
*Trabalho de Conclusão de Curso*

# Frequência escolar por meio de reconhecimento facial utilizando Redes Neurais Residuais

**John Davi Dutra Canuto Pires**

**orientado por**
**Prof. Dr. Marcelo Costa Oliveira**

Universidade Federal de Alagoas
Instituto de Computação
Maceió, Alagoas
24 de novembro de 2023

UNIVERSIDADE FEDERAL DE ALAGOAS
Instituto de Computação

# FREQUÊNCIA ESCOLAR POR MEIO DE RECONHECIMENTO FACIAL UTILIZANDO REDES NEURAIS RESIDUAIS

Trabalho de Conclusão de Curso apresentado ao Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação.

John Davi Dutra Canuto Pires

***Orientador: Prof. Dr. Marcelo Costa Oliveira***

**Banca Examinadora:**

| | |
|---|---|
| Marcelo Costa Oliveira | Prof. Dr., IC-UFAL |
| Andressa Martins Oliveira | Prof. Msc., IC-UFAL |
| Erick de Andrade Barboza | Prof. Dr., IC-UFAL |

Maceió, Alagoas
24 de novembro de 2023

**Catalogação na fonte**
**Universidade Federal de Alagoas**
**Biblioteca Central**
**Divisão de Tratamento Técnico**
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

P667f Pires, John Davi Dutra Canuto.
Frequência escolar por meio de reconhecimento facial utilizando redes neurais residuais / John Davi Dutra Canuto Pires. – 2023.
44 f. : il. color.

Orientador: Marcelo Costa Oliveira.
Monografia (Trabalho de Conclusão de Curso em Engenharia da Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2023.

Bibliografia: f. 41-44.

1. Reconhecimento facial. 2. Frequência escolar – Automatização. 3. Redes neurais. I. Título.

CDU: 004.8

**UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL**
**Instituto de Computação – IC**
Campus A. C. Simões - Av. Lourival de Melo Mota, BL 12
Tabuleiro do Martins, Maceió/AL - CEP: 57.072-970
Telefone: (082) 3214-1401

# Trabalho de Conclusão de Curso – TCC

## Formulário de Avaliação

Curso: Engenharia de Computação

Nome do Aluno

| J | O | H | N | | D | A | V | I | | D | U | T | R | A | | C | A | N | U |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | O | | P | I | R | E | S | | | | | | | | | | | | |

Nº de Matrícula

| 1 | 8 | 1 | 1 | 2 | 2 | 0 | 6 | | | | | - | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Título do TCC (Tema)

Frequência automática por meio de reconhecimento facial utilizando Redes Neurais Residuais

Banca Examinadora

| | |
|---|---|
| Marcelo Costa Oliveira<br>Nome do Orientador | Documento assinado digitalmente gov.br MARCELO COSTA OLIVEIRA Data: 24/11/2023 16:20:19-0300 Verifique em https://validar.iti.gov.br |
| Andressa Martins Oliveira<br>Nome do Professor | Documento assinado digitalmente gov.br ANDRESSA MARTINS OLIVEIRA Data: 24/11/2023 16:43:01-0300 Verifique em https://validar.iti.gov.br |
| Erick de Andrade Barboza<br>Nome do Professor | Documento assinado digitalmente gov.br ERICK DE ANDRADE BARBOZA Data: 24/11/2023 18:01:40-0300 Verifique em https://validar.iti.gov.br<br>Assinatura |

Data da Defesa: 24 / 11 / 2023

Nota Obtida: 9,5 ( Nove e meio )

Observações

**Coordenador do Curso De Acordo**

Documento assinado digitalmente gov.br LEANDRO DIAS DA SILVA Data: 27/11/2023 14:05:43-0300 Verifique em https://validar.iti.gov.br

Assinatura

# Agradecimentos

Primeiramente, eu gostaria de agradecer minha bisavó, Lizete Canuto Soares, e minhas avós, Sônia e Terezinha, por acreditarem na minha educação desde o começo. A jornada foi longa e algumas vezes quase impossível, mas não houve um só momento em que eu estivesse sozinho, minha família sempre me apoiou e acreditou em mim, independente do caminho que eu seguisse, agradeço à minha mãe, Rosilane, ao meu pai, Ari Júnior, e aos meus irmãos, Júlio e Juan por me lembrarem que nunca estive só.

Acredito que posso falar sem hesitar que tenho ótimos amigos, aqueles que conheci no começo da minha vida, Marcus e Vitor, outros que conheci no meio dela, Arnaldo, Jefferson, Matheus, Lucas Mattheus, Luís Eduardo, Lyvia, Samile e Raíssa, obrigado por todas as tardes, as risadas e aos aprendizados ao lado de vocês. Devo muito também aos amigos que conheci durante a caminhada na universidade: Luana Júlia, Lucas Massa, Hugo, Mateus Fernando, Bruna, João Pedro, Ruan, Hiago, Jhonnye, Igor, Cabral e Derek; obrigado por estarem ao meu lado em todos os momentos difíceis. Além disso, agradeço meu orientador, professor Marcelo Costa, por não apenas ser uma pessoa atenciosa e prestativa, mas também por me orientar durante todo esse final de curso; e aqueles que ofereceram moradia para mim quando mais precisei, minha Tia Ariana Bruna e a família do Vitor.

Quero agradecer também minha namorada e futura esposa, Luana Soares Lima, que esteve comigo em cada segundo desse trabalho e me apoiou de uma forma que só ela conseguiria, me fazendo crer que sem ela eu talvez não conseguisse passar por isso.

Por fim, quero agradecer àqueles que não foram nomeados e a mim mesmo, por suportar todo esse trajeto e conseguir se manter de cabeça erguida apesar de tudo.

Obrigado.

Maceió, 24 de novembro de 2023.

# Resumo

A prática de registrar e monitorar a frequência dos alunos é uma ação fundamental em diversos contextos, especialmente no ambiente escolar, apesar de frequentemente consumir parte significativa do tempo de aula devido ao seu processo manual. Neste estudo, é proposto o desenvolvimento de um modelo de reconhecimento facial com o propósito de automatizar o controle de frequência escolar. Baseado nas contribuições de King (2009) e nas técnicas de Redes Neurais Convolucionais (CNN), o modelo é projetado para superar os desafios relacionados à defasagem dos dispositivos de captura de imagem e às variações de ambiente na sala de aula.A base de dados é composta por cinco sessões de aula provenientes do estudo conduzido por Mery et al. (2019) e uma sessão de aula coletada em uma unidade escolar do Serviço Social da Indústria (SESI), totalizando quase 100 imagens para análise. Em comparação com os modelos Facenet512, Facenet e ArcFace, em relação à precisão na marcação de frequência final, o modelo apresentou um desempenho superior, com um aumento de cerca de 51% na acurácia em relação aos demais, atingindo uma marca satisfatória de 76% de acurácia, além de demonstrar uma especificidade superior. É importante destacar que, neste estágio de desenvolvimento, o modelo tem o potencial de beneficiar tanto os professores quanto os alunos, contribuindo para economizar tempo e reduzir erros na gestão da frequência escolar.

***Palavras-chave*: Aprendizagem profunda, Visão Computacional, Sistemas Ciberfísicos, Tecnologias na Educação, Reconhecimento facial**

# Abstract

The practice of recording and monitoring student attendance is a fundamental action in various contexts, especially in the school environment, although it often consumes a significant part of class time due to its manual process. This study proposes the development of a facial recognition model with the aim of automating school attendance control. Based on the contributions of King (2009) and Convolutional Neural Network (CNN) techniques, the model is designed to overcome the challenges related to the lag of image capture devices and variations in the classroom environment.The database consists of five class sessions from the study conducted by Mery et al. (2019) and one class session collected at a school unit of the Serviço Social da Indústria (SESI), totaling almost 100 images for analysis. Compared to the Facenet512, Facenet and ArcFace models, in terms of final frequency marking accuracy, the model showed superior performance. With an increase of around 51% in accuracy compared to the others, reaching a satisfactory 76% accuracy mark, as well as demonstrating superior specificity. It is important to note that, at this stage of development, the model has the potential to benefit both teachers and students, helping to save time and reduce errors in school attendance management.

***Palavras-chave*: Deep Learning, Computer Vision, Cyber-Physical Systems, Technologies in Education, Face Recognition**

# Lista de Figuras

# Lista de Tabelas

# Lista de Símbolos

| | |
|---|---|
| $x_i^a$ | Vetor âncora referente a uma imagem |
| $x_i^p$ | Vetor positivo referente a uma imagem |
| $x_i^n$ | Vetor negativo referente a uma imagem |
| $\alpha$ | Margem entre os vetores positivos e negativos |
| $f$ | Função de ativação que reside entre as camadas da rede neural |
| $q_i$ | Coordenada i do ponto q do plano |
| $p_i$ | Coordenada i do ponto p do plano |

# Lista de Abreviações

**ANN** *Artificial Neural Network*

**CNN** *Convolutional Neural Network*

**LFW** *Labeled Faces in the Wild*

**HOG** *Histogram of Oriented Gradients*

**RTSP** *Real Time Streaming Protocol*

**SGD** *Stochastic Gradient Descent*

**SESI** *Serviço Social da Indústria*

**ROC** *Receiver Operating Characteristic*

**AUC** *Area Under the ROC Curve*

**ReLU** *Rectified Linear Unit*

# Sumário

# Capítulo 1

# Introdução

A frequência escolar como atividade está conectada intrinsecamente ao dia-a-dia estudantil e por conta disso, quase não são discutidas novas formas de realizá-la. Existem diversas motivações para a frequência escolar acontecer, não apenas para contabilizar assiduidade dos alunos nas aulas, mas também, em estágios iniciais de desenvolvimento, servir como uma ferramenta socializadora entre os alunos, segundo Thornton (2013), sendo crucial neste estágio.

Segundo a Lei de Diretrizes e Bases da Educação Nacional (LDB) do Governo Federal do Brasil (1996), a quantidade mínima de horas de aula é 800 distribuídas em 200 dias de aula num ano letivo, supondo que em uma sala com 40 alunos são necessários 5 minutos para realizar a frequência diariamente, em uma semana serão gastos 25 minutos de aula, totalizando aproximadamente 17 horas gastas com a frequência escolar num ano letivo. Conforme o estudo feito pela National Center of Education Statistics, NCES (2009), do ponto de vista da coleta de informações sobre os alunos, a frequência escolar não apenas é um indicador de assiduidade, mas também um critério relacionado ao desempenho do aluno, ou seja, quanto maior a frequência, melhores os resultados. Por conta disso, a frequência deixa de ser apenas um registro e se torna uma das informações mais precisas para análise dos alunos e que de acordo com Uskov et al. (2015) pode ajudar a construir a ideia de uma sala de aula do futuro ou *Next Generation Smart Classrooms*.

Este trabalho se dá conta da existência das câmeras de vigilância presentes na sala de aula, que tem o papel de observar atividades suspeitas, ao elevar a responsabilidade do dispositivo dando a ele a capacidade de realizar a frequência escolar, corroborando o conceito de *Smart Classrooms*. Como consequência, a inserção de tecnologias nas câmeras de vigilância favorece o enfraquecimento do panorama onde a autoridade do gestor é questionada e dada à câmera (GZH (2018)), ao entregar ao dispositivo uma funcionalidade pedagógica, atenuando seu tom inquisitivo em sala de aula e evitando que interfira na relação entre professor e aluno. No ano de 2011, a presença de câmeras nas salas de aula era uma realidade em 72% das escolas particulares em todo o Brasil (Branca Nunes (2011)) e se alia ao fato de que 96% dos professores concordam que o uso da tecnologia em

sala de aula amplia as capacidades do professor ao invés de tirar sua autonomia, conforme a Biblioteca Educação Já (2021).

Diante deste contexto, o objetivo principal deste trabalho foi o desenvolvimento um sistema viável baseado nos estudos de King (2009) e Geitgey (2016) que realize o processo de reconhecimento facial no ambiente escolar utilizando apenas os recursos disponíveis nas escolas com o intuito de realizar a frequência escolar de forma automática. Além disso, os esforços deste trabalho também buscam estudar a transformação do ambiente escolar num ambiente ciberfísico que ressignifique o papel das câmeras em sala de aula.

## 1.1 Objetivo geral

O objetivo que move este trabalho é desenvolver um sistema que envolva a detecção e reconhecimento facial por meio de imagens de câmeras de segurança visando realizar automatizadamente a frequência escolar. Além de trabalhar numa solução que se aproxime do patamar de inferência em tempo real, permitindo que a ação da frequência escolar seja instantânea e não-intrusiva.

### 1.1.1 Objetivos Específicos

1. Verificar se a arquitetura apresentada neste trabalho gera bons resultados utilizando, no mínimo, uma única imagem do aluno como base;

2. Analisar o tempo de inferência do modelo a fim de refletir se a frequência pode ser feita em tempo real;

3. Estudar como o modelo trabalhado é afetado pelo qualidade dos dados dispostos a ele;

## 1.2 Organização do trabalho

Este trabalho foi organizado em capítulos que detalham os passos seguidos durante a pesquisa e modelagem do modelo utilizando CNN, detalhando conceitos computacionais e definições matemáticas que servem como base para este modelo.

O Capítulo 2 aprofunda-se em conceitos de redes neurais, nas diferentes propostas de reconhecimento facial, tais quais são Eigenfaces, Fisherfaces e o uso de CNN — e por fim — nos diferentes métodos de processamento digital de imagem utilizados.

O Capítulo 3 detalha a modelagem do sistema de detecção e reconhecimento facial feito usando o poder conjunto de uma CNN e uma versão alternativa da ResNet34, juntamente com os desafios encontrados e a apresentação de detalhes do funcionamento dos modelos pela biblioteca Dlib (King (2009)).

O Capítulo 4 traz a análise e a discussão dos principais resultados da aplicação do modelo no ambiente-alvo, comparando sempre a frequência real com a solução dada pelo modelo de CNN, analisando a forma de aquisição dos dados, a forma que o modelo foi usado e a avaliação da precisão do modelo.

O Capítulo 5 traz os pensamentos finais em relação ao desenvolvimento e aos resultados, discutindo o panorama do modelo em relação à frequência no ambiente escolar e as perspectivas futuras da aplicação.

# Capítulo 2

# Fundamentação Teórica

## 2.1 Trabalhos relacionados

Seguindo as tendências atuais da tecnologia da informação, as soluções automáticas de reconhecimento facial estão se tornando uma opção padrão em muitas áreas, inclusive na educação (Zhang et al. (2022)). O principal problema com esses métodos é a necessidade de dispositivos externos, que às vezes dependem da ação do aluno para funcionar e consomem tempo ao adicionarem mais camadas de logística para a ação da frequência. A maioria dos sistemas modernos depende muito da visão computacional e dos algoritmos de aprendizado de máquina; essas soluções podem ser mais flexíveis e reduzir o erro humano.

O estudo realizado por Serengil and Ozpinar (2020) apresenta uma revisão sistemática relacionando diversas soluções de reconhecimento facial e unindo-as num único *framework*, o *DeepFace*, visando tornar estes modelos de reconhecimento facial acessíveis para estudo ao sintetizarem todas as soluções para utilizar menos recursos computacionais. Os modelos de reconhecimento facial contém quatro estágios: detectar, alinhar, representar e verificar; o trabalho citado reduz a complexidade dos modelos acima ao retirar a responsabilidade dos modelos para com a detecção, alinhamento e verificação, tornando-os apenas responsáveis pela representação. Nos modelos que estão presentes estão: VGG-Face (Parkhi et al. (2015)), FaceNet (Schroff et al. (2015)), OpenFace (Baltrušaitis et al. (2016)), DeepFace (Taigman et al. (2014)), DeepID (Sun et al. (2014)) e Dlib (King (2009)). Como resultado, o trabalho cria com sucesso um*framework* leve, que possibilita diversos estudos comparativos, além de trazer um conjunto de soluções no estado-da-arte.

O modelo com maior acurácia segundo o trabalho acima, o FaceNet (Schroff et al. (2015)), desenvolvido pela Google Inc. (Schroff et al. (2015)) propõe uma arquitetura baseada em CNN utilizando como fundamentos o modelo feito por Zeiler and Fergus (2013) e as redes *Inception* baseadas no trabalho de Szegedy et al. (2014). A sua rede neural é constituída por uma CNN que recorre a dois otimizadores — SGD e *AdaGrad* — além do uso da *Triplet Loss* como poderio de verificação facial e o cálculo de distância

euclidiana como medidor de similaridade entre os rostos. Seus resultados mais expressivos foram obtidos utilizando a base LFW (Huang et al. (2008)) e diante de duas formas de alinhamento facial, as acurácias obtidas foram de $98.87\% \pm 0.15$ e $99.63\% \pm 0.09$, para os dados puros da base e para dados utilizando um detector facial similar ao Picasa (Chen et al. (2014)), respectivamente.

Na pesquisa de King (2009) é apresentada uma solução *cross-platform* que combina diversas ferramentas de *machine learning* e álgebra linear com o intuito de diminuir a complexidade de uso e aumentar a facilidade de integração entre ferramentas. Após anos, Geitgey (2016) trouxe uma nova abordagem para a ferramenta ao criar um framework utilizando uma versão alternativa da ResNet-34 (He et al. (2015)), que alcançou 99.38% de acurácia utilizando a LFW (Huang et al. (2008)) como base de dados. No entanto, as imagens que compõem esta base de dados, utilizada por Schroff et al. (2015) e King (2009), são imagens nítidas e ideias, tiradas em momentos propensos à boa qualidade de imagem, não refletindo a situação de um ambiente real, tal como uma sala de aula.

## 2.2 Aprendizado de máquina

Inteligência Artificial (IA) é a capacidade que sistemas têm de resolverem problemas complexos encontrando relações e padrões diante de observações, é baseada em modelos analíticos que geram predições, regras, recomendações ou alguma outra resposta similar. As primeiras tentativas de criar modelos analíticos sustentados apenas por relações conhecidas, de forma procedural e envolvendo decisões lógicas foram chamadas de sistemas especialistas (Russell and Norvig (2010)).

Na atualidade, com a presença de novos *frameworks*, com a disponibilidade de grandes volumes de dados e amplo acesso ao poder computacional, modelos analíticos são construídos utilizando a chamada "Aprendizagem de máquina" (do inglês *Machine Learning* (ML)) (Goodfellow et al. (2016)). A aprendizagem de máquina é sustentada pela ideia de substituir o papel da formalização do conhecimento de forma acessível para uma máquina, permitindo assim o desenvolvimento de sistemas mais eficientes.

Durante as últimas décadas, a área de estudo do aprendizado de máquina trouxe uma variedade enorme de avanços em algoritmos de aprendizado e técnicas de pré-processamento eficientes. Um desses avanços foi a evolução das redes neurais artificiais em direção às arquiteturas de redes neurais profundas com capacidades de aprendizado melhoradas, resumidas no termo aprendizado profundo (Goodfellow et al. (2016)). A figura 2.1 torna visual a hierarquia entre as áreas citadas acima.

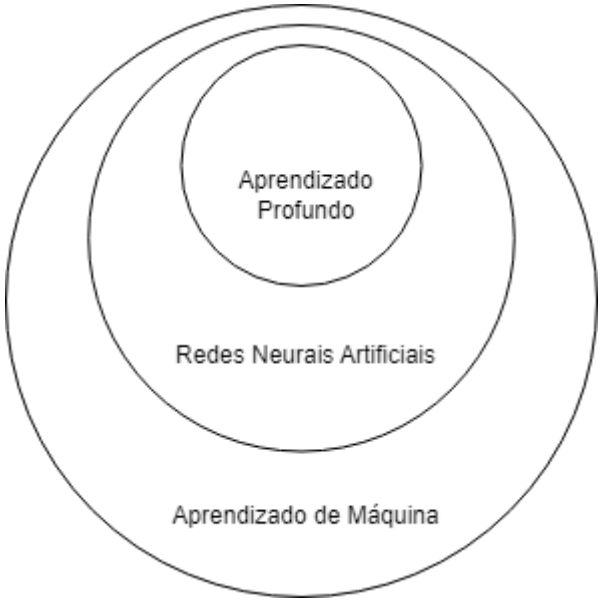


Figura 2.1: Representação em diagrama da hierarquia na área de aprendizado de máquina. Modificado de Goodfellow et al. (2016).

### 2.2.1 Redes Neurais e Aprendizado Profundo

As redes neurais artificiais (do inglês *Artificial neural networks* (ANN)) são um tipo de processo de *machine learning* que consiste em modelos de computação que se baseiam na estrutura neural do cérebro (Anderson and McNeill (1992)). De forma mais específica, esse processo usa nós ou neurônios interconectados capazes de aprender com os erros e se aprimorar continuamente usando cálculos simples; o uso de diversas camadas de neurônios interconectados tem como nome aprendizado profundo ou *deep learning* (figura 2.2).

Um nó ou neurônio representa a unidade primordial de processamento de informações necessária para o correto funcionamento de uma rede neural. Dentro desse contexto, o neurônio realiza a computação da soma ponderada dos sinais de entrada, considerando seus pesos associados. O resultado desta soma é então submetido à ação de uma função de ativação, cujo resultado é distribuído para todas as saídas pertinentes ao neurônio em questão (Haykin (2009)).

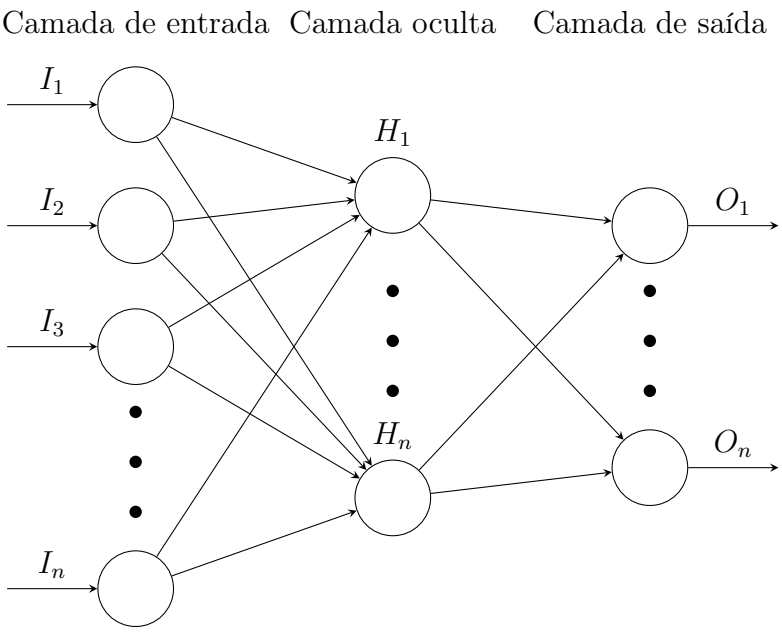


Figura 2.2: Representação gráfica de uma rede neural *feedforward* chamada *multilayer perceptron*. Fonte: autor, 2023

O intuito das redes neurais (RN) na totalidade é analisar uma quantidade finita de dados e aprender os padrões presentes neles por meio da regulagem dos valores dos pesos presentes nos neurônios para que no fim a RN se torne generalista e consiga extrair saídas satisfatórias a partir das entradas fornecidas. Por conta dessa capacidade de generalização de problemas, segundo Chua and Yang (1988), as redes neurais são propostas para resolver problemas em diversas áreas, como otimização, programação linear e não linear, reconhecimento de padrões e visão computacional.

Por fim, cabe dizer que as redes neurais, em específico os neurônios, podem ter diferentes tipos formas de interpretar o resultado de seu cálculo, diante disso, existem diversas funções incumbidas de modificarem esses valores e torná-los coerentes com o problema proposto, estas funções são chamadas de funções de ativação. Entre as funções de ativação mais comuns estão a função sigmoide, a função ReLU e a função tangente hiperbólica (Sharma et al. (2017)).

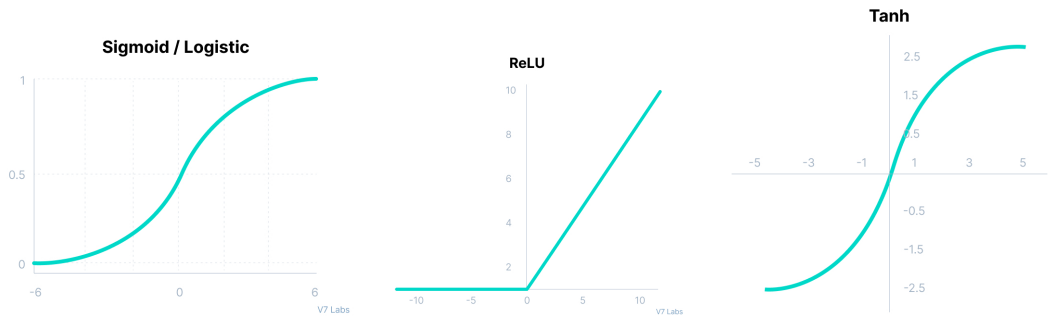


Figura 2.3: Gráficos contendo as representações das funções sigmoide, ReLU e tangente hiperbólica. Fonte: Baheti (2021)

### Otimizadores

De acordo com Taqi et al. (2018), uma rede neural necessita que as suas variáveis, presentes nas camadas, sejam alteradas de forma que o modelo funcione melhor no processo de classificação. Uma das formas de se obter essa melhoria nos valores das variáveis é por meio dos otimizadores, que funcionam com o intuito de buscar o decréscimo dos valores de perda obtidos pelo modelo, realizando cálculos de minimização encontrando o gradiente descendente daquela função.

Uma das formas de alcançar o valor mínimo global para aquela função é a estocástica, que consiste na variação aleatória de valores, compondo o algoritmo *Stochastic Gradient Descent* (SGD) (Sutskever et al. (2013)). O SGD executa uma atualização de parâmetros aleatória para cada amostra de treino e rótulo, tornando ele geralmente um algoritmo muito rápido que pode ser usado para aprendizado online. Um parâmetro presente no algoritmo SGD é o *momentum*, um método que ajuda a acelerar o método SGD a encontrar a direção correta de convergência mais rapidamente e apresentar menos oscilações.

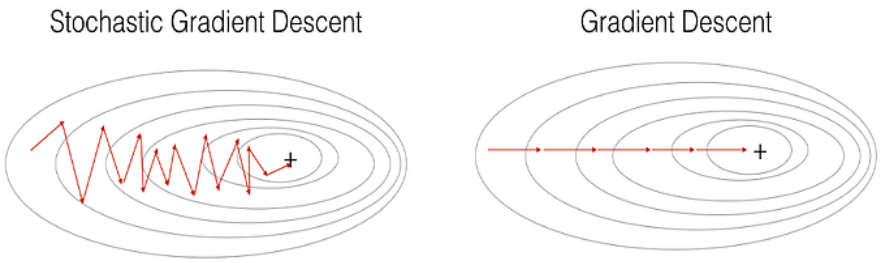


Figura 2.4: Esquematização visual da busca por parâmetros feita pelo algoritmo SGD em comparação à busca comum pelo mínimo global. Fonte: Goyal (2021)

### 2.2.2 Visão Computacional e Reconhecimento facial

Visão computacional é a área de estudo das ciências de computação que lida com a criação de um sistema análogo ao olho humano em toda sua completude, em outras palavras, é o estudo que visa permitir que os computadores identifiquem e processem objetos em imagens e vídeos da mesma forma que humanos fazem.

Uma das formas de se utilizar a visão computacional é como classificadora de imagens, que consiste no processo de prever uma classe ou uma *label* específica através da categorização de grupos de píxeis e vetores no contexto de uma imagem. O aprendizado pode ser de forma supervisionada ou não supervisionada, onde, o algoritmo aprende padrões por si só e aprende utilizando amostras previamente classificadas, respectivamente.

No domínio das primeiras técnicas de reconhecimento facial, tais como aquelas que utilizam abordagens de extração de características, como a Análise de Componentes Principais (PCA), uma das técnicas mais utilizadas é a Eigenfaces (Wahyu Mulyono et al. (2019); Khan et al. (2019)) (figura 2.5). Essa técnica não possui muita complexidade, com um bom desempenho quando as condições de iluminação são ideais e há variações mínimas nas expressões faciais. Entretanto, objetos com elevado contraste de cor podem dificultar o reconhecimento.

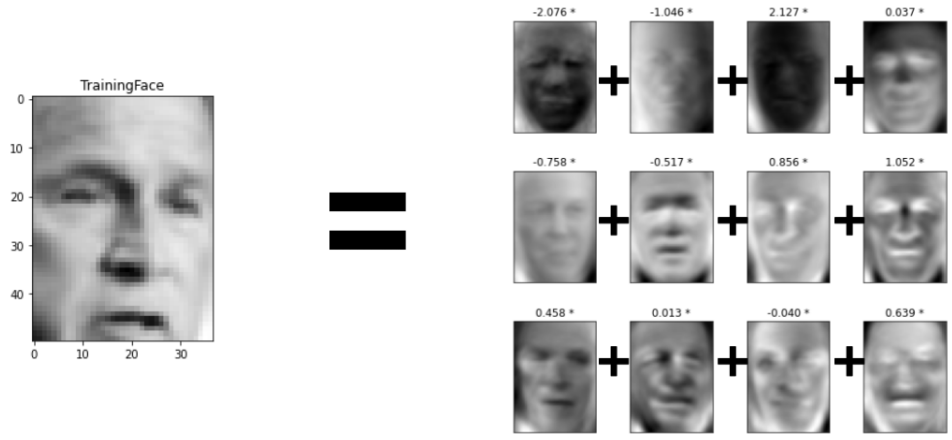


Figura 2.5: Disposição de um rosto e suas Eigenfaces à direita. Fonte: GeeksforGeeks (2021).

Os métodos Eigenface e Fisherface (Souza (2014)) utilizam a PCA e a análise discriminante linear (LDA), respectivamente, como base. Isso difere do *Local Binary Pattern* (LBP) e do *Local Binary Pattern Histogram* (LBPH), os quais são métodos baseados na criação de um histograma criado por subconjuntos de pixeis filtrados por uma matriz. Sendo o número de histogramas gerados a principal diferença entre o LBP e o LBPH (figura 2.6).

O principal problema do LBP e do LBPH (Deeba et al. (2019)) é a flexibilidade, o mesmo problema enfrentado pelas abordagens Eigenfaces e Fisherfaces. No entanto, esse

problema pode ser facilmente resolvido por métodos semelhantes, como a restrição da distância de Hamming (Yang and Wang (2007)), que lida estritamente com faces frontais e cenários ideais.

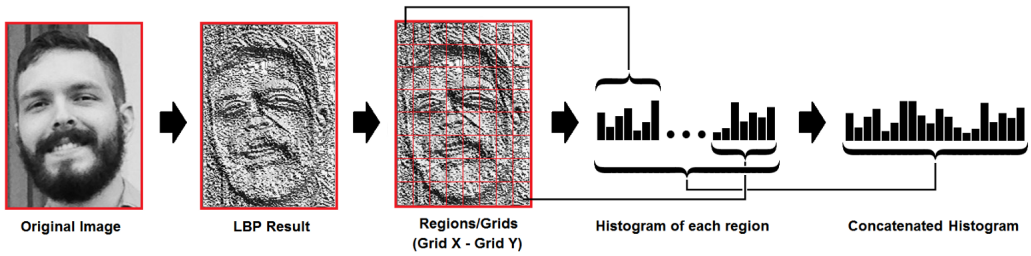


Figura 2.6: Esquematização dos passos que comandam o processo do algoritmo LBPH. Fonte: Salton (2017).

O que une todos esses métodos é o objetivo de mapear o rosto humano em características relevantes, o *facial landmarking*, sendo o processo de gerar um valor digital que mapeia uma estrutura baseada em seus pontos fiduciais (olhos, boca e nariz). Porém, diante dos estados e dos ambientes onde o ser humano está inserido, o mapeamento pode ser gerado com menos eficácia.

Portanto, a maior dificuldade enfrentada por todos os métodos acima citados é a flexibilidade, o qual é considerado o ponto-chave das redes neurais aliadas ao aprendizado profundo, pois com a presença de diferentes camadas e combinações com funções de ativação, o modelo final formado pela rede neural pode abarcar diversos cenários num só algoritmo. Para isso, se torna necessário entender como as CNN funcionam, visto que são estas que lidam com imagens e consequentemente, vídeos.

### 2.2.3 Detecção de objetos

A obtenção de um entendimento completo do conteúdo de um conjunto de imagens vai além da simples classificação com base em características específicas. Em muitos casos, é fundamental estimar a localização dos objetos presentes em cada imagem, caracterizando a tarefa conhecida como "detecção de objetos" (Zhao et al. (2019)). Na Figura 2.7, podemos visualizar um exemplo do resultado desejado ao utilizar um algoritmo de detecção de objetos. Essa abordagem não apenas identifica os objetos nas imagens, mas também fornece informações precisas sobre sua localização espacial, o que é essencial para uma compreensão mais completa e detalhada do conteúdo visual.

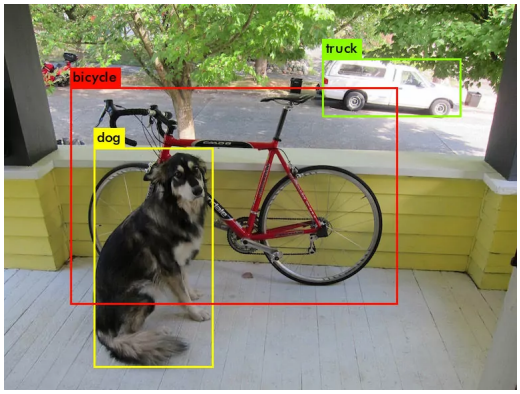


Figura 2.7: Demonstração de um problema típico de detecção de entidades. Fonte: Zhao et al. (2019).

O processo de detecção de objetos pode ser dividido em três etapas básicas (Zhao et al. (2019)):

- Seleção da região informativa — refere-se à decisão de escolher a posição e o formato da área que contém um objeto de interesse em uma imagem. Essa tarefa é desafiadora devido à variabilidade na localização, tamanho e formato dos objetos na imagem. Uma abordagem ineficiente seria percorrer toda a imagem usando filtros de diferentes tamanhos e formas, devido à complexidade computacional envolvida. Por outro lado, aplicar um número limitado de filtros pode resultar em resultados imprecisos. Encontrar o equilíbrio adequado entre eficiência computacional e precisão na seleção da região informativa é um desafio crucial nesse contexto.

- Extração de características — essa etapa lida com a extração e reconhecimento das características presentes nos objetos, a fim de criar uma representação robusta deles. A dificuldade primordial reside nas variações nas condições de iluminação e no contexto (plano de fundo) em que o objeto pode estar inserido, tornando desafiador o desenvolvimento de um extrator de características capaz de descrever o objeto de maneira eficaz e consistente. perfeitamente qualquer objeto em qualquer situação.

- Classificação — essa fase envolve o uso de um algoritmo que, com base na região informativa e nas características extraídas, consegue determinar a qual classe específica, de um conjunto pré-definido de categorias, o objeto pertence. Em qualquer problema de detecção de objetos, é necessário pelo menos definir duas classes distintas, sendo uma delas correspondente ao plano de fundo da imagem e a outra à classe de interesse, que abrange os objetos que se deseja identificar. No contexto do problema abordado neste trabalho, essas duas classes essenciais são o plano de fundo e a classe de interesse, sendo o foco principal da detecção de objetos.

**Detecção voltada para frequência escolar**

Essencialmente, a tarefa de criar um sistema de verificação facial eficiente e em grande escala é uma tarefa de projetar funções de perda apropriadas que melhor detectem as classes escolhidas. Por conta disso, o processo de reconhecimento facial normalmente envolve receber como entrada uma imagem e retornar um rótulo para aquela imagem indicando a localização de uma determinada pessoa, criando uma instância da rede para cada rosto, tornando-o um modelo obsoleto para rostos fora do escopo ou em estados diferentes do entregue para o modelo, em outras palavras, extraindo padrões de rosto de um determinado ambiente, nesse caso, o escolar.

Para contornar essa situação, este trabalho recorre ao *Deep Metric Learning*, que consiste em usar redes neurais para aprender automaticamente características de discriminação das imagens e, em seguida, calcular a métrica que melhor representa aquele objeto detectado na imagem (Zhao et al. (2021)). Visando facilitar o reconhecimento cruzado, pois ao determinar um vetor conhecido para o rosto de um aluno e outro para o rosto encontrado em sala de aula, para todos as amostras seguintes, resta apenas calcular as operações de similaridade (figura 2.8).

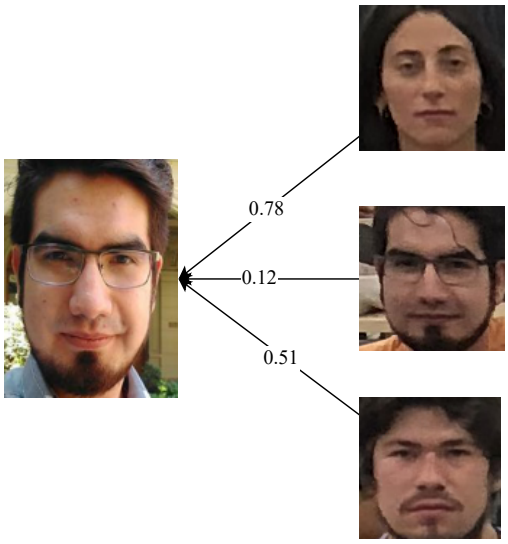


Figura 2.8: Exemplificação de como o mesmo rosto, por ter um valor matemático único, pode ser comparado diversas vezes. Fonte: autor, 2023.

### 2.2.4 Redes Neurais Convolucionais

Do inglês *Convolutional Neural Networks* (CNNs), as CNNs são um tipo específico de ANN, proposta pelo pesquisador francês Lecun et al. (1998). As CNNs se mostraram, desde a sua criação, serem muito eficazes para resolver problemas de classificação, se mostrando uma alternativa viável aos métodos tradicionais para esse tipo de problema.

Uma das desvantagens da CNN é o fato de existir a necessidade de uma abundância

de dados rotulados para a extração dos padrões, ou features. Basicamente, para extração dessas *features*, podem existir três componentes básicos em uma CNN, camada de convolução, *pooling* e totalmente conectada. Qualquer arquitetura básica é composta por esses blocos. Um exemplo da rede criada por Yann LeCun, a LeNet pode ser visto abaixo (figura 2.9):

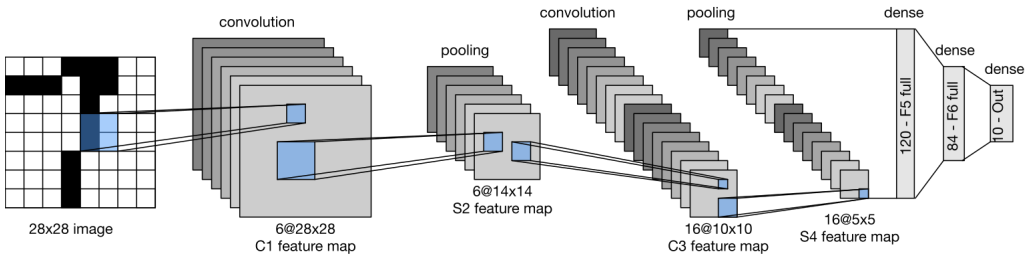


Figura 2.9: Arquitetura LeNet proposta por Yann LeCun. Fonte: Zhang et al. (2021).

### 2.2.5 Métricas

As métricas são um protocolo de avaliação de desempenho baseado em predições corretas e incorretas e suas disposições num processo supervisionado. Neste trabalho foram utilizadas, além da curva ROC e do valor AUC, os métodos consequentes da matriz de confusão: acurácia, precisão, *recall* e *F1-Score*.

#### Matriz de confusão

A matriz de confusão é uma ferramenta voltada para a visualização de problemas de classificação. Ela é composta por uma tabela bidimensional que ilustra o número de predições corretas e incorretas em cada classe, fornecendo informações detalhadas sobre o desempenho de um classificador num conjunto de dados. Com base na matriz de confusão, é possível examinar, quantitativamente, quais as classes o algoritmo de aprendizado tem maior dificuldade em classificar (Russell and Norvig (2010)).

A matriz é baseada no número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Sendo considerando o verdadeiro positivo ($VP$) o elemento positivo classificado como positivo; como verdadeiro negativo ($VN$) o elemento negativo que é negativo sendo classificado corretamente como negativo; como falso positivo ($FP$) o elemento falso que é definido como positivo; e, por fim, o falso negativo ($FN$) é aquele que é positivo, mas é classificado como negativo (figura 2.10).

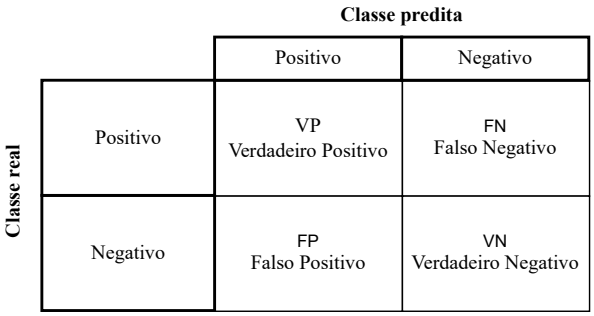


Figura 2.10: Exemplo da estrutura de uma matriz de confusão. Fonte: autor, 2023.

Partindo da ideia estabelecida acerca da matriz de confusão, algumas métricas podem ser geradas, neste trabalho os resultados serão medidos com base nas seguintes métricas:

- Acurácia: o propósito dela é estimar a quantidade de classes classificadas corretamente, para que isso aconteça é feita a divisão de todos os acertos pelo total de classes (da Fontoura Costa and Cesar (2000)).

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.1}$$

- Precisão: é um indicador que gera a relação entre as previsões positivas realizadas corretamente com o somatório de todas as previsões positivos, incluindo as falsas (da Fontoura Costa and Cesar (2000)).

$$Precisão = \frac{VP}{VP + FP} \tag{2.2}$$

- *Recall*: faz um paralelo entre a quantidade de amostras classificadas corretamente como positivas e aquelas de fato positivas. (da Fontoura Costa and Cesar (2000)).

$$Recall = \frac{VP}{VP + FN} \tag{2.3}$$

- *F1-Score*: é uma relação entre a precisão e o *recall* por meio da média harmônica de seus valores. (da Fontoura Costa and Cesar (2000)).

$$F1 - Score = \frac{Precisão \times Recall}{Precisão + Recall} \tag{2.4}$$

- ROC: é essencialmente um gráfico que mostra o desempenho de um modelo de classificação em todos os limiares de classificação considerando dois parâmetros: a taxa de verdadeiro positivo e a taxa de falso positivo (Google (2022)).

$$TVP = Recall = \frac{VP}{VP + FN} \, e \, TFP = \frac{FP}{FP + VN} \tag{2.5}$$

- AUC: é a medição de toda a área bidimensional abaixo da curva ROC inteira de (0,0) a (1,1), oferecendo uma medida agregada de desempenho em todos os limites de classificação possíveis. (Google (2022)).

### 2.2.6 Redes Residuais

As redes neurais profundas trabalham naturalmente com a intuição de que o processamento de um dado exige a participação de todas as camadas e a realização de seus devidos cálculos, gerando um determinado resultado. Seguindo esse princípio, a adição de mais camadas resultaria num melhoramento do aprendizado de forma proporcional e direta (figura 2.11).

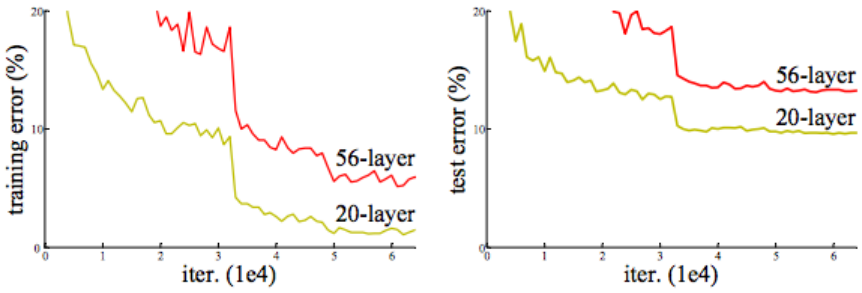


Figura 2.11: Gráficos comparando o progresso do treino e teste entre uma CNN com 20 camadas e outra com 56. Fonte: He et al. (2015).

No entanto, He et al. (2015) e seu trabalho apresentam este gráfico, desafiando a ideia de que adicionar mais camadas criaria uma função mais complexa e, portanto, a falha seria atribuída ao excesso de ajuste. Se esse fosse o caso, parâmetros e algoritmos de regularização adicionais, como *dropout* ou normalização L2, seriam uma abordagem bem-sucedida para corrigir essas redes. No entanto, o gráfico mostra que o erro de treinamento da rede de 56 camadas é maior do que o da rede de 20 camadas, destacando um fenômeno diferente que explica sua falha.

A falha da CNN de 56 camadas pode ser atribuído à função de otimização, à inicialização da rede ou ao problema do gradiente de desaparecimento, ou do inglês *Vanishing Gradient*. Observando o desempenho das redes, He et al. (2015) define uma arquitetura que contorna possíveis situações onde o problema citado acontece (figura 2.12), o bloco residual:

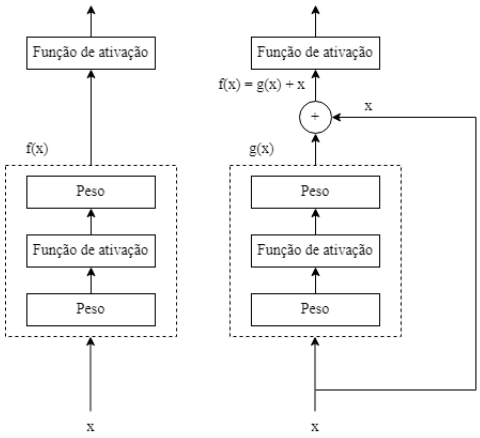


Figura 2.12: Diferença entre a estrutura por trás do bloco direto de uma rede neural e do bloco residual de uma rede neural residual. Modificado de Zhang et al. (2021).

Dado que haja uma operação $F(x)$ entre camadas, que $x$ seja a identidade, o dado original, e que $G(x)$ seja o resíduo da operação, ou seja,

$$G(x) \equiv F(x) - x \tag{2.6}$$

Se o mapeamento desejado é

$$F(x) \equiv x \tag{2.7}$$

Tem-se que

$$G(x) \equiv 0 \tag{2.8}$$

Portanto, dessa forma, se torna mais fácil para o modelo aprender, pois nesse caso não existe resíduo, tornando a função no caso citado, a própria função identidade. Seguindo a figura 2.12, no bloco residual existe um salto que leva da identidade $x$ para o resultado $F(x)$, o que acontece é — ao invés do modelo aprender a função $F(x)$ e otimizá-la, a rede aprende

$$G(x) \tag{2.9}$$

Por fim, ao seguir esse caminho, o modelo lida com uma otimização simples, uma vez que os gradientes são menos propensos a desaparecer ou explodir, pois o modelo está visando ajustar apenas as pequenas diferenças entre a entrada e a saída desejada e não mais toda a entrada. Enquanto métodos mais leves como Eigenfaces e *Fisherfaces* tendem a sacrificar precisão em favor da eficiência computacional, os modelos baseados em CNN oferecem alta precisão, mas frequentemente demandam recursos computacionais substanciais. Portanto, é essencial desenvolver uma abordagem que mantenha o equilíbrio

ideal entre tamanho e desempenho, evitando problemas como o "*Gradient Vanishing*" (Pascanu et al. (2012)). Nesse contexto, a arquitetura que se destacou foi a *Residual Network* (ResNet), especificamente a variante da ResNet34 (He et al. (2015)) utilizada por Geitgey (2016), a ResNet29, que demonstrou conseguir lidar eficazmente com esse panorama juntamente com o desafio de criar uma rede leve e precisa para reconhecimento facial.

# Capítulo 3

# Metodologia

Dada a existência de métodos de reconhecimento facial, que variam desde abordagens mais leves, como Eigenfaces (Wahyu Mulyono et al. (2019)) e *Fisherfaces* (Souza (2014)), até métodos mais robustos, como as Redes Neurais Convolucionais (CNN) (Lecun et al. (1998)), é crucial encontrar um equilíbrio entre a eficiência computacional e a precisão na execução. Este desafio torna-se particularmente relevante ao abordar tarefas de grande escala e concorrentes, como o monitoramento da frequência escolar.

## 3.1 Processo de frequência automática

O processo de obtenção da frequência automática feito neste trabalho é baseado ser apenas uma ferramenta colocada entre o banco de dados e ele mesmo, em outras palavras, ela trabalha com a obtenção dos dados e a devolução do resultado da chamada na mesma fonte. Este processo acontece em 4 etapas:

- Obtenção dos dados: Nesta etapa é onde os dados utilizados nas próximas etapas são obtidos, sejam eles do banco de dados ou das câmeras em sala de aula. Além disso, eles são padronizados em determinadas dimensões a fim de manter um padrão que equilibre qualidade e espaço de armazenamento.

- Detecção facial: Este estágio acontece recorrendo a uma CNN, a *Max-Margin Object Detector* (MMOD), e ocorre apenas nos dados das câmeras das salas de aula, por ser a fase em que há a detecção de objetos, onde os rostos presentes na filmagem são encontrados e separados para que futuramente sejam vetorizados.

- Vetorização facial: Todo e qualquer rosto destacado passa por esta fase, pois após os dados serem obtidos, os rostos dos alunos e os rostos encontrados pela MMOD são vetorizados ao serem entregues a rede principal deste sistema, a ResNet29. Gerando assim, uma transformação dos dados tridimensionais em dados unidimensionais, facilitando o desenvolvimento de cálculos.

- Verificação de similaridade: Por último, o estágio que define se há presença ou não, acontece por meio do cálculo utilizando a equação da distância euclidiana entre os vetores transformados na etapa anterior. A verificação ocorre ao comparar cada rosto dos alunos com todos os outros encontrados no *frame* capturado da câmera.

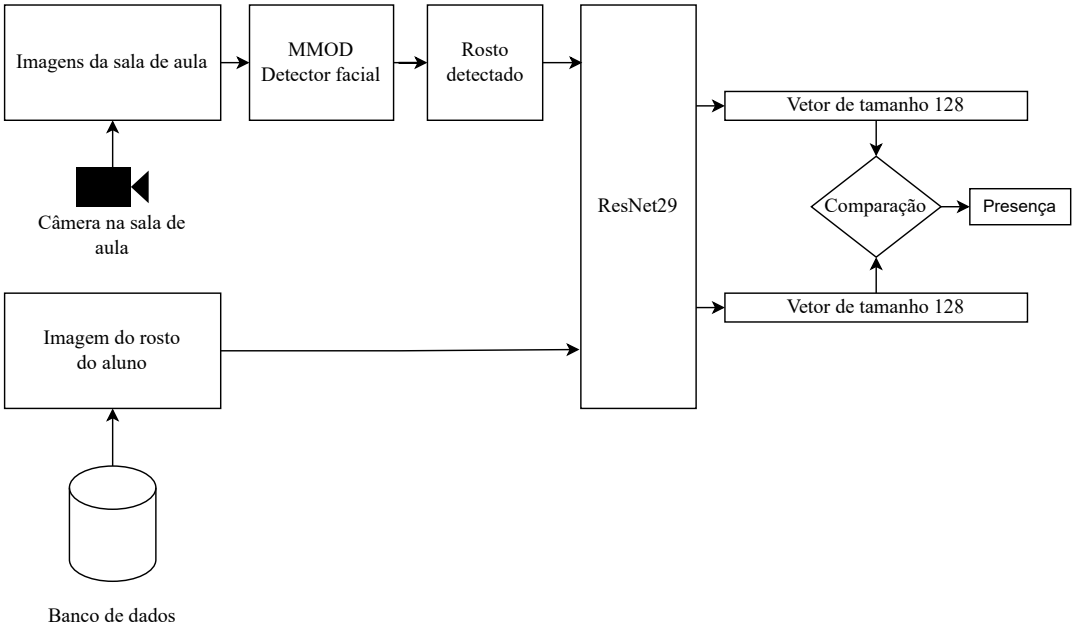


Figura 3.1: Esquemática contendo todos os processos necessários para a execução da frequência automática. Fonte: autor, 2023.

Portanto, tal qual a figura 3.1, anteriormente exibida, mostra, a sequência de funcionamento deste modelo baseia-se em gerar um vetor de características (ou *embeddings*) para os rostos desconhecidos, encontrar rostos na imagem a ser detectada, gerar vetores para estes rostos desconhecidos e comparar com os rostos conhecidos.

### 3.1.1 Composição dos dados

Independente da capacidade do modelo de detectar e reconhecer os estudantes, é majoritariamente necessário analisar o processo inicial do modelo, os dados. Esse cenário foi montado a partir de um conjunto de imagens organizados em 5 aulas coletadas por Mery et al. (2019), juntamente com dados coletados do SESI, como mostra a tabela 3.1.

| Fonte | Sessões | Número de alunos | Tamanho da imagem da sala de aula |
|---|---|---|---|
| SESI | 1 | 34 | 2304×1296 |
| Mery et al. (2019) | 5 | 67 | 2000×2000 |

Tabela 3.1: Tabela contendo a composição dos dados das duas fontes analisadas neste trabalho. Fonte: autor, 2023.

Para a realização das ações de detecção e reconhecimento por parte do modelo para a realização da frequência, foi feita a estruturação das imagens em dois tipos: imagens dos alunos e imagens das câmeras. As imagens dos alunos são organizadas para incluir pelo menos 1 imagem contendo o rosto do aluno, sendo esse valor proporcional à qualidade, dado que mais variações no rosto permitem um maior leque para o reconhecimento, enquanto as imagens das câmeras, utilizadas no formato de vídeo, são recebidas pelo modelo através do protocolo RTSP referente às câmeras de vigilância presentes na sala de aula. A figura 3.2 exibe a forma com que os dados dos rostos dos alunos são padronizados para o modelo.

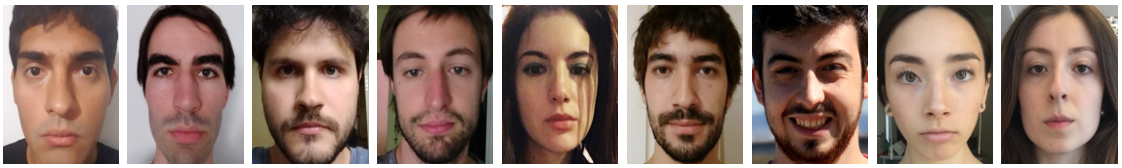

Figura 3.2: Composição das imagens correspondentes aos rostos dos alunos no formato mais favorável ao modelo, isolando o rosto, tendo sua dimensão padronizada no valor de 200×200. Fonte: Mery et al. (2019)

Por conseguinte, as imagens da câmera podem se portar de diversas formas, caso priorizem a visualização dos rostos da turma em sua totalidade. A composição dos dados não precisa ser necessariamente através do protocolo RTSP, podendo ser também um arquivo bruto de vídeo ou um leque de *frames*, como a figura 3.3.

Figura 3.3: Captura da sala de aula utilizando um *smartphone* com resolução 3024×4032. Fonte: Mery et al. (2019)

Os dados utilizados nos testes são compostos por duas fontes: SESI e o estudo feito por Mery et al. (2019). Ambas as fontes contém as fotos dos estudantes e imagens da sala de aula, sendo os dados do estudo de Mery et al. (2019) capturados pela câmera de um *smartphone*, mais precisamente um *iPhone 8* com iOS 11.2.6 e 12 MP e os dados do SESI originados das câmeras de vigilância presentes na sala de aula, câmeras essas IP L42 da Clear CFTV com 2 MP, usando uma turma como teste para o modelo.

### 3.1.2 Detecção facial

O primeiro passo no processo de reconhecimento facial é encontrar as faces presentes na imagem a ser analisada. Dentre os diversos métodos disponíveis para detecção facial, como HOG (Dalal and Triggs (2005)) e o *framework* Viola-Jones (Viola and Jones (2001)), mesmo que apresentem resultados satisfatórios em relação a detectar faces, não apresentam flexibilidade em diferentes situações de pose e iluminação. O método escolhido foi uma CNN, mais especificamente o modelo *Max-margin Object Detection* desenvolvido baseado no estudo de King (2015), que retorna uma matriz contendo as *bounding boxes* delimitando os rostos encontrados, as classes detectadas e a pontuação de confiança. A figura 3.4 exibe a estrutura do modelo de detecção facial.

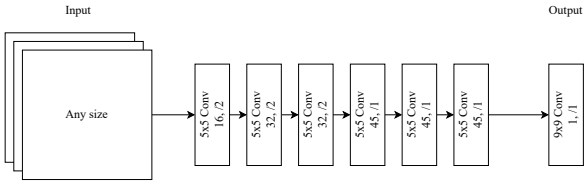


Figura 3.4: Estrutura da rede neural convolucional que realiza a detecção facial. Fonte: autor, 2023

O modelo apresentado na figura 3.4 foi pré-treinado e contém 7 camadas convolucionais, sendo 3 voltadas para *downsampling*, com o intuito de diminuir a dimensionalidade das imagens de entrada, e 4 camadas comuns de convolução que culminam na entrega das coordenadas de localização na imagem para classe encontrada e uma métrica auxiliar de acurácia ao detectar.

### 3.1.3 Reconhecimento facial

O *Deep Metric Learning* diz respeito às técnicas de obtenção de um *embedding* para determinado objeto. Diante do exposto, o modelo desenvolvido neste trabalho realiza a transformação das imagens em vetores de 128 características, compondo um mapeamento do rosto da pessoa presente no dado de entrada. Como processo de diferenciação para corroborar a capacidade de assimilação do modelo, o treinamento funciona com a análise de 3 imagens de rostos por vez:

1. Carrega uma imagem de rosto de treinamento de uma pessoa conhecida;
2. Carrega outra imagem da mesma pessoa conhecida;
3. Carrega uma imagem de uma pessoa totalmente diferente.

Em seguida, o algoritmo analisa as medições que está gerando atualmente para cada uma dessas três imagens e ajusta ligeiramente a rede neural de modo a garantir que as medidas geradas para a n.º 1 e a n.º 2 estejam ligeiramente mais próximas e que as medidas para a n.º 2 e a n.º 3 estejam ligeiramente mais distantes, para que isso acontecesse, foram feitas modificações na rede ResNet34 (He et al. (2015)), criando uma ResNet com 29 camadas e que recorre à *Triplet Loss*, sendo este o modelo principal desta aplicação.

#### Implementação do modelo: ResNet29 e *Triplet Loss*

A arquitetura das ResNets estão dispersas entre diversas combinações de camadas e blocos residuais, sendo as mais notáveis suas versões com 34, 50 e 101 camadas. A estruturação do modelo apresentado neste trabalho foi feita utilizando como base a ResNet34,

realizando a remoção de algumas camadas e metade dos filtros por camada com o intuito de tornar ainda menos custoso o treinamento e consequentemente a inferência.

Como pode ser visto na figura 3.5, a arquitetura contém 14 blocos residuais, que se entrelaçam em saltos para ser feito o uso do cálculo do resíduo:

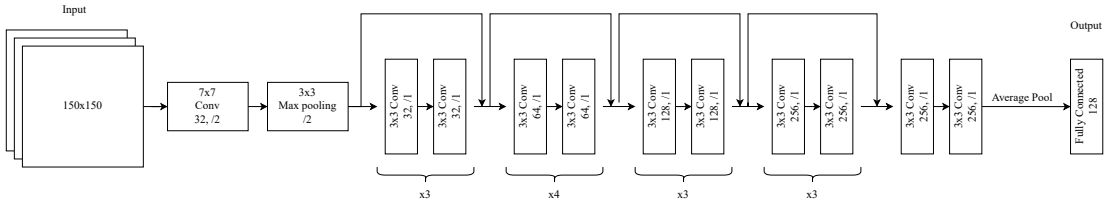


Figura 3.5: Estrutura da rede neural convolucional que realiza o reconhecimento facial, a ResNet29. Fonte: autor, 2023

O modelo em questão foi treinado com a base LFW (Huang et al. (2008)), que abarca 3 milhões de imagens que contém ao todo 7845 rostos individuais e é fortemente usada em processos de *face verification*. A etapa de treino deste modelo foi executada com seus pesos inicializados aleatoriamente e valor de *learning rate* em 0.1 aliado a um *momentum* de 0.9 recorrendo ao otimizador SGD (Sutskever et al. (2013)) e para cada iteração, esse *learning rate* diminui em 10% até que chegue no valor de *learning rate* de 0.0001. Com essa configuração, foi obtido um valor de acurácia de $99,3833\%$ com desvio padrão de $0,0272732$.

Para que esse valor de acurácia fosse alcançado, foi feito o uso de uma função de perda onde o ponto crucial dela fosse minimizar a distância euclidiana dos *embeddings* da mesma pessoa e, ao mesmo tempo, maximizar a mesma distância entre os *embeddings* de pessoas diferentes, esta função de perda é chamada de *Triplet Loss*. A *triplet loss* é definida pela seguinte estrutura apresentada na equação 3.1:

$$\|x_i^a - x_i^p\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2, \forall (x_i^a, x_i^p, x_i^n) \in \tau \tag{3.1}$$

Sendo a equação 3.2 a função de perda a ser minimizada:

$$Loss = \sum_i^N [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha] \tag{3.2}$$

Onde $\alpha$ é uma margem aplicada entre pares positivos e negativos; $x_i^a$ é o *embedding* âncora sendo a referência facial; $x_i^p$ é o *embedding* positivo com a mesma identidade da âncora; $x_i^n$ é a referência negativa e $\tau$ é o conjunto de todos os grupos de três ou *triplets* da base de dados. A figura 3.6 representa visualmente o processo de aprendizado do modelo ao utilizar *triplet loss*:

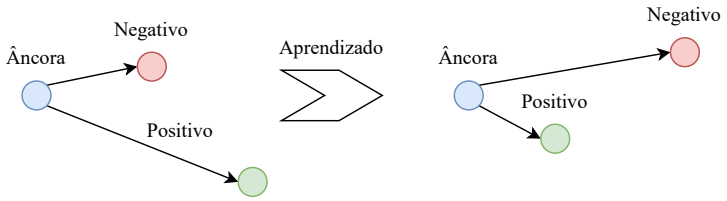


Figura 3.6: Esquematização gráfica do processo de aprendizado fornecido pela *Triplet Loss*. Modificado de: Menezes et al. (2020)

Para verificar se dois rostos são da mesma pessoa, foi utilizado um limite (*threshold*) baseado na distância euclidiana dos *embeddings* dos rostos, como mostra a equação 3.3. Se a distância estiver abaixo do limite, os *embeddings* são da mesma pessoa. Caso contrário, elas pertencem a pessoas diferentes. Esse limite foi definido inicialmente com base em experimentos com o conjunto de dados LFW (Huang et al. (2008)).

$$d\left(p,q\right) = \sqrt{\sum_{i=1}^{n}\left(q_i - p_i\right)^2} \tag{3.3}$$

Essa abordagem de comparação de distâncias com apenas um exemplo por classe funciona semelhantemente a um classificador *K-Nearest Neighbor* desenvolvido por Fix and Hodges (1951), no qual o K é definido como 1. Diante do caso específico onde dois indivíduos estão abaixo do limite estabelecido, um método que pode ser aplicado ao contexto da verificação facial usando essa arquitetura é que, se tivermos uma situação em que mais de um aluno tem uma distância euclidiana menor do que o inferior ao limite para uma identidade específica, presumimos que a identidade real do aluno é quando a distância é mínima (mais próxima de zero).

# Capítulo 4

# Resultados e Discussões

Diante do sistema pronto e capaz de realizar a ação da frequência, foram utilizados estes mesmos dados para comparação deste modelo com os modelos no estado-da-arte, sendo eles: Facenet512, Facenet (Schroff et al. (2015)) e Arcface (Deng et al. (2019)). A comparação foi feita ao mensurar a distância euclidiana entre os vetores (ou *embeddings*) resultantes para cada rosto encontrado e limitá-la em um *threshold* de 0,5, onde abaixo desse valor a distância confirma o reconhecimento de um rosto. Portanto, foi verificado primeiramente os dados da pesquisa de Mery et al. (2019) no escopo das métricas: acurácia, precisão, *recall*, *F1-Score*, ROC e AUC; na forma de média entre as 5 sessões de aula.

| Modelo | Acurácia(%) | Precisão(%) | *Recall*(%) | *F1-Score*(%) |
|---|---|---|---|---|
| Facenet512 | 46,86 $\pm 0,175$ | 90,67 $\pm 0,034$ | 43,54 $\pm 0,224$ | 55,45 $\pm 0,233$ |
| Facenet | 29,55 $\pm 0,057$ | 87,66 $\pm 0,067$ | 25,08 $\pm 0,107$ | 37,36 $\pm 0,144$ |
| Arcface | 41,79 $\pm 0,135$ | 90,51 $\pm 0,054$ | 39,15 $\pm 0,196$ | 50,84 $\pm 0,232$ |
| ResNet29 | 76,71 $\pm 0,055$ | 90,19 $\pm 0,101$ | 83,00 $\pm 0,071$ | 85,64 $\pm 0,036$ |

Tabela 4.1: Tabela contendo as médias das métricas obtidas por modelo perante os dados de Mery et al. (2019). Fonte: autor, 2023.

Denota-se que o modelo apresentado neste trabalho, demonstra uma melhoria percentual média no quesito acurácia de cerca de 51,35% para com os outros modelos na entrega de uma frequência correta. Para uma questão como a frequência, todos os modelos se saíram bem ao entregar boas métricas de precisão, indicando que para este problema os reconhecedores faciais conseguem reconhecer bem alunos que estão na sala de aula, errando pouco em confundir rostos. No entanto, na métrica *recall*, o modelo ResNet29 se saiu muito melhor ao trazer uma taxa percentual de confiança de 43,27% mais alta ao dizer que determinado aluno está presente na aula. Por conseguinte, o *F1-Score* da ResNet29 apresentou também valores melhores, com uma média de 55,9% de superioridade

perante os outros modelos.

Em vista das métricas selecionadas e dos resultados obtidos, os valores de ROC e AUC representam o equilíbrio do modelo perante a sua classificação. Como pode ser visualizado na figura 4.1, o modelo apresentado neste trabalho entrega o maior equilíbrio geralmente, conseguindo alcançar um índice de separabilidade favorável.

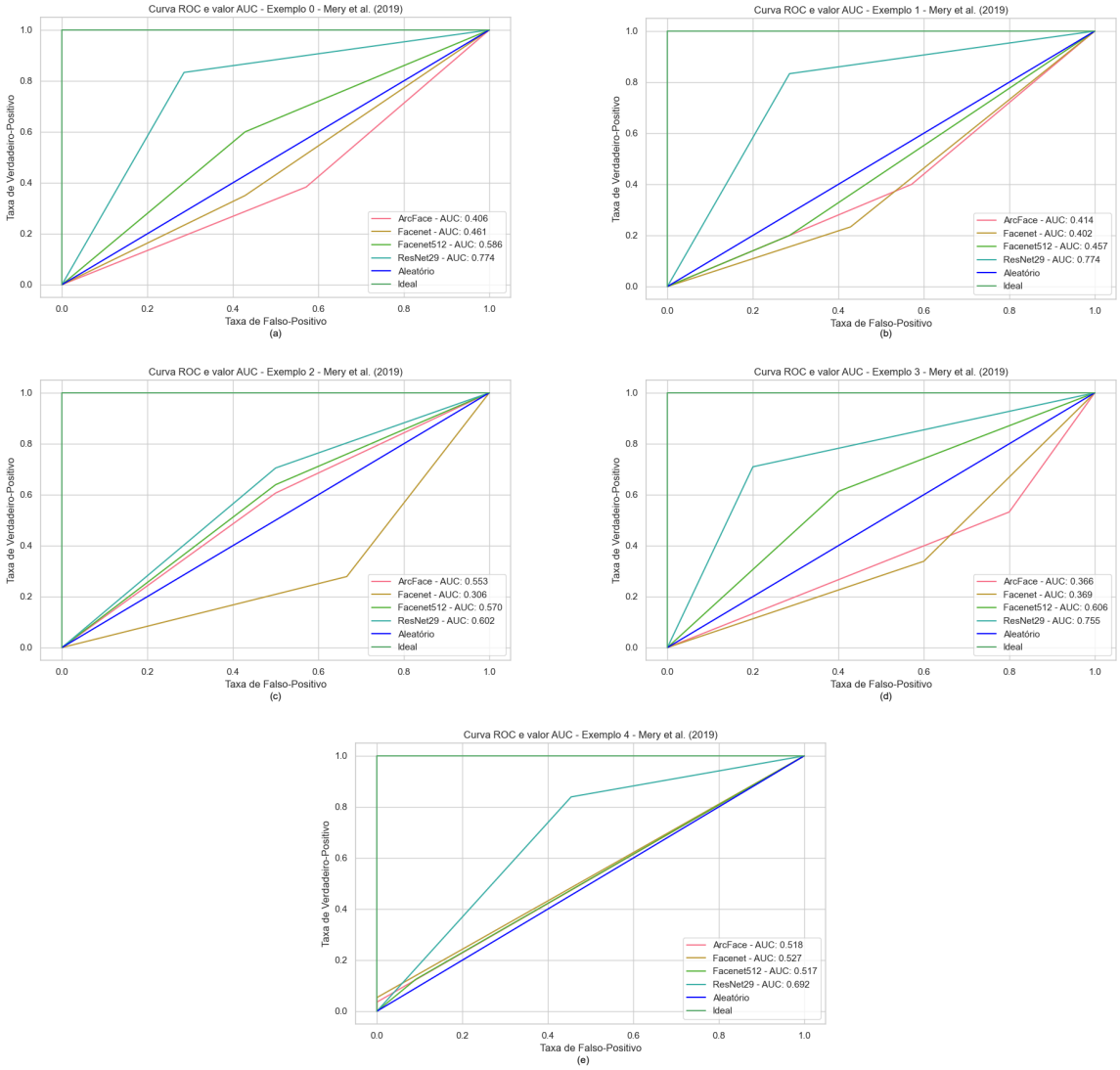


Figura 4.1: Gráficos contendo as curvas ROC e os valores AUC para cada sessão de aula do trabalho de Mery et al. (2019). Fonte: autor, 2023

Dando sequência aos testes, os dados do SESI possuem tamanho maior, porém uma resolução pior por utilizarem câmeras de vigilância comuns com apenas 3 megapixels disponíveis para captura. A tabela 4.2 contabiliza valores mais modestos, por conta da qualidade de captura das imagens da sala de aula, resultando numa queda geral nas métricas.

| Modelo | Acurácia(%) | Precisão(%) | *Recall*(%) | *F1-Score*(%) |
|---|---|---|---|---|
| Facenet512 | 24,24 | 100,00 | 19,35 | 32,43 |
| Facenet | 27,27 | 100,00 | 22,85 | 36,84 |
| Arcface | 15,15 | 100,00 | 9,67 | 17,64 |
| ResNet29 | 75,75 | 96,00 | 77,41 | 85,71 |

Tabela 4.2: Tabela contendo as métricas obtidas por modelo perante os dados do SESI. Fonte: autor, 2023.

Analogamente aos dados fornecidos pelo trabalho de Mery et al. (2019), os dados do SESI presentes na figura 4.2 também entregaram métricas melhores, cerca de 3,63 vezes mais acurácia para a ResNet29, além de valores de ROC e AUC superiores aos outros modelos. No entanto, dado o cenário no qual as imagens tem uma resolução menor, a desempenho geral foi levemente comprometida, assemelhando-se ao caso (4.1 (c)), onde o modelo pende para verdadeiros-positivos, mas sem muita acentuação, indicando dificuldade ao classificar a presença.

Por fim, para realizar a detecção e o reconhecimento, a ResNet29 levou aproximadamente 43 segundos para analisar 55 *frames* em comparação aos outros modelos que levaram em média 30 minutos, devido suas arquiteturas realizarem a análise e a detecção serialmente, sendo necessário verificar todos os rostos de uma imagem por vez para reconhecer cada indivíduo.

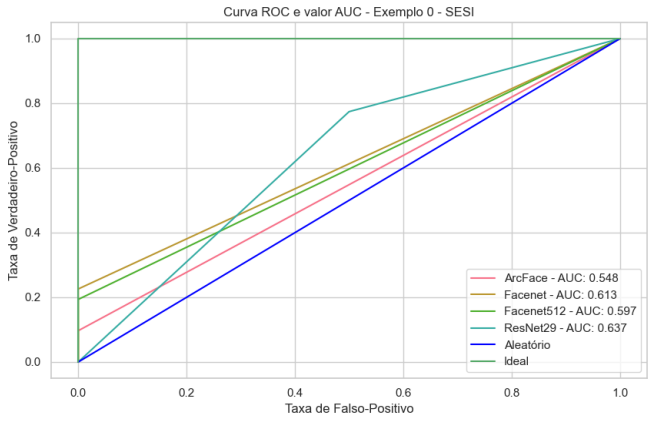


Figura 4.2: Gráficos contendo as curvas ROC e os valores AUC para uma sessão de aula do SESI. Fonte: autor, 2023

# Capítulo 5

# Conclusão

Nesta pesquisa, conforme os resultados obtidos e com o panorama educacional brasileiro, conclui-se que o uso do modelo ResNet29 é uma aproximação viável para a tarefa de reconhecimento facial voltada para frequência escolar. Foi possível analisar o desempenho de diferentes modelos diante não apenas de um cenário ideal onde a quantidade de imagens para análise é baixa com boa qualidade, mas também perante um cenário onde a qualidade das imagens não é ideal e os estudantes não estão voltados para câmera que capta suas imagens.

Para uma frequência escolar, mais importante que dizer se algum aluno está presente é saber quem não está, neste quesito o modelo se sobressai, porém, apresenta espaço para melhoria, especialmente no caso em que foram utilizados os dados do SESI. Em termos de comparação, a ResNet29 foi capaz de não apenas reconhecer com mais acurácia os alunos presentes, mas também distinguir os presentes e os faltosos superiormente ao apresentar uma especificidade maior geralmente, com uma taxa de verdadeiros negativos mais alta que os outros modelos segundo os gráficos contendo a curva ROC.

O curto tempo necessário para analisar 55 imagens revela que o modelo consegue fornecer uma visão abrangente da sala de aula em uma média de apenas 1,28 segundo. Essa eficiência é resultado da arquitetura combinada do modelo CNN de detecção facial e da ResNet29. A maneira como essa arquitetura de detecção e reconhecimento opera, juntamente com seu rápido tempo de inferência para identificar a presença, abre a oportunidade para a utilização de *streaming* de forma simultânea durante o progresso da aula em lotes.

Isso reforça a natureza não intrusiva do sistema, uma vez que não requer que a escola armazene dados adicionais para verificações posteriores. Além disso, a entrega quase que em tempo real das informações de presença coincide com o término da aula, contribuindo para uma gestão mais ágil da frequência dos alunos, além de conseguir criar uma conexão da sala de aula para o ambiente virtual, corroborando a ideia de Uskov et al. (2015).

Concluindo, a modelagem do sistema obteve sucesso ao apresentar boas métricas ao passo em que se mostra ser leve e capaz de realizar ações simultâneas indiretamente, sem

a necessidade da intrusão dos professores ou de máquinas além daquelas que captam as imagens. Portanto, a abordagem discutida evidencia a capacidade do modelo de trabalhar bem com poucos recursos e de substituir uma ação humana promissoramente, melhorando a produtividade no ambiente escolar sem a necessidade da obtenção de novos recursos financeiros e logísticos.

## 5.1 Trabalhos futuros

Considerando os resultados satisfatórios em relação ao registro de frequência em comparação com outros modelos, identificam-se oportunidades significativas para aprimorar o desempenho e a robustez do modelo por meio do aumento da complexidade da arquitetura do sistema como um todo.

A adoção da estrutura de redes residuais possibilita a incorporação de mais camadas e filtros, bem como a exploração de sinergias com outros modelos de detecção facial, contribuindo para aprimorar a robustez do sistema, com atenção ao custo computacional. Portanto, como um passo subsequente promissor, considera-se a expansão da complexidade global do sistema, visando desbloquear todo o potencial inerente à arquitetura das redes residuais.

# Referências Bibliográficas


Anderson, D. and McNeill, G. (1992). Artificial neural networks technology.

Baheti, P. (2021). Activation functions in neural networks [12 types  use cases].

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

Biblioteca Educação Já (2021). O que pensam os professores brasileiros sobre a tecnologia digital em sala de aula?

Branca Nunes, V. (2011). Diferença abissal entre ensino público e privado no brasil também se reflete no quesito segurança.

Chen, D., Ren, S., Wei, Y., Cao, X., and Sun, J. (2014). Joint cascade face detection and alignment. In *European Conference on Computer Vision*.

Chua, L. and Yang, L. (1988). Cellular neural networks: theory. *IEEE Transactions on Circuits and Systems*, 35(10):1257–1272.

da Fontoura Costa, L. and Cesar, R. (2000). *Shape Analysis and Classification: Theory and Practice*. Image Processing Series. Taylor & Francis.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.

Deeba, F., Memon, H., Dharejo, F. A., Ahmed, A., and Ghaffar, A. (2019). Lbph-based enhanced real-time face recognition. *International Journal of Advanced Computer Science and Applications*, 10(5).

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Fix, E. and Hodges, J. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine.

GeeksforGeeks (2021).

Geitgey, A. (2016). Machine learning is fun! part 4: Modern face recognition with deep learning.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Google (2022).

Governo Federal do Brasil (1996). Lei de diretrizes e bases da educação nacional. [Accessed: 2023-08-07].

Goyal, C. (2021). Complete guide to gradient-based optimizers in deep learning.

GZH (2018). Câmeras na escola: salas de aula precisam ser espaços de confiança.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Number v. 10 in Neural networks and learning machines. Prentice Hall.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Huang, G., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. *Tech. rep.*

Khan, M., Chakraborty, S., Astya, R., and Khepra, S. (2019). Face detection and recognition using opencv. In *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 116–119.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758.

King, D. E. (2015). Max-margin object detection.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Menezes, A. G., Sá, J. M. D. d. C., Llapa, E., and Estombelo-Montesco, C. A. (2020). Automatic attendance management system based on deep one-shot learning.

Mery, D., Mackenney, I., and Villalobos, E. (2019). Student attendance system in crowded classrooms using a smartphone camera. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 857–866.

NCES (2009). Why does attendance matter?

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

Pascanu, R., Mikolov, T., and Bengio, Y. (2012). Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition.

Salton, K. (2017). Reconhecimento facial: Como funciona o lbph.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832.

Serengil, S. I. and Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5.

Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316.

Souza, F. L. d. (2014). Classificador fisherface fuzzy para o reconhecimento de faces. Master's thesis.

Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1891–1898, Los Alamitos, CA, USA. IEEE Computer Society.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA. PMLR.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708.

Taqi, A. M., Awad, A., Al-Azzo, F., and Milanova, M. (2018). The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 140–145.

Thornton, M. (2013). Persistent absenteeism among irish primary school pupils. *Educational Review*, 65:488 – 501.

Uskov, V., Bakken, J., and Pandey, A. (2015). *The Ontology of Next Generation Smart Classrooms*, volume 41, pages 3–14.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I.

Wahyu Mulyono, I. U., Ignatius Moses Setiadi, D. R., Susanto, A., Rachmawanto, E. H., Fahmi, A., and Muljono (2019). Performance analysis of face recognition using eigenface approach. In *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 1–5.

Yang, H. and Wang, Y. (2007). A lbp-based face recognition method with hamming distance constraint. In *Fourth International Conference on Image and Graphics (ICIG 2007)*, pages 645–649.

Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhang, X., Zhang, X., and Dolah, T. D. J. B. (2022). Intelligent classroom teaching assessment system based on deep learning model face recognition technology. *Scientific Programming*.

Zhao, W., Rao, Y., Wang, Z., Lu, J., and Zhou, J. (2021). Towards interpretable deep metric learning with structural matching.

Zhao, Z.-Q., Zheng, P., tao Xu, S., and Wu, X. (2019). Object detection with deep learning: A review.