convenientes para realizar comparativos posteriores.

Os rótulos presentes nas bases de dados foram convertidos de maneira a demonstrar este tipo de representação. Nos rótulos da base PAT2Math, o rótulo foi considerado presença quando a emoção relatada era a mesma emoção que o modelo estava tentando classificar ou ausência em caso contrário. Nos rótulos da base DAiSEE, as intensidades da emoção 0 e 1 foram consideradas ausência da emoção, enquanto as intensidades 2 e 3 foram consideradas presença da emoção.

Por se tratar de um classificador binário, todos os modelos utilizaram função de perda *categorical crossentropy* e função de ativação *softmax*. O otimizador predominantemente utilizado foi o *Adam* com *learning rate* (LR) 10^{-4}, embora alguns modelos utilizaram *SGD*. A métrica utilizada durante o treinamento foi a acurácia.

O treinamento foi realizado utilizando redução de LR (*ReduceLROnPlateau*) em um fator de 0.2 a cada cinco *epochs* onde a função de perda não apresentasse evolução. O treinamento foi interrompido nos modelos (*EarlyStopping*) quando a função de perda não apresentava evolução durante 10 *epochs*.

7 RESULTADOS

Foram desenvolvidos diversos modelos, explicados na Seção 6.3, 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 e personalidade nos modelos.

Neste capítulo, são descritos as arquiteturas e características específicas dos modelos desenvolvidos, assim como os resultados obtidos.

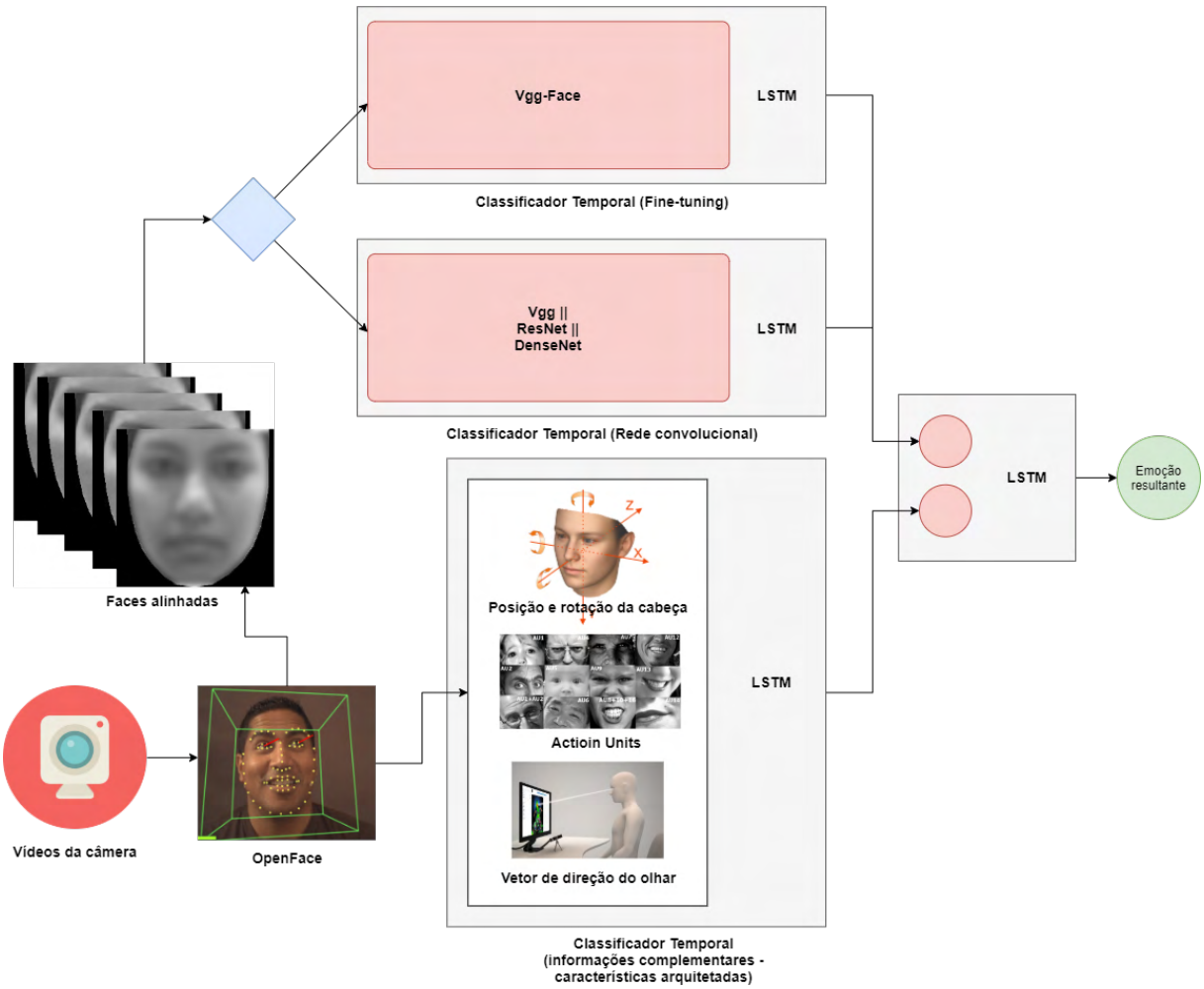
7.1 Primeira geração de modelos

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 19 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 primeira geração utilizaram uma janela de tempo de dez segundos para as camadas temporais (Conv3D e GRU). A escolha deste valor foi determinado durante testes de diferentes topologias de redes. Como, em geral, redes de primeira geração não apresentaram resultados promissores, o desenvolvimento de redes mais complexas (de segunda geração) foi preferível ao invés de realizar ajustes finos na largura das camadas temporais destes modelos.

Os modelos de características complementares de primeira geração usam redes recorrentes do tipo LSTM para extrair padrões temporais das características complementares das imagens.

Figura 19: Arquitetura dos modelos de primeira geração: treinamento de redes convolucionais temporais e rede temporal de características arquitetadas. Fusão usando saídas dos modelos escolhidos.



Fonte: Elaborado pelo autor.

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.

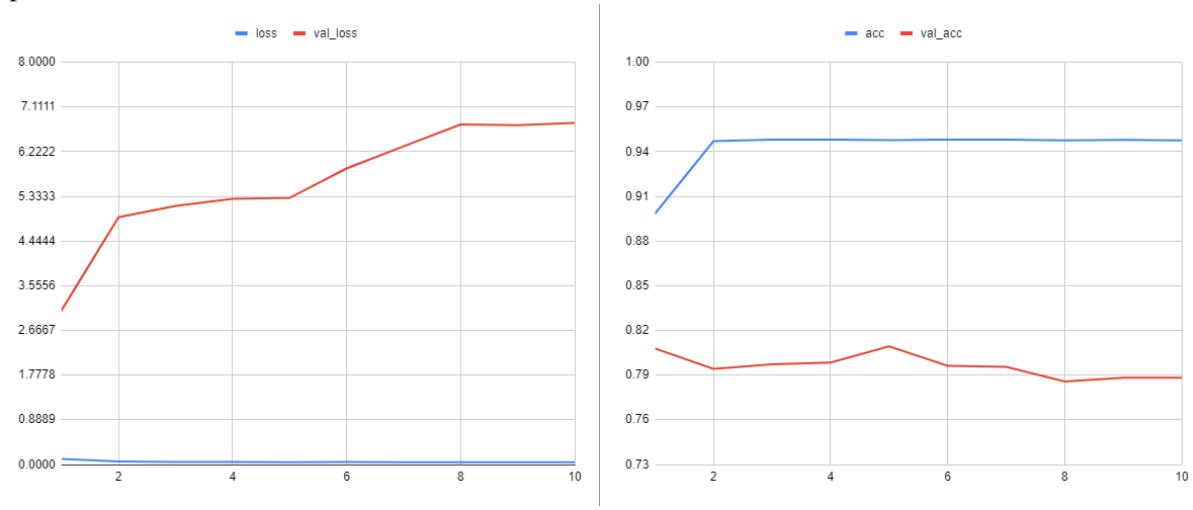
Na primeira etapa, a principal métrica utilizada para determinação do melhor modelo foi a acurácia, que é a métrica presente na maioria dos trabalhos correlatos analisados.

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 in-

divíduo 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.

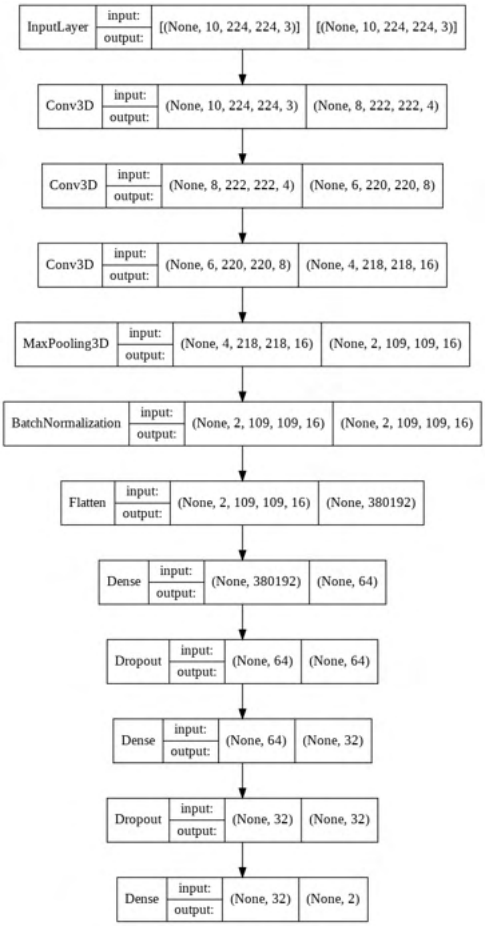
Dos modelos de primeira geração, poucos obtiveram resultados apresentáveis, pois na maioria dos modelos, as curvas de perda e acurácia não convergiam. Destaque pode ser dado para um modelo que realiza o ajuste fino de uma rede pré-treinada na base VGG-Face, e utiliza a arquitetura da rede VGG16. Este modelo utilizou os pesos da rede pré-treinada e ligou a rede VGG-Face uma camada *TimeDistributed* para tratar os vídeos. A saída desta camada foi ligada em uma camada do tipo GRU, que deu sequência à dois pares de camadas *Dropout* e *Dense* (esta última parte responsável pela classificação). O modelo obteve 63,62% de acurácia. Outro modelo implementado nesta etapa utilizou três camadas do tipo Conv3D (TRAN et al., 2015) sequenciais, seguidas de *MaxPooling*, *BatchNormalization* e *Flatten*. Esta última achata as características extraídas da rede em um vetor unidimensional. A partir daí, três pares de camadas *Dropout* e *Dense* foram utilizados para realizar a classificação. A Figura 21 demonstra as camadas deste modelo. Este tipo de rede realiza convoluções em entradas de dados de três dimensões ao invés de duas, como as camadas convolucionais tradicionais. No presente caso, 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. A Figura 20 demonstra um exemplo de uma rede desenvolvida que não demonstra melhora nas suas curvas de acurácia e perda (*loss*) ao longo das épocas de treinamento.

Figura 20: Curvas de *loss* e acurácia em um modelo típico da primeira e segunda geração: falha no aprendizado.



Fonte: Elaborado pelo autor.

Figura 21: Camadas de modelo Conv3D de primeira geração.



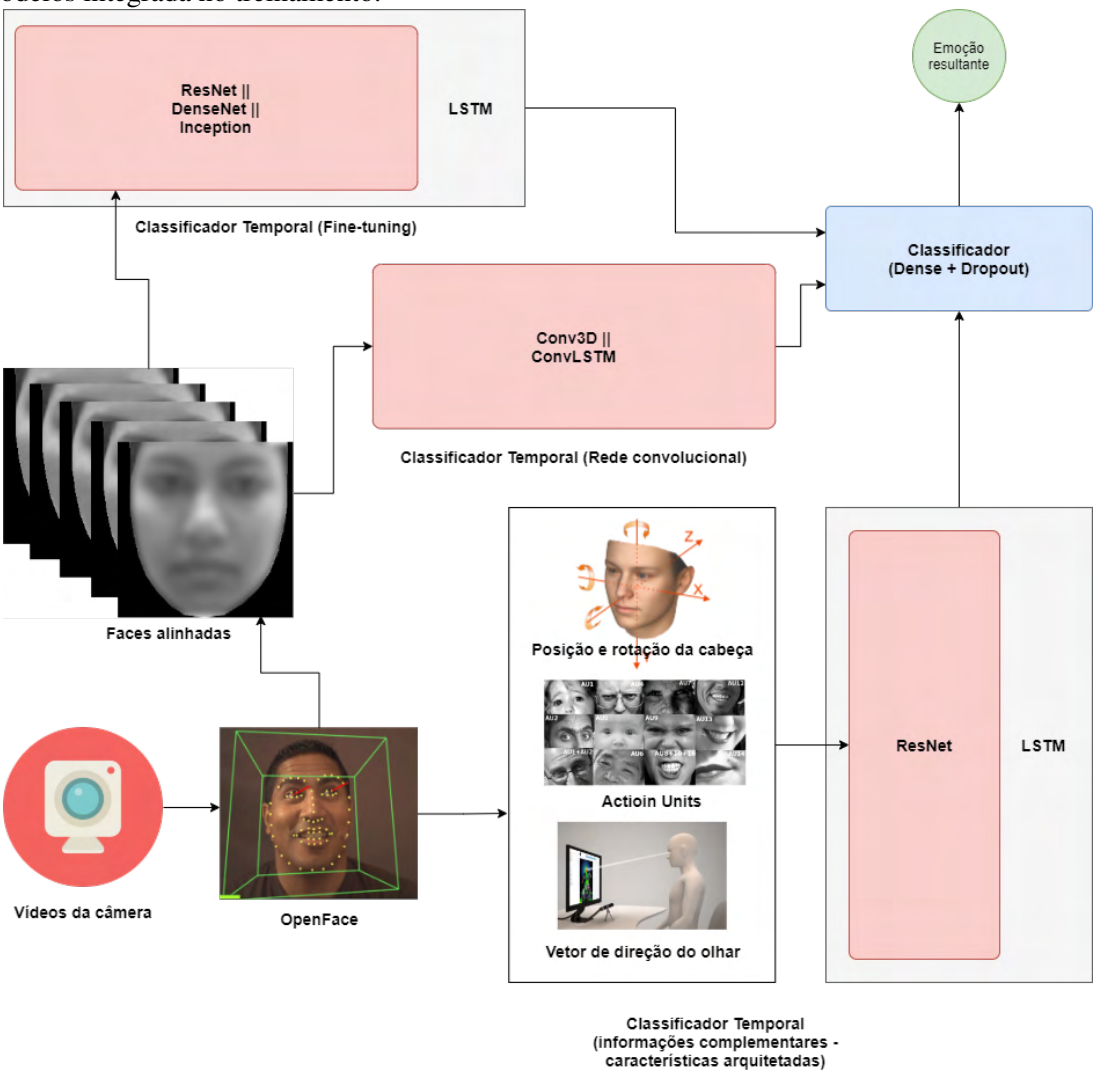
Fonte: Elaborado pelo autor.

7.2 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 22.

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

Figura 22: Arquitetura dos modelos de segunda geração: redes inspiradas em arquitetura ResNet e fusão de modelos integrada no treinamento.

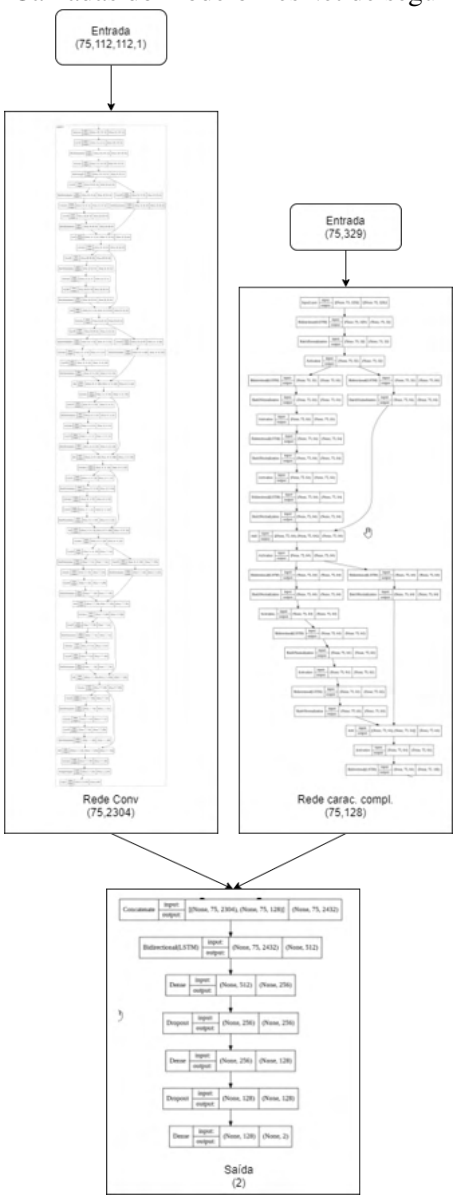


Fonte: Elaborado pelo autor.

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 Figura 23 demonstra as camadas de uma rede de segunda geração, evidenciando o tamanho das entradas e saídas de cada um dos modelos, enquanto a Figura 24 demonstra a arquitetura geral do modelo.

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

Figura 23: Camadas do modelo ResNet de segunda geração.

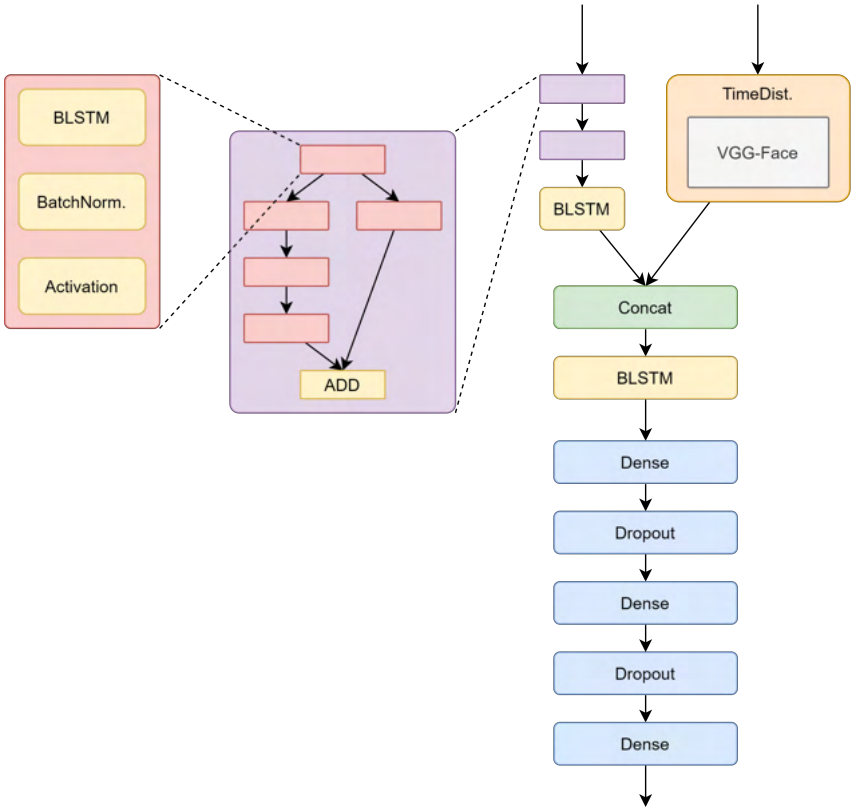


Fonte: Elaborado pelo autor.

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. Foi utilizada uma janela de 50 quadros nas camadas temporais desta rede, na expectativa que ao empregar um maior quantidades de dados temporais, e e consequentemente o aumento da demanda computacional, a rede demonstrasse melhores resultados que as redes de primeira geração.

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

Figura 24: Modelo com ajuste fino da rede VGG-Face .



Fonte: Elaborado pelo autor.

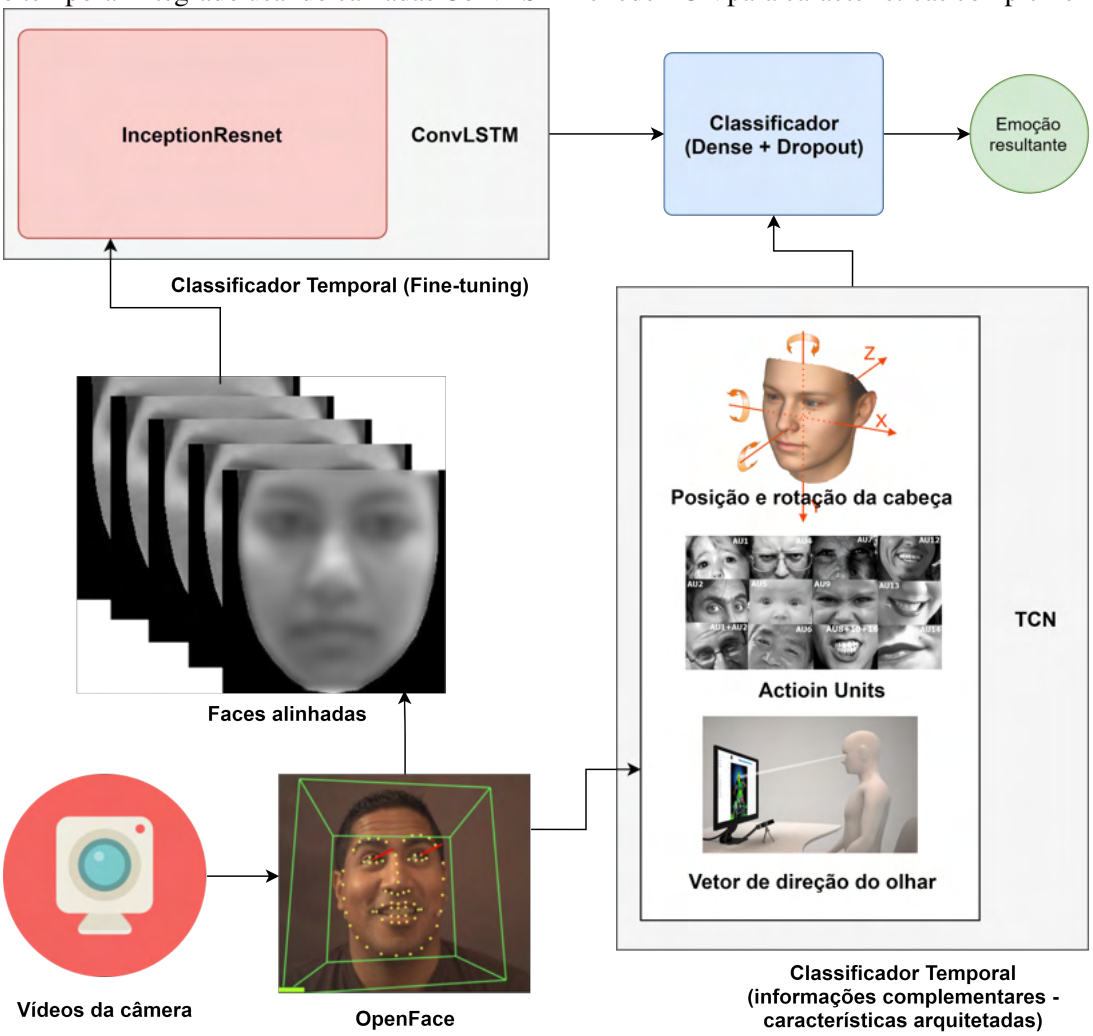
OpenFace, e obteve 85,03% de acurácia.

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. Utilizando uma janela de 75 quadros dos vídeos, 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.

7.3 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, conforme explicado na Seção 4.1.1. Para todos modelos implementados que utilizam a métrica F1, foi utilizado o método *f1_score* da biblioteca *sklearn* do *Python*. A Figura 25 demonstra uma visão geral dos modelos de terceira geração.

Figura 25: Arquitetura dos modelos de terceira geração: ajuste fino de redes mais complexas, treinamento temporal integrado usando camadas ConvLSTM e rede TCN para características complementares.



Fonte: Elaborado pelo autor.

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 (SZEGEDY et al., 2017), escolhida por conta do seu desempenhos nos testes de *benchmark*

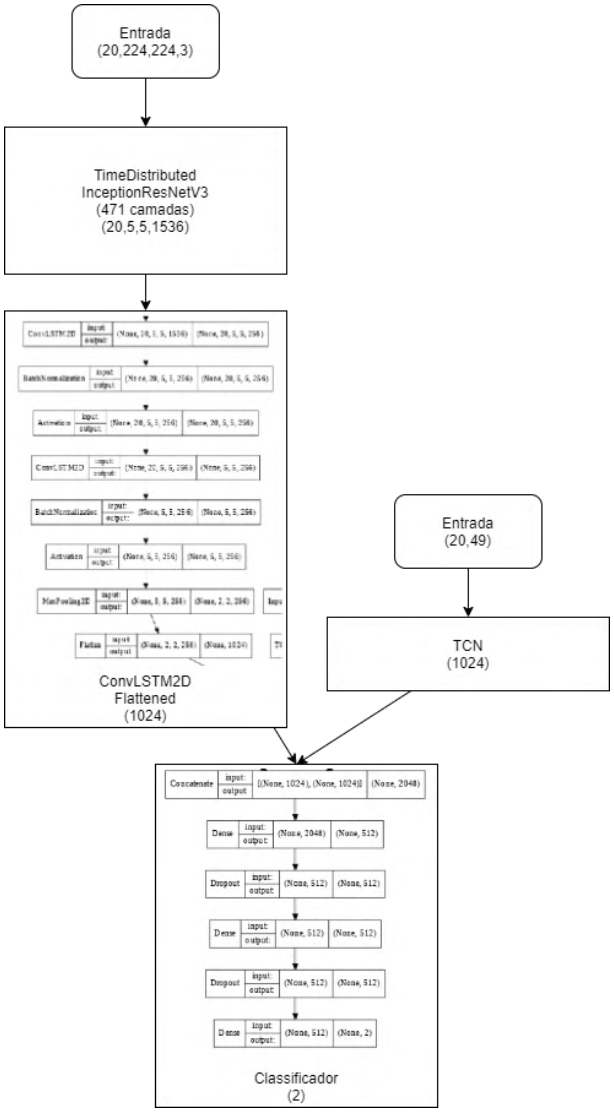
realizados (BIANCO et al., 2018). Para o tratamento dos vídeos, foram escolhidas camadas ConvLSTM2D (HU et al., 2020), que seguem uma lógica semelhante à rede Conv3D, com a diferença que na rede ConvLSTM2D a descoberta das características temporais é realizada usando a mesma estratégia de uma rede recorrente do tipo LSTM (com memória e esquecimento de informação), ao invés de por convoluções, como na camada Conv3D. Nas redes desta geração em diante optou-se pela redução para 20 os quadros analisados pelas camadas temporais. Esta decisão foi tomada porque ao decidir treinar uma rede do tipo InceptionResNet em sequência à camadas ConvLSTM, a quantidade de poder computacional necessário para o treinamento aumentou muito em comparação aos modelos treinados até então. Este aumento da demanda fez necessário o uso de máquinas com processadores gráficos mais avançados que os utilizados anteriormente (plano PRO+ do Google Colab). Sendo assim, 20 quadros temporais foi a quantidade que a máquina conseguiu trabalhar sem que esgotasse a memória disponível (32GB).

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. A Figura 26 demonstra um resumo das partes que compões o melhor modelo de terceira geração implementado evidenciando a largura das camadas de saída de cada ponto-chave do modelo, enquanto a Figura 27 apresenta uma descrição mais detalhada da composição de cada tipo de camada do modelo.

Conforme observado na Figura 25, 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 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 13 mostra a relação de desempenho entre os treinamentos dos mesmos modelos nas diferentes bases de dados.

Figura 26: Dimensões de saída de cada etapa das redes que compõe o melhor modelo de terceira geração.



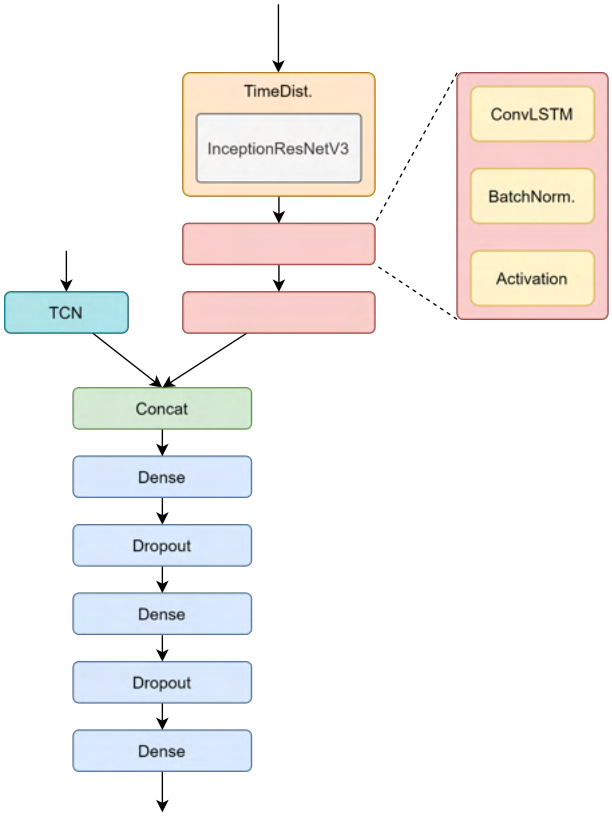
Fonte: Elaborado pelo autor.

7.4 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 7.3 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 e personalidade. 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,

Figura 27: Arquitetura do melhor modelo de terceira geração.



Fonte: Elaborado pelo autor.

Tabela 13: Desempenho dos modelos de terceira geração utilizando diferentes conjuntos de dados para treinamento e teste

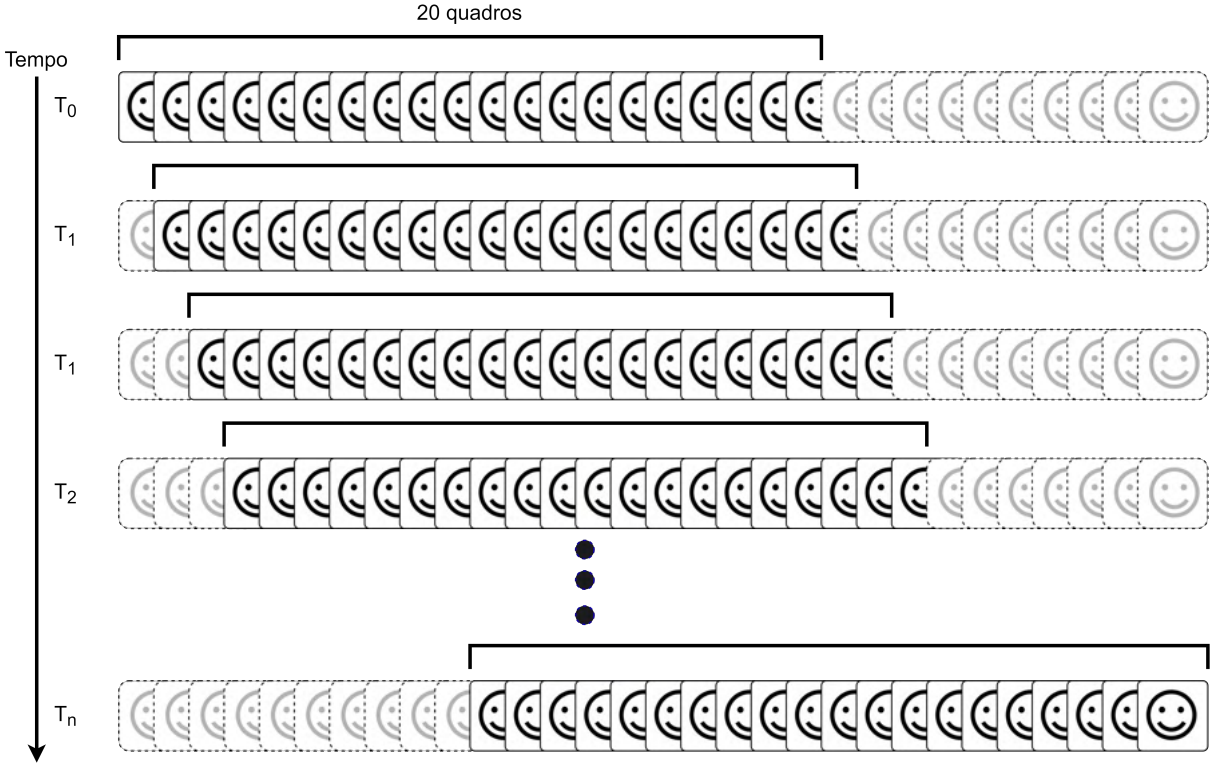
Base de dados	**Acurácia**	**F1**
DAiSEE	0,9417	0,5122
PAT2Math	0,5972	0,5859
DaiSEE + PAT2Math	0,9573	0,6882

Fonte: Elaborado pelo autor.

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. Esta rede recebeu como entrada a presença ou ausência da emoção sendo treinada pelo modelo durante uma janela de 20 quadros, a mesma utilizada na característica temporal das outras redes, isto é, as camadas ConvLSTM da rede convolucional e a camada TCN de características complementares. A Figura **??** demonstra como a rede foi treinada para utilizar o histórico de emoções do estudante funciona. A saída desta rede foi concatenada às saídas das outras redes e integradas ao classificador, conforme mostra a Figura 29. 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 7.3. Este resultado demonstra a vantagem na utilização de informações históricas na detecção de emoções.

Figura 28: Treinamento das redes que usam histórico de emoções: Os rótulos das emoções experienciadas no passado são usadas como entrada do modelo.



Fonte: Elaborado pelo autor.

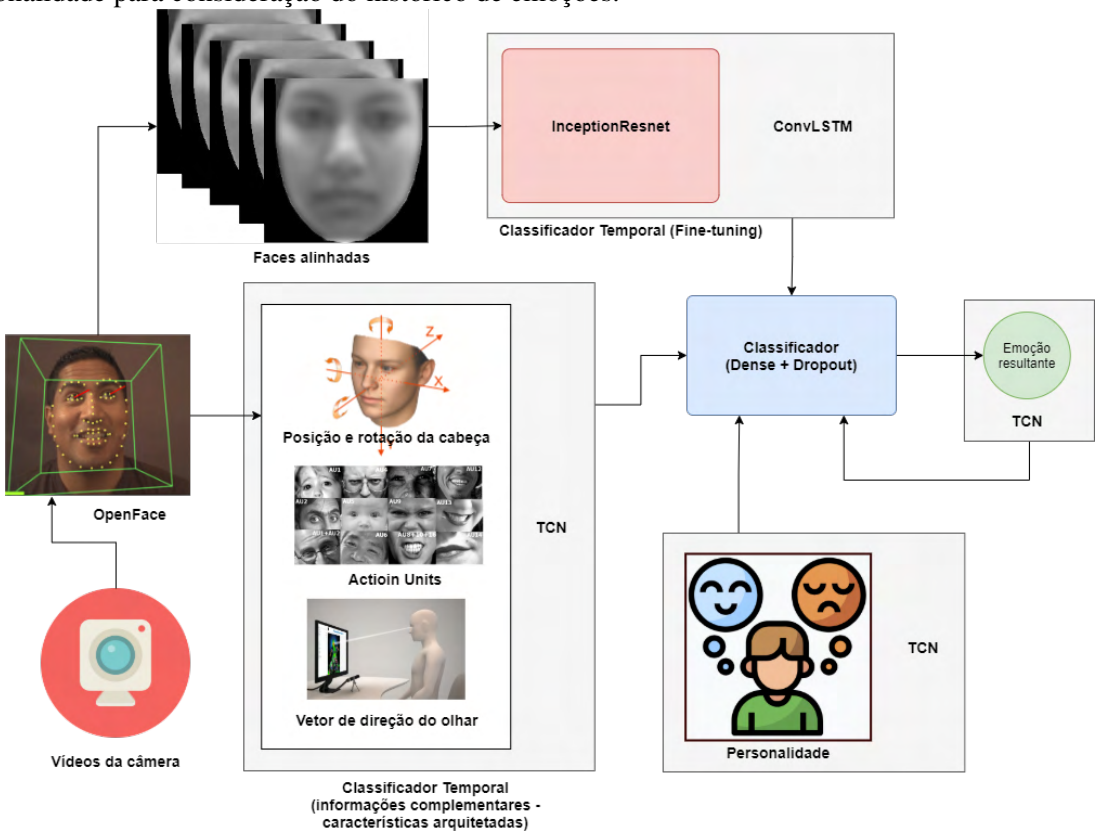
Cada subsequente geração de modelos trouxe uma nova melhoria comparativamente aos modelos de gerações anteriores. A Tabela 14 demonstra o desempenho obtida pelo melhor modelo em cada geração com a característica específica que o diferencia dos demais. A partir desta comparação, é possível observar a evolução dos modelos ao longo de seu desenvolvimento, que, com as figuras que os descrevem (Figura 19, Figura 22, Figura 25, e Figura 29), demonstram seu funcionamento, nível de complexidade e desempenho.

Tabela 14: Desempenho dos melhores modelos obtidos em cada geração.

	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.

Fonte: Elaborado pelo autor.

Figura 29: Arquitetura dos modelos de quarta geração: utilização de rótulos de predições anteriores e personalidade para consideração do histórico de emoções.



Fonte: Elaborado pelo autor.

7.4.1 Modelos usando personalidade

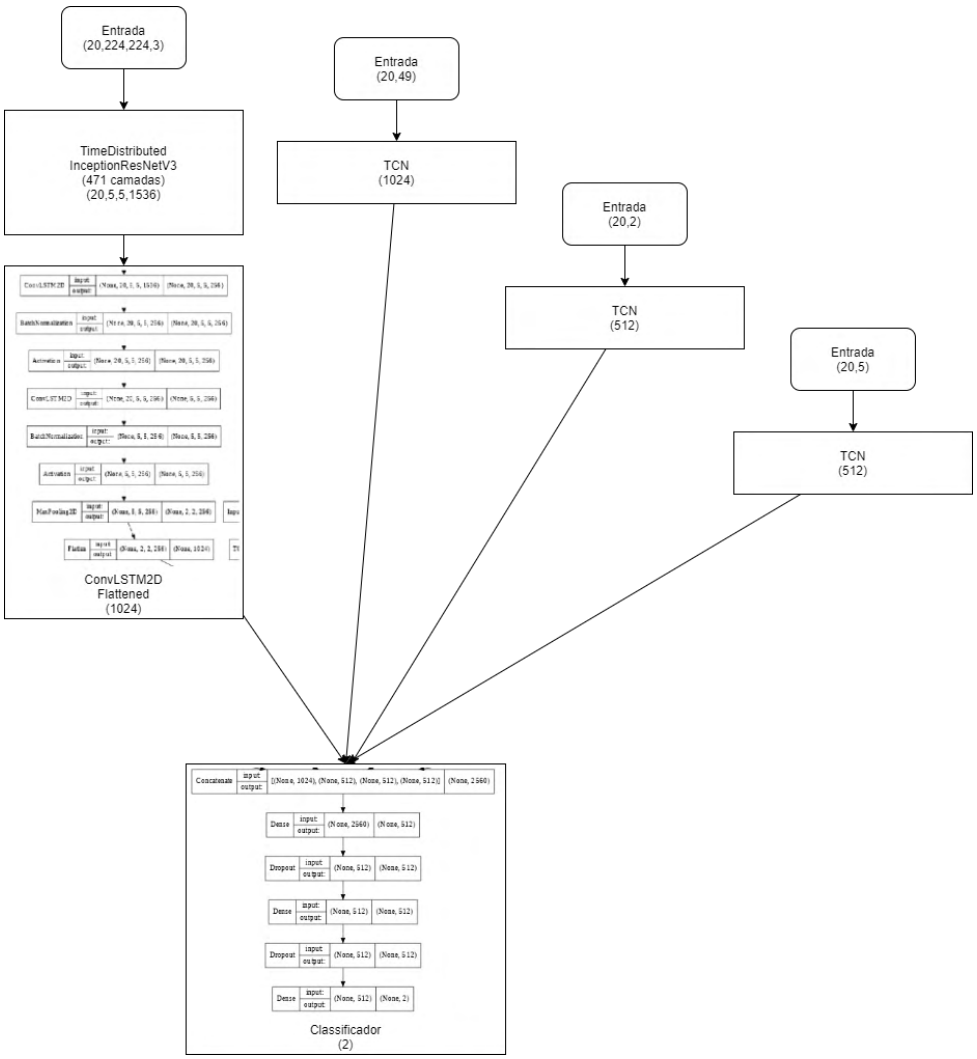
Ainda na quarta geração foram testados modelos incorporando a personalidade do estudante. Porém, como somente a base PAT2Math dispunha de tais informações, somente treinamentos com esta base puderam ser realizados.

Para a incorporação da personalidade, de maneira semelhante ao realizado para incorporação do histórico de emoções detectadas, as informações de personalidade foram encapsuladas em uma rede do tipo TCN (conforme visto na Figura 29), e as características descobertas concatenadas com as características das outras redes para alimentar o classificador.

Para fins comparativos, três versões foram treinadas usando somente a base PAT2Math: (i) Um modelo com a mesma arquitetura vista no último modelo de terceira geração, conforme visto na Seção 7.3 e Figura 25; (ii) Um modelo incorporando o histórico de emoções, da mesma forma que o modelo descrito no início desta seção; (iii) E um modelo utilizando o histórico de emoções e também a personalidade, conforme visto na Figura 29 e na Figura 30.

O modelo (i) obteve $45,09\%$ de acurácia e $F1 = 0,4505$, enquanto o modelo (ii) obteve $88,96\%$ de acurácia e $F1 = 0,8741$, e, finalmente, o modelo (iii) obteve $90,18\%$ de acurácia e $F1 = 0,8870$. A Tabela 15 descreve tais resultados e demonstra que existe um considerável ganho de desempenho ao se utilizar de informações temporais durante o treinamento de emoções

Figura 30: Camadas do modelo de quarta geração treinado para reconhecer engajamento usando o histórico de emoções e personalidade.



Fonte: Elaborado pelo autor.

em vídeo, sobretudo de histórico de emoções e personalidade. A Figura 31 mostra a evolução ao longo das épocas de treinamento dos valores de perda e acurácias curvas para os modelos treinados utilizando a base de dados PAT2Math.

7.4.2 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

Tabela 15: Melhores modelos treinados usando somente a base de dados PAT2Math

	Acurácia	**F1**
Normal (3ª Ger.) (i)	0,4509	0,4505
Histórico de emoções (ii)	0,8896	0,8741
Histórico + Personalidade (iii)	0,9018	0,8871

Fonte: Elaborado pelo autor.

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 16 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. A Figura 32 demonstra, para cada um destes modelos, a evolução ao longo do treinamento das curvas de perda e acurácia.

Tabela 16: 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

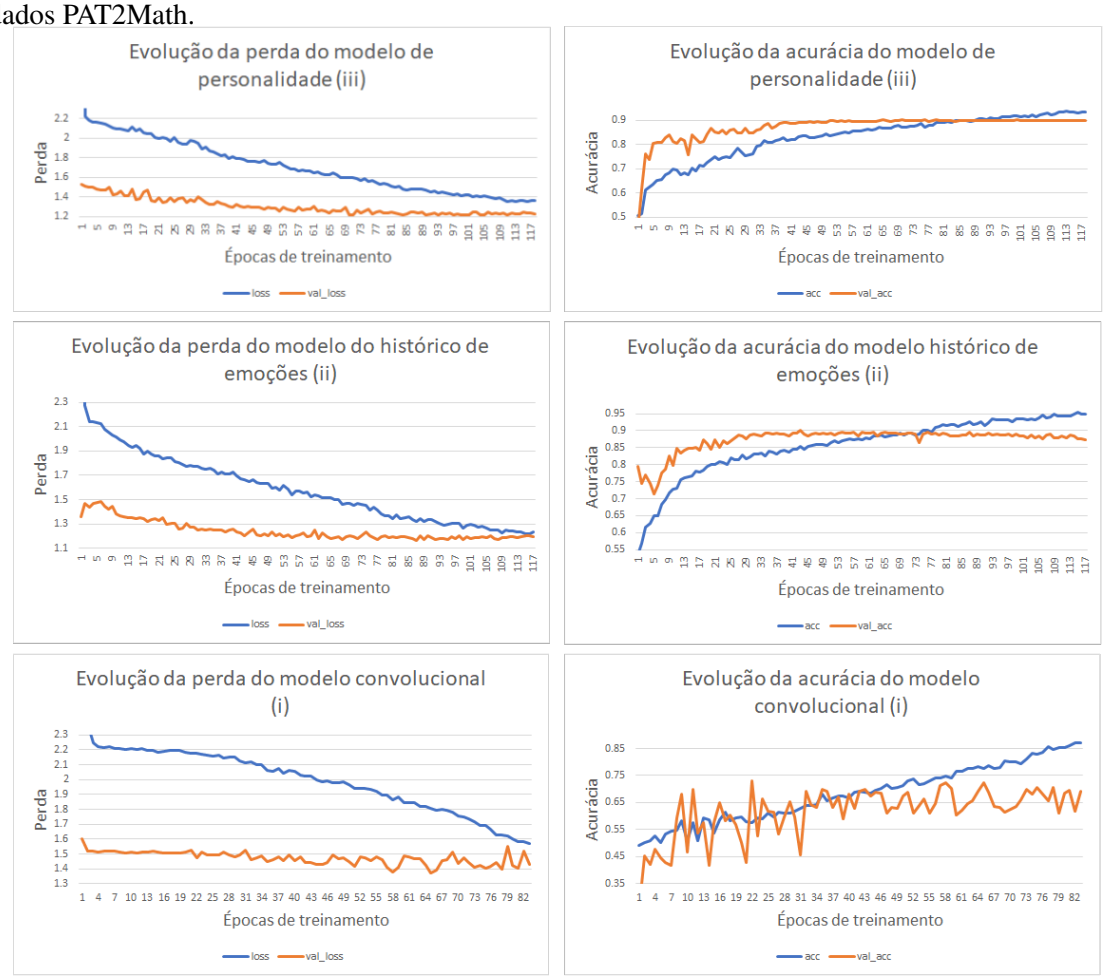
Fonte: Elaborado pelo autor.

7.4.3 Comparativo com o estado-da-arte

Uma vez realizados comparativos que demonstram a vantagem que o emprego do histórico de emoções e personalidade trazem para a detecção de emoções acadêmicas através dos vídeos da face, para a obtenção de uma noção do avanço sobre o estado-da-arte torna-se necessário a realização de comparativos entre os modelos desenvolvidos e diferentes modelos e métodos encontrados em outros trabalhos.

Em vista disso, pode-se observar na Tabela 17 um comparativo entre os trabalhos relacionados apresentados na seção 5.4 e os modelos desenvolvidos para a emoção **engajamento**. Pelos resultados apresentados nota-se que o modelo de quarta geração desenvolvido foi o que obteve melhores resultados considerando a métrica F1, que uma métrica mais justa quando se trata de dados desbalanceados. Este resultado superior pode ser atribuído ao uso do histórico de emoções durante a etapa de treinamento, sendo um modelo mais complexo, que utiliza uma camada extra de abstração sobre o fluxo das emoções comparativamente a modelos que utilizam

Figura 31: Evolução da função de perda e acurácia durante o treinamento dos modelos usando a base de dados PAT2Math.



Fonte: Elaborado pelo autor.

somente informações dos vídeos. Um comparativo sobre o uso da personalidade não pôde ser realizado com outros trabalhos por não haver nenhum outro que utilizasse esta característica e treinado na base de dados PAT2Math.

Ainda sobre o desempenho dos modelos que participaram do comparativo, quando considerado a acurácia, o modelo de terceira geração desenvolvido obteve o segundo lugar dentre os trabalhos analisados, ficando atrás somente do trabalho de Zhang et al. (2019). Cabe ressaltar que não foi possível obter a matriz de confusão deste trabalho para analisar se os resultados que os autores obtiveram de 98,82% para acurácia representa realmente um potencial superior de generalização. Outro ponto muito importante a ser considerado é que o trabalho de Zheng et al. (2021) foi deixado de fora deste comparativo por apresentar alguns problemas metodológicos e de falta de clareza quanto ao modelo desenvolvido, conforme evidenciado na seção 5.4.

Para as outras emoções acadêmicas, infelizmente poucos trabalhos foram encontrados, e, por consequência, a comparação dos resultados foi mais restrita. Para a métrica acurácia, os modelos de quarta geração obtiveram os melhores resultados em comparação aos trabalhos analisados, enquanto que, para a métrica F1, o modelo desenvolvido não obteve o melhor desem-

Tabela 17: Comparativo entre modelo desenvolvido e trabalhos relacionados para a emoção engajamento, usando base DAiSEE.

Trabalho	**Acurácia**	**F1**
(GUPTA et al., 2016a)	0,6283	0,6125
(DEWAN et al., 2018)	0,9323	?
(ABEDI; KHAN, 2021)	0,9221	0,5580
(ZHANG et al., 2019)	**0,9882**	?
Modelo 3ª geração	**0,9417**	0,5122
Modelo 4ª geração	0,8541	**0,6468**

Fonte: Elaborado pelo autor.

penho somente para a emoção tédio. Os resultados deste comparativo para estas três emoções podem ser observados na Tabela 18.

Tabela 18: Comparativo entre modelo desenvolvido e trabalhos relacionados para as emoções.

Trabalho	**Emoções**	**Acurácia**	**F1**
(GUPTA et al., 2016a)	Confusão, Frustração e Tédio	TED: 0,537; FRU: 0,735; CON: 0,723	?
(LEONG, 2020)	Frustração e Tédio	TED: 0,5878; FRU: 0,7309	TED: 0,696, FRU: 0,419
Modelo 4ª geração	Confusão, Frustração e Tédio	TED: **0,7807**; FRU: **0,9546**; CON: **0,8874**	TED: 0,5428; FRU: **0,4991**; CON: 0,6123

Fonte: Elaborado pelo autor.

Figura 32: Evolução da função de perda e acurácia durante o treinamento dos modelos de quarta geração para cada emoção.

Fonte: Elaborado pelo autor.

8 CONCLUSÃO E TRABALHOS FUTUROS

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 traços de personalidade e a temporalidade das emoções dos estudantes (tanto sequencia de imagens do vídeo quanto histórico de emoções do estudante) 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 da tese 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 no Capítulo 5. Uma das explicações para esse desempenho superior é o uso de características não muito comuns de serem vistas em trabalhos da área como, por exemplo, o histórico de emoções e a personalidade.

O **segundo objetivo** específico da tese era 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. 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 14) 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 da tese 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.

O **terceiro objetivo** específico era verificar a influência da inclusão da informação de traço da personalidade do aluno na acurácia da detecção das emoções de aprendizagem quando essa

informação é incluída no treinamento dos algoritmos de aprendizagem profunda de reconhecimento facial de emoções. Também pôde-se observar que a inclusão das características de personalidade dos indivíduos no treinamento das redes aumentou o desempenho do modelo em cerca de $1,48\%$ (de $0,8741$ para $0,8871$ F1, conforme Tabela 15) para o engajamento usando a métrica F1 na base de dados PAT2Math. No entanto, os ganhos obtidos foram mais modestos que os ganhos verificados com a consideração da temporalidade das emoções. Diante de todas essas considerações, os diferenciais do modelo desenvolvido visivelmente resultam em ganhos de desempenho se comparado tanto aos algoritmos estado-da-arte, quanto às mesmas versões de modelos sem estes diferenciais.

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, as abordagens utilizadas para este tipo de detecção foram a consideração da temporalidade da sequência das emoções, bem como a personalidade do aluno. Essas escolhas se justificam pelo fato das emoções de aprendizagem experimentadas por um estudante dependem de sua personalidade e das emoções manifestadas anteriormente (D'MELLO et al., 2014; REIS et al., 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. Além disso, a utilização da personalidade como uma característica auxiliar também demonstrou aumentar a precisão da detecção no modelo temporal, pois de acordo com Reis et al. (2018), o tempo que cada pessoa permanece em uma determinada emoção é influenciado pela sua personalidade.

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, das únicas duas bases de dados públicas existentes que tratam de emoções de aprendizagem (EmotiW e DAiSEE), ambas enfrentam alguns problemas, como, por exemplo, 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, e também possui informação sobre traços de personalidade, além de ser consideravelmente menos desbalanceada.

Sugere-se, portanto, a ampliação da investigação a respeito dos efeitos da consideração da personalidade dos alunos na detecção das emoções, cujos resultados reportados no presente trabalho foram limitados devido à ausência de maior quantidade de dados de entrada que reportem tal informação. Uma maior disponibilidade de informações nesse âmbito iria consequentemente aumentar a qualidade e quantidade de trabalhos desenvolvidos com foco na personalidade e emoções presentes no aprendizado, além de proporcionar um maior potencial comparativo entre trabalhos desenvolvidos e treinados nas mesmas bases.

8.1 Limitações

Embora os resultados atingidos nesse trabalho foram relevantes para a detecção automática de emoções, é importante salientar algumas limitações dos modelos desenvolvidos. Essas limitações se referem à necessária especificidade dos dados de entrada, restrições de contexto para emprego do modelo e tamanho das amostras. A seguir, são descritas as principais limitações identificadas no desenvolvimento deste trabalho.

- Os últimos modelos desenvolvidos, que apresentaram melhores resultados, foram treinados para o reconhecimento da emoção engajamento. Usando os mesmos parâmetros, os modelos foram retreinados para o reconhecimento das emoções confusão, frustração e tédio. Para um melhor desempenho para estas outras emoções, o ideal seria uma pesquisa de parâmetros específicos para cada emoção.

- As bases utilizadas para treinamento foram compostas majoritariamente de alunos asiáticos, e uma pequena parcela de alunos brasileiros, o que pode comprometer o viés de generalização do modelo.

- O treinamento dos modelos para uso da personalidade na detecção da emoção se deu em uma base de dados menor, o que pode afetar a eficácia da abordagem quando utilizado em um ambiente real.
- O treinamento foi utilizado com base somente no rosto do estudante, incluindo suas expressões faciais, posição e rotação da cabeça. Portanto, outros sinais do corpo não são considerados pelo modelo.
- O reconhecimento facial é limitado quando o rosto do aluno não está totalmente visível na câmera, afetando a capacidade do algoritmo de reconhecer a emoção.
- Embora treinado para reconhecimento de emoções em ambientes de diferentes níveis de luminosidade, o algoritmo não consegue reconhecer o rosto em ambientes de luminosidade extremamente baixa.

REFERÊNCIAS

ABEDI, A.; KHAN, S. S. Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network. In: CONFERENCE ON ROBOTS AND VISION (CRV), 2021., 2021. **Anais...** [S.l.: s.n.], 2021. p. 151–157.

ACKERMANN, P.; KOHLSCHEIN, C.; BITSCH, J. Á.; WEHRLE, K.; JESCHKE, S. EEG-based automatic emotion recognition: feature extraction, selection and classification methods. In: IEEE 18TH INTERNATIONAL CONFERENCE ON E-HEALTH NETWORKING, APPLICATIONS AND SERVICES (HEALTHCOM), 2016., 2016. **Anais...** [S.l.: s.n.], 2016. p. 1–6.

AHONEN, T.; HADID, A.; PIETIKÄINEN, M. Face recognition with local binary patterns. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2004. **Anais...** [S.l.: s.n.], 2004. p. 469–481.

AMMAR, M. B.; NEJI, M.; ALIMI, A. M.; GOUARDÈRES, G. The affective tutoring system. **Expert Systems with Applications**, [S.l.], v. 37, n. 4, p. 3013–3023, 2010.

ANDERSON, J. R. The expert module. **Foundations of intelligent tutoring systems**, [S.l.], p. 21–53, 1988.

ANDONI, A.; PANIGRAHY, R.; VALIANT, G.; ZHANG, L. Learning polynomials with neural networks. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 2014. **Anais...** [S.l.: s.n.], 2014. p. 1908–1916.

BALTRUSAITIS, T.; ZADEH, A.; LIM, Y. C.; MORENCY, L.-P. Openface 2.0: facial behavior analysis toolkit. In: IEEE INTERNATIONAL CONFERENCE ON AUTOMATIC FACE & GESTURE RECOGNITION (FG 2018), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 59–66.

BÄNZIGER, T.; SCHERER, K. R. Introducing the geneva multimodal emotion portrayal (gemep) corpus. **Blueprint for affective computing: A sourcebook**, [S.l.], p. 271–294, 2010.

BIANCO, S.; CADENE, R.; CELONA, L.; NAPOLETANO, P. Benchmark analysis of representative deep neural network architectures. **IEEE access**, [S.l.], v. 6, p. 64270–64277, 2018.

BOSCH, N.; D'MELLO, S. The affective experience of novice computer programmers. **International journal of artificial intelligence in education**, [S.l.], v. 27, n. 1, p. 181–206, 2017.

BRADLEY, M. M.; LANG, P. J. Measuring emotion: the self-assessment manikin and the semantic differential. **Journal of behavior therapy and experimental psychiatry**, [S.l.], v. 25, n. 1, p. 49–59, 1994.

BRADSKI, G. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, [S.l.], 2000.

BROOKS, S. Does personal social media usage affect efficiency and well-being? **Computers in Human Behavior**, [S.l.], v. 46, p. 26–37, 2015.

CALVO, R. A.; D'MELLO, S. Affect detection: an interdisciplinary review of models, methods, and their applications. **IEEE Transactions on affective computing**, [S.l.], v. 1, n. 1, p. 18–37, 2010.

CAMBRIA, E. Affective computing and sentiment analysis. **IEEE Intelligent Systems**, [S.l.], v. 31, n. 2, p. 102–107, 2016.

CARREIRA, J.; ZISSERMAN, A. Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2017. **Anais...** [S.l.: s.n.], 2017. p. 6299–6308.

CASTILLO, J. C.; CASTRO-GONZÁLEZ, Á.; ALONSO-MARTÍN, F.; FERNÁNDEZ-CABALLERO, A.; SALICHS, M. Á. Emotion detection and regulation from personal assistant robot in smart environment. In: **Personal Assistants**: emerging computational technologies. [S.l.]: Springer, 2018. p. 179–195.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, [S.l.], v. 16, p. 321–357, 2002.

CHO, K.; VAN MERRIËNBOER, B.; GULCEHRE, C.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. **arXiv preprint arXiv:1406.1078**, [S.l.], 2014.

COOTES, T. F.; TAYLOR, C. J. et al. **Statistical models of appearance for computer vision**. [S.l.]: Technical report, University of Manchester, 2004.

DAHMANE, M.; MEUNIER, J. Emotion recognition using dynamic grid-based HoG features. In: FACE AND GESTURE 2011, 2011. **Anais...** [S.l.: s.n.], 2011. p. 884–888.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR'05), 2005., 2005. **Anais...** [S.l.: s.n.], 2005. v. 1, p. 886–893.

DALGLEISH, T.; POWER, M. **Handbook of cognition and emotion**. [S.l.]: John Wiley & Sons, 2000.

DARWIN, C.; PRODGER, P. **The expression of the emotions in man and animals**. [S.l.]: Oxford University Press, USA, 1998.

DEN UYL, M.; VAN KUILENBURG, H. The FaceReader: online facial expression recognition. In: OF MEASURING BEHAVIOR, 2005. **Proceedings...** [S.l.: s.n.], 2005. v. 30, n. 2, p. 589–590.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: a large-scale hierarchical image database. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2009., 2009. **Anais...** [S.l.: s.n.], 2009. p. 248–255.

DÉNIZ, O.; BUENO, G.; SALIDO, J.; TORRE, F. De la. Face recognition using histograms of oriented gradients. **Pattern Recognition Letters**, [S.l.], v. 32, n. 12, p. 1598–1603, 2011.

DEWAN, M. A. A.; LIN, F.; WEN, D.; MURSHED, M.; UDDIN, Z. A Deep Learning Approach to Detecting Engagement of Online Learners. In: IEEE SMARTWORLD, UBIQUITOUS INTELLIGENCE & COMPUTING, ADVANCED & TRUSTED COMPUTING, SCALABLE COMPUTING & COMMUNICATIONS, CLOUD & BIG DATA COMPUTING, INTERNET OF PEOPLE AND SMART CITY INNOVATION (SMARTWORLD/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 1895–1902.

DEWAN, M. A. A.; MURSHED, M.; LIN, F. Engagement detection in online learning: a review. **Smart Learning Environments**, [S.l.], v. 6, n. 1, p. 1, 2019.

DHALL, A.; GOECKE, R.; LUCEY, S.; GEDEON, T. Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS (ICCV WORKSHOPS), 2011., 2011. **Anais...** [S.l.: s.n.], 2011. p. 2106–2112.

DHALL, A.; GOECKE, R.; LUCEY, S.; GEDEON, T. Acted facial expressions in the wild database. **Australian National University, Canberra, Australia, Technical Report TR-CS-11**, [S.l.], v. 2, p. 1, 2011.

DHALL, A.; KAUR, A.; GOECKE, R.; GEDEON, T. Emotiw 2018: audio-video, student engagement and group-level affect prediction. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2018., 2018. **Proceedings...** [S.l.: s.n.], 2018. p. 653–656.

D'MELLO, S.; CALVO, R. A. Beyond the basic emotions: what should affective computing compute? In: CHI'13 EXTENDED ABSTRACTS ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2013. **Anais...** [S.l.: s.n.], 2013. p. 2287–2294.

D'MELLO, S.; PICARD, R. W.; GRAESSER, A. Toward an affect-sensitive AutoTutor. **IEEE Intelligent Systems**, [S.l.], v. 22, n. 4, 2007.

D'MELLO, S.; GRAESSER, A. Dynamics of affective states during complex learning. **Learning and Instruction**, [S.l.], v. 22, n. 2, p. 145–157, 2012.

D'MELLO, S.; LEHMAN, B.; PEKRUN, R.; GRAESSER, A. Confusion can be beneficial for learning. **Learning and Instruction**, [S.l.], v. 29, p. 153–170, 2014.

EBRAHIMI KAHOU, S.; MICHALSKI, V.; KONDA, K.; MEMISEVIC, R.; PAL, C. Recurrent neural networks for emotion recognition in video. In: ACM ON INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2015., 2015. **Proceedings...** [S.l.: s.n.], 2015. p. 467–474.

EKMAN, P. An argument for basic emotions. **Cognition & emotion**, [S.l.], v. 6, n. 3-4, p. 169–200, 1992.

EKMAN, P. Basic emotions. **Handbook of cognition and emotion**, [S.l.], p. 45–60, 1999.

EKMAN, P. E.; DAVIDSON, R. J. **The nature of emotion**: fundamental questions. [S.l.]: Oxford University Press, 1994.

EKMAN, R. **What the face reveals**: basic and applied studies of spontaneous expression using the facial action coding system (facs). [S.l.]: Oxford University Press, USA, 1997.

ELRAHMAN, S. M. A.; ABRAHAM, A. A review of class imbalance problem. **Journal of Network and Innovative Computing**, [S.l.], v. 1, n. 2013, p. 332–340, 2013.

FASEL, B.; LUETTIN, J. Automatic facial expression analysis: a survey. **Pattern recognition**, [S.l.], v. 36, n. 1, p. 259–275, 2003.

FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, [S.l.], v. 27, n. 8, p. 861–874, 2006.

FREDRICKSON, B. L. What good are positive emotions? **Review of General Psychology**, [S.l.], v. 2, n. 3, p. 300–319, 1998.

FREITAS PEREIRA, T. de; ANJOS, A.; DE MARTINO, J. M.; MARCEL, S. LBP- TOP based countermeasure against face spoofing attacks. In: ASIAN CONFERENCE ON COMPUTER VISION, 2012. **Anais...** [S.l.: s.n.], 2012. p. 121–132.

Frossard, Davi. **VGG in TensorFlow**. [Online; acessado em 4 de Outubro de 2019].

GOLDBERG, L. R. An alternative"description of personality": the big-five factor structure. **Journal of personality and social psychology**, [S.l.], v. 59, n. 6, p. 1216, 1990.

GOLDBERG, L. R.; JOHNSON, J. A.; EBER, H. W.; HOGAN, R.; ASHTON, M. C.; CLONINGER, C. R.; GOUGH, H. G. The international personality item pool and the future of public-domain personality measures. **Journal of Research in personality**, [S.l.], v. 40, n. 1, p. 84–96, 2006.

GOLDONI, D.; REIS, H.; JAQUES, P. A. Modelagem Estatística do Tempo de Permanência de Estudantes no Estado de Confusão Através de Análise de Sobrevivência Multivariada. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2022. **Anais...** SBC, 2022.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.

GOODFELLOW, I. J.; ERHAN, D.; CARRIER, P. L.; COURVILLE, A.; MIRZA, M.; HAMNER, B.; CUKIERSKI, W.; TANG, Y.; THALER, D.; LEE, D.-H. et al. Challenges in representation learning: a report on three machine learning contests. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING, 2013. **Anais...** [S.l.: s.n.], 2013. p. 117–124.

GOODMAN, L. A. Snowball sampling. **The annals of mathematical statistics**, [S.l.], p. 148–170, 1961.

GORDON, G.; SPAULDING, S.; WESTLUND, J. K.; LEE, J. J.; PLUMMER, L.; MARTINEZ, M.; DAS, M.; BREAZEAL, C. Affective personalization of a social robot tutor for children's second language skills. In: THIRTIETH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2016. **Anais...** [S.l.: s.n.], 2016.

GROSS, R.; MATTHEWS, I.; COHN, J.; KANADE, T.; BAKER, S. Multi-pie. **Image and Vision Computing**, [S.l.], v. 28, n. 5, p. 807–813, 2010.

Gross, Sam and Wilber, Michael. **Training and investigating Residual Nets**. [Online; acessado em 4 de Outubro de 2019].

GUPTA, A.; D'CUNHA, A.; AWASTHI, K.; BALASUBRAMANIAN, V. Daisee: towards user engagement recognition in the wild. **arXiv preprint arXiv:1609.01885**, [S.l.], 2016.

GUPTA, A.; JAISWAL, R.; ADHIKARI, S.; BALASUBRAMANIAN, V. N. DAISEE: dataset for affective states in e-learning environments. **arXiv**, [S.l.], p. 1–22, 2016.

HALL, C. S.; LINDZEY, G.; CAMPBELL, J. B. **Teorias da personalidade**. [S.l.]: Artmed Editora, 2000.

HARLEY, J. M.; BOUCHET, F.; HUSSAIN, M. S.; AZEVEDO, R.; CALVO, R. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. **Computers in Human Behavior**, [S.l.], v. 48, p. 615–625, 2015.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2015. **Proceedings...** [S.l.: s.n.], 2015. p. 1026–1034.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 770–778.

HECKATHORN, D. D. Respondent-driven sampling: a new approach to the study of hidden populations. **Social problems**, [S.l.], v. 44, n. 2, p. 174–199, 1997.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. **Neural computation**, [S.l.], v. 18, n. 7, p. 1527–1554, 2006.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, [S.l.], v. 9, n. 8, p. 1735–1780, 1997.

HOLMES, G.; DONKIN, A.; WITTEN, I. WEKA: a machine learning workbench. In: ANZIIS'94-AUSTRALIAN NEW ZEALND INTELLIGENT INFORMATION SYSTEMS CONFERENCE, 1994. **Proceedings...** [S.l.: s.n.], 1994. p. 357–361.

HU, W.-S.; LI, H.-C.; PAN, L.; LI, W.; TAO, R.; DU, Q. Spatial–spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification. **IEEE Transactions on Geoscience and Remote Sensing**, [S.l.], v. 58, n. 6, p. 4237–4250, 2020.

HUANG, F.-J.; LECUN, Y. Large-scale learning with svm and convolutional nets for generic object categorization. In: COMPUTER VISION AND PATTERN RECOGNITION CONFERENCE (CVPR'06), 2006. **Proceedings...** [S.l.: s.n.], 2006.

HUANG, G.; LIU, Z.; VAN DER MAATEN, L.; WEINBERGER, K. Q. Densely connected convolutional networks. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2017. **Proceedings...** [S.l.: s.n.], 2017. p. 4700–4708.

HUANG, W.; SONG, G.; LI, M.; HU, W.; XIE, K. Adaptive Weight Optimization for Classification of Imbalanced Data. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SCIENCE AND BIG DATA ENGINEERING, 2013. **Anais...** [S.l.: s.n.], 2013. p. 546–553.

HUTMACHER, F. Why is there so much more research on vision than on any other sensory modality? **Frontiers in psychology**, [S.l.], v. 10, p. 2246, 2019.

INZLICHT, M.; BARTHOLOW, B. D.; HIRSH, J. B. Emotional foundations of cognitive control. **Trends in cognitive sciences**, [S.l.], v. 19, n. 3, p. 126–132, 2015.

JAQUES, P. A.; NUNES, M. A. S. Computação Afetiva aplicada à Educação. In: PIMENTEL MARIANO; SAMPAIO, F. F. (Ed.). **Informática na Educação**: técnicas e tecnologias computacionais. Porto alegre: [s.n.], 2019. (Informática na Educação, v. 3). Disponível em: <http://ieducacao.ceie-br.org/computacaoafetiva>.

JAQUES, P. A.; SEFFRIN, H.; RUBI, G. L.; MORAIS, F. de; GHILARDI, C.; BITTENCOURT, I. I.; ISOTANI, S. Rule-based expert systems to support step-by-step guidance in algebraic problem solving: the case of the tutor pat2math. **Expert Systems with Applications**, [S.l.], v. 40, n. 14, p. 5456–5465, 2013.

JONES, M.; VIOLA, P. Fast multi-view face detection. **Mitsubishi Electric Research Lab TR-20003-96**, [S.l.], v. 3, n. 14, p. 2, 2003.

JYOTI, S.; SHARMA, G.; DHALL, A. A Single Hierarchical Network for Face, Action Unit and Emotion Detection. In: DIGITAL IMAGE COMPUTING: TECHNIQUES AND APPLICATIONS (DICTA), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 1–8.

Kaggle. **Challenges in Representation Learning**: facial expression recognition challenge. [Online; accessed 17-July-2019].

KAUR, A.; MUSTAFA, A.; MEHTA, L.; DHALL, A. Prediction and localization of student engagement in the wild. In: DIGITAL IMAGE COMPUTING: TECHNIQUES AND APPLICATIONS (DICTA), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 1–8.

KHORRAMI, P.; PAINE, T.; HUANG, T. Do deep neural networks learn facial action units when doing expression recognition? In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS, 2015. **Proceedings...** [S.l.: s.n.], 2015. p. 19–27.

KIM, D. H.; BADDAR, W. J.; JANG, J.; RO, Y. M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. **IEEE Transactions on Affective Computing**, [S.l.], v. 10, n. 2, p. 223–236, 2017.

KING, D. E. Dlib-ml: a machine learning toolkit. **Journal of Machine Learning Research**, [S.l.], v. 10, n. Jul, p. 1755–1758, 2009.

KLEINGINNA, P. R.; KLEINGINNA, A. M. A categorized list of emotion definitions, with suggestions for a consensual definition. **Motivation and emotion**, [S.l.], v. 5, n. 4, p. 345–379, 1981.

KOLLIAS, D.; NICOLAOU, M. A.; KOTSIA, I.; ZHAO, G.; ZAFEIRIOU, S. Recognition of affect in the wild using deep neural networks. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, 2017. **Proceedings...** [S.l.: s.n.], 2017. p. 26–33.

KONSTANTINOVA, M.; KAZIMIROVA, E.; PEREPELKINA, O. **Classification of affective and social behaviors in public interaction for affective computing and social signal processing**. [S.l.]: PeerJ Preprints, 2018.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, [S.l.], v. 5, n. 4, p. 221–232, 2016.

KREIFELTS, B.; WILDGRUBER, D.; ETHOFER, T. Audiovisual integration of emotional information from voice and face. In: **Integrating face and voice in person perception**. [S.l.]: Springer, 2013. p. 225–251.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2012. **Anais...** [S.l.: s.n.], 2012. p. 1097–1105.

KUMARI, J.; RAJESH, R.; POOJA, K. Facial expression recognition: a survey. **Procedia Computer Science**, [S.l.], v. 58, p. 486–491, 2015.

LAI, M.-L.; TSAI, M.-J.; YANG, F.-Y.; HSU, C.-Y.; LIU, T.-C.; LEE, S. W.-Y.; LEE, M.-H.; CHIOU, G.-L.; LIANG, J.-C.; TSAI, C.-C. A review of using eye-tracking technology in exploring learning from 2000 to 2012. **Educational research review**, [S.l.], v. 10, p. 90–115, 2013.

LANE, R. D.; NADEL, L. **Cognitive neuroscience of emotion**. [S.l.]: Oxford University Press, 2002.

LAZARUS, R. S. Thoughts on the relations between emotion and cognition. **American psychologist**, [S.l.], v. 37, n. 9, p. 1019, 1982.

LAZARUS, R. S. The cognition-emotion debate: a bit of history. **Handbook of cognition and emotion**, [S.l.], v. 5, n. 6, p. 3–19, 1999.

LEA, C.; FLYNN, M. D.; VIDAL, R.; REITER, A.; HAGER, G. D. Temporal convolutional networks for action segmentation and detection. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2017. **Anais...** [S.l.: s.n.], 2017. p. 156–165.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, [S.l.], v. 521, n. 7553, p. 436, 2015.

LEONG, F. H. Deep learning of facial embeddings and facial landmark points for the detection of academic emotions. In: INTERNATIONAL CONFERENCE ON INFORMATION AND EDUCATION INNOVATIONS, 5., 2020. **Proceedings...** [S.l.: s.n.], 2020. p. 111–116.

LI, J.; LAM, E. Y. Facial expression recognition using deep neural networks. In: IEEE INTERNATIONAL CONFERENCE ON IMAGING SYSTEMS AND TECHNIQUES (IST), 2015., 2015. **Anais...** [S.l.: s.n.], 2015. p. 1–6.

LI, S.; DENG, W. Deep facial expression recognition: a survey. **arXiv preprint arXiv:1804.08348**, [S.l.], 2018.

LI, S.; DENG, W.; DU, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2017. **Proceedings...** [S.l.: s.n.], 2017. p. 2852–2861.

LIU, C.; TANG, T.; LV, K.; WANG, M. Multi-feature based emotion recognition for video clips. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2018., 2018. **Proceedings...** [S.l.: s.n.], 2018. p. 630–634.

LONGADGE, R.; DONGRE, S. Class imbalance problem in data mining review. **arXiv preprint arXiv:1305.1707**, [S.l.], 2013.

LU, C.; ZHENG, W.; LI, C.; TANG, C.; LIU, S.; YAN, S.; ZONG, Y. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2018., 2018. **Proceedings...** [S.l.: s.n.], 2018. p. 646–652.

LUCEY, P.; COHN, J. F.; KANADE, T.; SARAGIH, J.; AMBADAR, Z.; MATTHEWS, I. The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION-WORKSHOPS, 2010., 2010. **Anais...** [S.l.: s.n.], 2010. p. 94–101.

MAJUMDER, N.; PORIA, S.; GELBUKH, A.; CAMBRIA, E. Deep learning-based document modeling for personality detection from text. **IEEE Intelligent Systems**, [S.l.], v. 32, n. 2, p. 74–79, 2017.

MARCUS, G. Deep learning: a critical appraisal. **arXiv preprint arXiv:1801.00631**, [S.l.], 2018.

MARTINEZ, B.; VALSTAR, M. F.; JIANG, B.; PANTIC, M. Automatic analysis of facial actions: a survey. **IEEE transactions on affective computing**, [S.l.], 2017.

MASE, K. Recognition of facial expression from optical flow. **IEICE TRANSACTIONS on Information and Systems**, [S.l.], v. 74, n. 10, p. 3474–3483, 1991.

MAVADATI, S. M.; MAHOOR, M. H.; BARTLETT, K.; TRINH, P.; COHN, J. F. Disfa: a spontaneous facial action intensity database. **IEEE Transactions on Affective Computing**, [S.l.], v. 4, n. 2, p. 151–160, 2013.

MCDUFF, D.; MAHMOUD, A.; MAVADATI, M.; AMR, M.; TURCOT, J.; KALIOUBY, R. e. AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In: CHI CONFERENCE EXTENDED ABSTRACTS ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2016., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 3723–3726.

MIKOLOV, T.; KARAFIÁT, M.; BURGET, L.; ČERNOCKỲ, J.; KHUDANPUR, S. Recurrent neural network based language model. In: ELEVENTH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, 2010. **Anais...** [S.l.: s.n.], 2010.

MITCHELL, T.; BUCHANAN, B.; DEJONG, G.; DIETTERICH, T.; ROSENBLOOM, P.; WAIBEL, A. Machine learning. **Annual review of computer science**, [S.l.], v. 4, n. 1, p. 417–433, 1990.

MOLLAHOSSEINI, A.; CHAN, D.; MAHOOR, M. H. Going deeper in facial expression recognition using deep neural networks. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV), 2016., 2016. **Anais...** [S.l.: s.n.], 2016. p. 1–10.

MONKARESI, H.; BOSCH, N.; CALVO, R. A.; D'MELLO, S. K. Automated detection of engagement using video-based estimation of facial expressions and heart rate. **IEEE Transactions on Affective Computing**, [S.l.], v. 8, n. 1, p. 15–28, 2016.

MORAIS, F. de; KAUTZMANN, T. R.; BITTENCOURT, I. I.; Patricia A. Jaques. EmAP-ML: a protocol of emotions and behaviors annotation for machine learning labels. In: EC-TEL, 2019, Netherlands. **Anais...** Springer, 2019.

NARDELLI, M.; VALENZA, G.; GRECO, A.; LANATA, A.; SCILINGO, E. P. Recognizing emotions induced by affective sounds through heart rate variability. **IEEE Transactions on Affective Computing**, [S.l.], v. 6, n. 4, p. 385–394, 2015.

NeuroData Lab. **Comparing Emotion Recognition Tech**: microsoft, neurodata lab, amazon, affectiva. [Online; accessed 3-June-2019].

NEZAMI, O. M.; DRAS, M.; HAMEY, L.; RICHARDS, D.; WAN, S.; PARIS, C. Automatic Recognition of Student Engagement using Deep Learning and Facial Expression. **arXiv preprint arXiv:1808.02324**, [S.l.], 2018.

NG, A. et al. Sparse autoencoder. **CS294A Lecture notes**, [S.l.], v. 72, n. 2011, p. 1–19, 2011.

NIU, X.; HAN, H.; ZENG, J.; SUN, X.; SHAN, S.; HUANG, Y.; YANG, S.; CHEN, X. Automatic Engagement Prediction with GAP Feature. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2018., 2018. **Proceedings...** [S.l.: s.n.], 2018. p. 599–603.

NUNES, M. Computação Afetiva personalizando interfaces, interações e recomendações de produtos, serviços e pessoas em ambientes computacionais. **DCOMP e PROCC: Pesquisas e Editora UFS: São Cristóvão**, [S.l.], p. 115–151, 2012.

NWANA, H. S. Intelligent tutoring systems: an overview. **Artificial Intelligence Review**, [S.l.], v. 4, n. 4, p. 251–277, 1990.

OCUMPAUGH, J. 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**, [S.l.], v. 60, 2015.

OLAH, C. **The repeating module in an LSTM**. [S.l.: s.n.], 2015.

OSÓRIO, F. S.; BITTENCOURT, J. R. Sistemas Inteligentes baseados em redes neurais artificiais aplicados ao processamento de imagens. In: I WORKSHOP DE INTELIGÊNCIA ARTIFICIAL, 2000. **Anais...** [S.l.: s.n.], 2000.

PANTIC, M.; VALSTAR, M.; RADEMAKER, R.; MAAT, L. Web-based database for facial expression analysis. In: IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO, 2005., 2005. **Anais...** [S.l.: s.n.], 2005. p. 5–pp.

PASCANU, R.; MIKOLOV, T.; BENGIO, Y. On the difficulty of training recurrent neural networks. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 2013. **Anais...** [S.l.: s.n.], 2013. p. 1310–1318.

PEKRUN, R. Emotions as drivers of learning and cognitive development. In: **New perspectives on affect and learning technologies**. [S.l.]: Springer, 2011. p. 23–39.

PEKRUN, R. Emotions and learning. In: **Educational practices series**. [S.l.]: IEA, IBE, 2014.

PEKRUN, R.; ELLIOT, A. J.; MAIER, M. A. Achievement goals and achievement emotions: testing a model of their joint relations with academic performance. **Journal of educational Psychology**, [S.l.], v. 101, n. 1, p. 115, 2009.

PICARD, R. W. **Affective computing**. [S.l.]: MIT press, 2000.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. **Machine learning**, [S.l.], v. 68, n. 3, p. 169–171, 2011.

RAMANAN, D.; ZHU, X. Face detection, pose estimation, and landmark localization in the wild. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2012., 2012. **Anais...** [S.l.: s.n.], 2012. p. 2879–2886.

REIS, H.; ALVARES, D.; JAQUES, P.; ISOTANI, S. Analysis of permanence time in emotional states: a case study using educational software. In: INTERNATIONAL CONFERENCE ON INTELLIGENT TUTORING SYSTEMS, 2018. **Anais...** [S.l.: s.n.], 2018. p. 180–190.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M. et al. Imagenet large scale visual recognition challenge. **International journal of computer vision**, [S.l.], v. 115, n. 3, p. 211–252, 2015.

SARIYANIDI, E.; GUNES, H.; CAVALLARO, A. Automatic analysis of facial affect: a survey of registration, representation, and recognition. **IEEE transactions on pattern analysis and machine intelligence**, [S.l.], v. 37, n. 6, p. 1113–1133, 2014.

SCHERER, K. R. Appraisal theory. **Handbook of cognition and emotion**, [S.l.], p. 637–663, 1999.

SCHERER, K. R. What are emotions? And how can they be measured? **Social science information**, [S.l.], v. 44, n. 4, p. 695–729, 2005.

SCHERER, K. R. et al. Psychological models of emotion. **The neuropsychology of emotion**, [S.l.], v. 137, n. 3, p. 137–162, 2000.

SCHMIDHUBER, J. Deep learning in neural networks: an overview. **Neural networks**, [S.l.], v. 61, p. 85–117, 2015.

SCHWARZ, N. Feelings as information: informational and motivational functions of affective states. **Handbook of Motivation ad Cognition: Foundations of Social Behaviour**, [S.l.], v. 2, p. 527–561, 1990.

SENECHAL, T.; MCDUFF, D.; KALIOUBY, R. Facial action unit detection using active learning and an efficient non-linear kernel approximation. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS, 2015. **Proceedings...** [S.l.: s.n.], 2015. p. 10–18.

SIDNEY, K. D.; CRAIG, S. D.; GHOLSON, B.; FRANKLIN, S.; PICARD, R.; GRAESSER, A. C. Integrating affect sensors in an intelligent tutoring system. In: AFFECTIVE INTERACTIONS: THE COMPUTER IN THE AFFECTIVE LOOP WORKSHOP AT, 2005. **Anais...** [S.l.: s.n.], 2005. p. 7–13.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, [S.l.], 2014.

SIQUEIRA-BATISTA, R.; VITORINO, R. R.; GOMES, A. P.; OLIVEIRA, A. d. P.; FERREIRA, R. d. S.; ESPERIDIÃO-ANTONIO, V.; SANTANA, L. A.; CERQUEIRA, F. R. Artificial neural networks and medical education. **Revista Brasileira de Educação Médica**, [S.l.], v. 38, n. 4, p. 548–556, 2014.

STRAPPARAVA, C.; MIHALCEA, R. Learning to identify emotions in text. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2008., 2008. **Proceedings...** [S.l.: s.n.], 2008. p. 1556–1560.

SUBRAMANIAN, R.; WACHE, J.; ABADI, M. K.; VIERIU, R. L.; WINKLER, S.; SEBE, N. ASCERTAIN: emotion and personality recognition using commercial sensors. **IEEE Transactions on Affective Computing**, [S.l.], v. 9, n. 2, p. 147–160, 2016.

SUNG, K. K.; POGGIO, T. **Example Based Learning for View-Based Human Face Detection.** [S.l.]: MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 1994.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; ALEMI, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: THIRTY-FIRST AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2017. **Anais...** [S.l.: s.n.], 2017.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2015. **Proceedings...** [S.l.: s.n.], 2015. p. 1–9.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the inception architecture for computer vision. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 2818–2826.

TAO, J.; TAN, T. Affective computing: a review. In: INTERNATIONAL CONFERENCE ON AFFECTIVE COMPUTING AND INTELLIGENT INTERACTION, 2005. **Anais...** [S.l.: s.n.], 2005. p. 981–995.

THOMAS, C.; NAIR, N.; JAYAGOPI, D. B. Predicting Engagement Intensity in the Wild Using Temporal Convolutional Network. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2018., 2018. **Proceedings...** [S.l.: s.n.], 2018. p. 604–610.

TRAN, D.; BOURDEV, L.; FERGUS, R.; TORRESANI, L.; PALURI, M. Learning spatiotemporal features with 3d convolutional networks. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2015. **Proceedings...** [S.l.: s.n.], 2015. p. 4489–4497.

VANLEHN, K. **Student Modelling**. [S.l.]: Carnegie Mellon University, 1987.

VANLEHN, K. The Behavior of Tutoring Systems. **International Journal of Artificial Intelligence in Education**, [S.l.], v. 16, p. 227–265, 2006.

VIOLA, P.; JONES, M. et al. Rapid object detection using a boosted cascade of simple features. **CVPR (1)**, [S.l.], v. 1, p. 511–518, 2001.

WANG, S.; LIU, W.; WU, J.; CAO, L.; MENG, Q.; KENNEDY, P. J. Training deep neural networks on imbalanced data sets. In: IJCNN), 2016., 2016. **Anais...** [S.l.: s.n.], 2016. p. 4368–4374.

WERLANG, P. **GitHub**: werlang/emolearn-ml-model. [Online; acessado 25-Setembro-2022], `https://github.com/werlang/emolearn-ml-model`.

WHITEHILL, J.; SERPELL, Z.; LIN, Y.-C.; FOSTER, A.; MOVELLAN, J. R. The faces of engagement: automatic recognition of student engagementfrom facial expressions. **IEEE Transactions on Affective Computing**, [S.l.], v. 5, n. 1, p. 86–98, 2014.

WOOLF, B. P. **Building Intelligent Interactive Tutors**: student-centered strategies for revolutionizing e-learning. 1. ed. [S.l.]: Morgan Kaufmann, 2008.

YAN, W.-J.; LI, X.; WANG, S.-J.; ZHAO, G.; LIU, Y.-J.; CHEN, Y.-H.; FU, X. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. **PloS one**, [S.l.], v. 9, n. 1, p. e86041, 2014.

YANG, J.; WANG, K.; PENG, X.; QIAO, Y. Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION, 2018., 2018. **Proceedings...** [S.l.: s.n.], 2018. p. 594–598.

ZAJONC, R. B. Feeling and thinking: preferences need no inferences. **American psychologist**, [S.l.], v. 35, n. 2, p. 151, 1980.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2014. **Anais...** [S.l.: s.n.], 2014. p. 818–833.

ZENG, N.; ZHANG, H.; SONG, B.; LIU, W.; LI, Y.; DOBAIE, A. M. Facial expression recognition via learning deep sparse autoencoders. **Neurocomputing**, [S.l.], v. 273, p. 643–649, 2018.

ZHANG, H.; XIAO, X.; HUANG, T.; LIU, S.; XIA, Y.; LI, J. An novel end-to-end network for automatic student engagement recognition. In: IEEE 9TH INTERNATIONAL CONFERENCE ON ELECTRONICS INFORMATION AND EMERGENCY COMMUNICATION (ICEIEC), 2019., 2019. **Anais...** [S.l.: s.n.], 2019. p. 342–345.

ZHANG, K.; ZHANG, Z.; LI, Z.; QIAO, Y. Joint face detection and alignment using multitask cascaded convolutional networks. **IEEE Signal Processing Letters**, [S.l.], v. 23, n. 10, p. 1499–1503, 2016.

ZHANG, L.; MISTRY, K.; JIANG, M.; NEOH, S. C.; HOSSAIN, M. A. Adaptive facial point detection and emotion recognition for a humanoid robot. **Computer Vision and Image Understanding**, [S.l.], v. 140, p. 93–114, 2015.

ZHAO, G.; PIETIKAINEN, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, [S.l.], n. 6, p. 915–928, 2007.

ZHENG, X.; HASEGAWA, S.; TRAN, M.-T.; OTA, K.; UNOKI, T. Estimation of Learners' Engagement Using Face and Body Features by Transfer Learning. In: INTERNATIONAL CONFERENCE ON HUMAN-COMPUTER INTERACTION, 2021. **Anais...** [S.l.: s.n.], 2021. p. 541–552.

ZIMMERMANN, P.; GUTTORMSEN, S.; DANUSER, B.; GOMEZ, P. Affective computing—a rationale for measuring mood with mouse and keyboard. **International journal of occupational safety and ergonomics**, [S.l.], v. 9, n. 4, p. 539–551, 2003.