UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

JOÃO VICTOR CAMPOS DE NEGREIRO

RECONHECIMENTO DE INDIVÍDUOS MULTIMODAL (FACE E VOZ): ANÁLISE COMPARATIVA ENTRE UMA ABORDAGEM DE APRENDIZADO DE MÁQUINA CLÁSSICA E UMA PROPOSTA UTILIZANDO REDE NEURAL PROFUNDA

Manaus
2022

JOÃO VICTOR CAMPOS DE NEGREIRO

RECONHECIMENTO DE INDIVÍDUOS MULTIMODAL (FACE E VOZ): ANÁLISE COMPARATIVA ENTRE UMA ABORDAGEM DE APRENDIZADO DE MÁQUINA CLÁSSICA E UMA PROPOSTA UTILIZANDO REDE NEURAL PROFUNDA

Dissertação apresentada ao Programa de Mestrado em Engenharia Elétrica da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica na área de concentração Controle e Automação de Sistemas.

Orientadora: Profa. Dra. Marly Guimarães Fernandes Costa
Coorientador: Prof. Dr. Cícero Ferreira Fernandes Costa Filho

Manaus
2022

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Negreiro, João Victor Campos de
N385r Reconhecimento de indivíduos multimodal (face e voz): análise comparativa entre uma abordagem de aprendizado de máquina clássica e uma proposta utilizando rede neural profunda / João Victor Campos de Negreiro . 2022
102 f.: il. color; 31 cm.

Orientadora: Marly Guimarães Fernandes Costa
Coorientador: Cícero Ferreira Fernandes Costa Filho
Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Amazonas.

1. Reconhecimento biométrico. 2. Aprendizado de máquina. 3. Rede neural profunda. 4. Aprendizado por quantização vetorial. 5. Rede neural convolucional. I. Costa, Marly Guimarães Fernandes. II. Universidade Federal do Amazonas III. Título

JOÃO VICTOR CAMPOS DE NEGREIRO

RECONHECIMENTO DE INDIVÍDUOS MULTIMODAL (FACE E VOZ): ANÁLISE COMPARATIVA ENTRE UMA ABORDAGEM DE APRENDIZADO DE MÁQUINA CLÁSSICA E UMA PROPOSTA UTILIZANDO REDE NEURAL PROFUNDA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica na área de concentração Controle e Automação de Sistemas.

Aprovado em 03 de novembro de 2022.

BANCA EXAMINADORA

Profª Dra. Marly Guimarães Fernandes Costa, Presidente
Universidade Federal do Amazonas

Prof. Dr. Jozias Parente de Oliveira, Membro
Universidade do Estado do Amazonas

Prof. Dr. José Raimundo Gomes Pereira, Membro
Universidade Federal do Amazonas

AGRADECIMENTOS

Primeiramente agradeço aos meus pais, Ademir e Edilza, que puderam me orientar, educar, direcionar e incentivar a estudar desde criança até a universidade.

Agradeço aos meus orientadores, Profª. Drª. Marly Guimarães Fernandes Costa, e, Prof. Dr. Cícero Ferreira Fernandes Costa Filho por caminharem comigo neste trabalho, pela orientação concedida, disponibilidade para sanar dúvidas, esclarecer as dificuldades encontradas, pelo aprendizado adquirido e, principalmente, pela confiança que me foi dada.

Agradeço à Áurea que me acompanha na minha vida, que esteve nos momentos de dificuldades me incentivando e não me deixando desistir.

Agradeço ao meu irmão Júlio, que sempre se prontificou em me ajudar, tirar dúvidas e contribuir com o trabalho no que esteve ao seu alcance.

Agradeço aos meus amigos e companheiros de universidade.

Agradeço a todas as pessoas que de alguma forma contribuíram para minha formação acadêmica.

À instituição FAPEAM pelo fornecimento da bolsa que permitiu dedicação exclusiva ao desenvolvimento da dissertação.

À Universidade Federal do Amazonas e, em especial, ao Centro de Tecnologia Eletrônica e da Informação – CETELI – pela concessão de toda infraestrutura para realização deste trabalho. Esta pesquisa, conforme previsto no Art. 48 do decreto nº 6.008/2006, foi parcialmente financiada pela Samsung Eletrônica da Amazônia Ltda, nos termos da Lei Federal nº 8.387/1991, através de convênio nº 004, firmado com o Centro de P&D em Eletrônica e Tecnologia da Informação da Universidade Federal do Amazonas - CETELI / UFAM.

RESUMO

Os seres humanos utilizam características do corpo como rosto, voz e olhos em conjunto com outras informações contextuais para se reconhecerem. O reconhecimento biométrico busca identificar um indivíduo utilizando características comportamentais, físicas ou psicológicas. Esse trabalho apresenta uma análise comparativa entre uma abordagem de aprendizado de máquina clássica e uma proposta utilizando rede neural profunda na atividade de reconhecimento de indivíduo. Utilizaram-se dois modos biométricos: face e voz. Estes dados foram obtidos da base de dados bimodal MOBIO (MCCOOL et al, 2012). Utilizaram-se 50 indivíduos, sendo 37 homens e 13 mulheres. Aplicou-se um pré-processamento nas imagens, extraindo a face, padronizando em 64x80 e convertendo para monocromática. Foi utilizado um *autoencoder* para obter uma representação reduzida dos dados da face. Para voz, optou-se por um detector de atividade para classificar trechos de áudios com ou sem voz. Extraíram-se coeficientes Mel-Cepstrais e seus coeficientes derivados, compondo 39 coeficientes. Foram desenvolvidos modelos unimodais e multimodais de identificação biométrica, totalizando 6 arquiteturas. O modelo multimodal com técnicas de aprendizagem de máquina possui uma etapa fusão à nível de pontuação e aprendizado por quantização vetorial (LVQ). O modelo multimodal com técnicas de aprendizado profundo de máquina possui uma fusão a nível de característica e uma rede neural convolucional (CNN). Testaram-se as arquiteturas propostas em diversos cenários de *clusters*, quantidade de *frames* de áudio, dimensão da camada de codificação, quantidade de coeficientes MFCCs, regularização e otimizadores. Avaliaram-se os sistemas através da área sobre a curva ROC (AUC-ROC), taxa de verdadeiros positivos e taxa de falsos positivos e o limiar do melhor ponto de operação. Além disso, mediu-se o tempo de treinamento e testes das redes elaboradas. Os resultados mostram que para a proposta multimodal com LVQ foi obtido AUC-ROC de 0,98 e a proposta multimodal com CNN teve um valor de AUC-ROC de 0,99. Os resultados apontaram que a utilização de aprendizagem profunda produz melhores desempenhos, além de treinamentos mais otimizados. Assim, as arquiteturas propostas neste trabalho podem constituir um bom ponto de partida para implementação de um sistema robusto de identificação automática de indivíduos.

Palavras-chave: reconhecimento biométrico, aprendizado de máquina, face-voz, rede neural profunda, aprendizado por quantização vetorial, rede neural convolucional.

ABSTRACT

Humans use body features such as face, voice and eyes in conjunction with other contextual information to recognize themselves. Biometric recognition seeks to identify an individual using behavioral, physical or psychological characteristics. This work presents a comparative analysis between a classical machine learning approach and a proposal using a deep neural network in the individual recognition activity. Two biometric modes were used: face and voice. These data were obtained from the MOBIO bimodal database (MCCOOL et al, 2012). Fifty individuals were used, 37 men and 13 women. A pre-processing was applied to the images, extracting the face, standardizing it in 64x80 and converting it to monochrome. An autoencoder was used to obtain a reduced face data representation. For voice, an activity detector was chosen to classify audio excerpts with or without voice. Mel Cepstral coefficients and their derived coefficients were extracted, composing 39 coefficients. Unimodal and multimodal models of biometric identification were developed, totaling 6 architectures. The multimodal model with machine learning techniques has a fusion step at the scoring level and Learning Vector Quantization (LVQ). The multimodal model with deep machine learning techniques has a feature level fusion and a Convolutional Neural Network (CNN). The proposed architectures were tested in different cluster scenarios, audio frames number, encoding layer dimension, MFCCs coefficients number, regularization and optimizers. The systems were evaluated through the area under the ROC curve (AUC-ROC), True Acceptance Rate (TAR) and False Acceptance Rate (FAR) and best operating point threshold. In addition, the training and testing time of networks was measured. The results show that for the multimodal proposal with LVQ, an AUC-ROC of 0.98 was obtained and the multimodal proposal with CNN reached an AUC-ROC value of 0.99. The results showed that deep learning produces better performances, in addition to more optimized training. Thus, the architectures proposed in this work can constitute a good starting point for implementing a robust system for automatic identification of individuals.

Keywords: biometric recognition, machine learning, face-voice, deep neural network, learning vector quantization, convolutional neural network.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

ABIS	*Automated Biometric Identification Solution*
ACC	*Accuracy*
ADAM	*Adaptive Moment Estimation*
AFIS	*Automated Fingerprint Identification System*
ANN	*Artificial Neural Network*
AUC-ROC	*Area Under the Curve - Receiver Operating Characteristics*
CAFe	Comunidade Acadêmica Federada
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CCA	*Canonical Correlation Analysis*
CETELI	Centro de Tecnologia Eletrônica e da Informação
CNN	*Convolutional Neural Network*
DAV	Detector de atividade de voz
DCA	*Discriminant Correlation Analysis*
DCT	*Discrete Cosine Transform*
DFT	*Discrete Fourier Transform*
DPCA	Análise Diagonal do Componente Principal
DPCA	*Diagonal Principal Component Analysis*
EER	*Equal Error Rate*
FAPEAM	Fundação de Amparo à Pesquisa do Estado do Amazonas
FAR	*False Acceptance Rate*
FFT	*Fast Fourier Transform*
FN	Falso Negativo
FP	Falso Positivo
FPGA	*Field-Programmable Gate Array*
FRR	*False Reject Rate*
GAP	*Global Average Pooling*
GMM	*Gaussian Mixture Model*
GMM-NAP	*Gaussian Mixture Model with Nuisance Attribute Projection*

GPU	*Graphics Processing Unit*
HOG	*Histogram of Oriented Gradients*
IEEE	*Institute of Electrical and Electronic Engineers*
IoT	*Internet of Things*
KNN	*K-Nearest Neighbors*
LBP	*Local Binary Pattern*
LGPD	Lei Geral de Proteção dos Dados
LLR	*Log-Likelihood Ratio*
LSA	*Latent Semantic Analysis*
LVQ	*Learning Vector Quantization*
MCC	*Matthews Correlation Coefficient*
MFCC	*Mel-frequency Cepstral Coefficients*
MOBIO	*Mobile Biometry*
MT	*Mean Time*
PCA	*Principal Component Analysis*
PDI	Processamento Digital de Imagens
PF	Polícia Federal
ReLU	*Rectified Linear Unit*
RGB	*Red, Green, and Blue*
RMSProp	*Root Mean Square Propagation*
ROC	*Receiver Operating Characteristic*
SGD	*Stochastic Gradient Descent*
SVM	*Support Vector Machine*
TAR	*True Acceptance Rate*
TPU	*Tensor Processing Unit*
TSE	Tribunal Superior Eleitoral
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
VR	*Verification Rate*
XJTU	*Xi'an Jiaotong University*

SUMÁRIO

1 INTRODUÇÃO

Os seres humanos utilizam características do corpo como rosto, voz e olhos em conjunto com outras informações contextuais para se reconhecerem (GEARY, 2004). A junção de todos os atributos compõe a identidade pessoal de um indivíduo. Um dos principais desafios de pesquisa relevante consiste no estabelecimento de uma relação entre o indivíduo e sua identidade pessoal. Assim, o reconhecimento biométrico busca identificar um indivíduo utilizando características comportamentais, físicas ou psicológicas (JAIN; ROSS; NANDAKUMAR, 2011).

Como características biométricas podem ser empregados traços fisiológicos (como íris ou face) ou comportamentais (como voz ou assinatura) que estejam associadas a um indivíduo. Para um traço ser considerado uma característica biométrica é necessário atender aos requisitos: universalidade, distinção, permanência e coletividade. Há diversas características biométricas e as mais utilizadas são: impressão digital, face, geometria da mão, íris e a voz (PRABHAKAR; PANKANTI; JAIN, 2003). Os recursos para o reconhecimento facial estão totalmente disponíveis em computadores e dispositivos móveis, o que desperta interesse na criação de aplicações como autenticação biométrica, vigilância, controles de fronteiras e redução de fraudes no comércio eletrônico, entre outros. O processo de reconhecimento facial é natural e não intrusivo e pode ser realizado a distância, colocando-o em vantagens em relação a impressão digital e o reconhecimento da íris (JAIN; LI, 2011). A voz pode ser considerada uma característica humana que combina aspectos biológicos e comportamentais. Um som produzido por uma pessoa é composto de um conjunto de aspectos físicos do corpo, como boca, nariz, lábios e cordas vocais. O som produzido é afetado pelo estado emocional, pela idade ou até mesmo por condições médicas. O reconhecimento da voz pode ser realizado tanto de forma passiva, através de conversas regulares sem interações com os indivíduos ou de forma ativa, falando uma determinada frase, com uma interação direta com o indivíduo. As vantagens da utilização da voz como característica biométrica é o baixo custo da tecnologia aliado ao mínimo grau de invasividade, exigindo apenas que os indivíduos forneçam uma amostra vocal (COUNCIL; COMMITTEE, 2010).

De acordo com Prabhakar, Pankanti e Jain (2003), os sistemas de reconhecimento biométrico podem ser divididos em dois: identificação e verificação/autenticação. No modo de identificação, o sistema busca responder à pergunta: “Quem é esta pessoa?”. Assim, o sistema pesquisa uma correspondência comparando as características do indivíduo com todas as demais

cadastradas no banco de dados (comparação 1:N). Porém, no modo de verificação, há apenas duas possibilidades de resultado: ou o usuário é aceito, pois é a identidade genuína, ou é rejeitado, pois o usuário foi considerado como impostor (comparação 1:1). Busca-se, então, responder à pergunta “Essa pessoa é o João?”, realizando autenticação, caso a identidade fornecida corresponda com a que foi pré-cadastrada no banco de dados.

O presente trabalho está voltado para a temática de sistema biométrico no modo de identificação.

Com respeito às aplicações dos sistemas biométricos, dividem-se em três grupos principais. O grupo 1, aplicações comerciais, subdivide-se em: *login* em rede de computadores, controle de acesso físico, segurança eletrônica de dados e telefones celulares. Já o grupo 2, aplicações governamentais, fragmenta-se em: carteira de identidade nacional, carteira de motorista, controle de fronteira e passaporte. Por fim, o grupo 3 compõe as aplicações forenses, dividido em identificação de cadáver, investigação criminal, identificação de terrorista e identificação de pessoa desconhecida (PRABHAKAR; PANKANTI; JAIN, 2003).

No Brasil, dois exemplos da utilização da identificação e verificação biométrica em larga escala podem ser citados. O primeiro é o programa de identificação biométrica do eleitor brasileiro, implementada pelo Tribunal Superior Eleitoral (TSE), que excluiu a possibilidade de intervenção humana, uma vez que a urna somente é liberada para votação após a verificação das impressões digitais do eleitor com as armazenadas no banco de dados da Justiça Eleitoral. Para realizar essa tarefa foi desenvolvido um sistema AFIS (do inglês *Automated Fingerprint Identification System*) que, durante o cadastramento, realiza em uma comparação automatizada das dez impressões digitais do eleitor com todas as impressões digitais cadastradas para garantir que o registro do eleitor seja único (TSE, 2018). Mais recentemente, em 2021, a Polícia Federal (PF) adquiriu o sistema ABIS (do inglês *Automated Biometric Identification Solution*), considerado uma evolução do sistema AFIS, possui suporte a impressão digital, facial, íris, voz e outras características. O sistema oferece a possibilidade da utilização multimodal através de fusão a nível de pontuação. Pode ser utilizado para identificação 1:N e verificação 1:1, possuindo uma alta escalabilidade com a capacidade de pesquisar e verificar milhões de registros. Dentre suas aplicações, o sistema atua na identificação biométrica, nos processos de documentos oficiais, investigação e identificação criminal, prevenção de fraudes, casos de pessoas desaparecidas e identificação de corpos (PF, 2021).

Os sistemas biométricos também se classificam como unimodal, que utiliza apenas uma única característica, e multimodal, que utiliza duas ou mais características. Sistemas com

múltiplas biometrias surgiram devido às limitações de utilizar apenas uma característica, estando sujeitas a distorções no sinal de aquisição ou outros fatores como baixa luminosidade, sujeira na impressão digital, ruídos sonoros e etc. Ademais, o sistema multimodal é considerado mais robusto e mais difícil de ser fraudado devido à necessidade de se falsificar dois ou mais traços biométricos de forma simultânea. Uma outra vantagem sobre os sistemas unimodais é resolução do problema de não universalidade, pois utilizar duas ou mais características biométricas garantem um cobertura populacional relevante (JAIN; ROSS, 2004).

O presente trabalho debruça-se na utilização de mais de um modo, mais precisamente, dois modos: face e voz.

Os sistemas biométricos normalmente possuem os módulos de aquisição de sinal, extração de características, *matching* e decisão. Os sistemas multimodais são acompanhados da etapa de fusão. Essa etapa pode ocorrer antes da etapa de *matching*, a nível de sensor e de característica, ou após o *matching*, a nível de pontuação, de rank e de decisão (JAIN; ROSS; NANDAKUMAR, 2011).

Com advento da internet, *smartphones* e *big data*, o ser humano está cada vez mais conectado aos diversos sistemas, além de produzir dados constantemente durante o seu dia a dia. Essa nova era da informação criou desafios para empresas e pesquisadores em prol da segurança da informação. Segundo a Interpol, em seu relatório do impacto do Covid-19 em cibercrimes, as principais ameaças foram golpes e *phishing* online, malware disruptivo (*ransomware*), *malware* de coleta de dados, domínios maliciosos e desinformação, sendo 12% dos relatos oriundos do continente Americano (INTERPOL, 2020).

O aprendizado de máquina pode ser definido como campo de estudo que habilita os computadores a aprenderem sem serem explicitamente programados, permitindo aprenderem com os dados. Abrangendo aplicações em diversas áreas do conhecimento, são excelentes em problemas complexos que não possuem boa solução com abordagens tradicionais, problemas com necessidade de adaptação de ambientes ou novos dados e problemas complexos com larga quantidade de dados. Os sistemas de aprendizado de máquinas são classificados quanto à supervisão humana: supervisionado, não supervisionado, semi-supervisionado e aprendizado por reforço. Além disso, a partir dos anos 1980, inseriu-se o conceito de aprendizado profundo através das redes neurais. Essas redes possuem camadas de neurônios escondidas responsáveis pelo aprendizado de máquina. Mais especificamente, as redes neurais convolucionais são úteis para processamento de dados bidimensionais, capazes de extrair característica automaticamente, possuem sua arquitetura inspirada do córtex visual (GÉRON, 2019).

O presente trabalho realiza uma comparação entre um método clássico de aprendizado de máquina, o aprendizado por quantização vetorial, e uma abordagem de aprendizado profundo, utilizando rede neural convolucional nos modelos de reconhecimento biométrico unimodais de face e de voz e multimodal nos modos face e voz conjuntamente. Foram propostas seis arquiteturas para o problema de classificação, sendo cada uma para o respectivo modo (face, voz e multimodal) e abordagem (aprendizado por quantização vetorial e rede neural convolucional). Foi realizada a identificação biométrica e então comparados os desempenhos das arquiteturas em diferentes cenários de generalização e otimização. A base de dados utilizada no trabalho foi a MOBIO (MCCOOL et al, 2012). Essa é uma base de dados bimodal que consiste em dados de face e voz.

1.1 OBJETIVO GERAL

Desenvolvimento de métodos multimodais (face e voz) de reconhecimento biométrico de indivíduos que utilizam uma abordagem de aprendizado de máquina clássica e uma abordagem com aprendizado de máquina profundo, buscando o avanço do estado da arte.

1.2 OBJETIVOS ESPECÍFICOS

- Avaliar o desempenho dos modelos unimodais de face, voz e multimodal biométrico em uma abordagem não profunda de reconhecimento, utilizando aprendizado por quantização vetorial.
- Avaliar o desempenho dos modelos unimodais de face, voz e multimodal biométrico em uma abordagem de aprendizado profundo de reconhecimento, utilizando rede neural convolucional.
- Comparar o desempenho de modelos unimodais (face e voz) com os modelos multimodais de reconhecimento biométrico.
- Avaliar se há avanço no desempenho do sistema multimodal biométrico através de uma abordagem de aprendizado profundo, comparado ao estado da arte nesse tema.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado conforme a divisão descrita a seguir:

- Capítulo 1: Introdução;
- Capítulo 2: Revisão Bibliográfica;
- Capítulo 3: Referencial Teórico;
- Capítulo 4: Materiais e Métodos;
- Capítulo 5: Resultados e discussão;
- Capítulo 6: Conclusões.

O Capítulo 1 contextualiza a tarefa de identificação biométrica, cita a importância da utilização de sistemas multimodais robustos, discorre sobre as etapas e desafios para elaboração de um sistema de identificação multimodal e apresenta os objetivos gerais e específicos deste trabalho de dissertação. O Capítulo 2 busca dar uma visão do estado da arte sobre o tema, através da análise de trabalhos publicados na literatura, com foco nas abordagens e técnicas exploradas pelos autores e, por fim, apresenta um quadro sumário desta análise. O Capítulo 3 apresenta os principais fundamentos teóricos para entendimento deste trabalho: redes *autoencoder*, aprendizado por quantização vetorial, coeficientes Mel-Cepstrais, redes neurais convolucionais e os métodos de otimizações. No Capítulo 4 são apresentados as características do banco de dados utilizado, a implementação das duas abordagens, aprendizagem de máquina clássica e rede neural profunda assim como as arquiteturas propostas e o fluxo do treinamento, validação e teste das mesmas. No Capítulo 5, são apresentados os resultados alcançados e a análise dos mesmos. Por fim, no Capítulo 6 são apresentadas as conclusões decorrentes deste trabalho.

2 REVISÃO DA LITERATURA

Com objetivo de obter o estado da arte sobre como os métodos multimodais para reconhecimento biométrico estão sendo implementados, foi realizada, como passo inicial desse projeto, uma revisão da literatura. As buscas bibliográficas foram realizadas nas bases literárias *Web of Science*, *IEEE* e *Engineering Village*, disponibilizadas através do Portal de Periódicos da CAPES e acessada remotamente pela Comunidade Acadêmica Federada - CAFe. Os termos de busca utilizados foram agrupados pelo uso de operadores Booleanos *AND* e *OR*), conforme a seguir:

> *Biometric AND (Multimodal OR Bimodal) AND (Verification* OR *Recognition* OR *Identification).*

Essa mesma estratégia de busca foi utilizada nas bases literárias citadas, realizando-se as devidas adequações e adaptações. Restringiu-se o período de buscas para o intervalo entre 2010 e 2021. Os resultados, então, foram refinados pelos seguintes termos: *fusion, voice* e *face*. Priorizaram-se os artigos de revista com maior confiabilidade e credibilidade científica e as publicações mais recentes. Assim, os artigos revisados foram os apresentados no Quadro 1.

2.1 ANÁLISE DOS TRABALHOS

Ao se analisar previamente os trabalhos obtidos que utilizavam no mínimo os modos face ou voz, percebe-se que alguns objetivaram a autenticação biométrica, outros se ocupavam da tarefa de identificação. Quanto ao padrão metodológico, observou-se que os trabalhos, com os devidos ajustes, eram compostos das seguintes etapas: pré-processamento, a extração das características, fusão dos dados e decisão.

Através da análise do Quadro 1, é notório que os autores optaram por métodos clássicos na extração de características como: Coeficientes Mel-Cepstrais, MFCC (do inglês *Mel-frequency Cepstral Coefficients*), Histograma de Gradientes Orientados, HOG (do inglês *Histogram of Oriented Gradients*), Padrão Binário Local, LBP (do inglês *Local Binary Pattern*) e análise de componente principal, PCA (do inglês *Principal Component Analysis*). Dentre as opções de métodos de fusão, predominaram-se técnicas de fusão ao nível de características e ao de pontuação. Distância Euclidiana e o classificador k vizinhos mais próximos se destacaram dentre as técnicas utilizadas na etapa de decisão. Das diversas métricas de avaliação, pode-se citar a Taxa de Erro Igual (EER do inglês *Equal Error Rate*) como principal ferramenta utilizada

para medir os desempenhos dos sistemas. A seguir apresentam-se as análises individuais dos artigos cujos resumos estão exibidos no Quadro 1.

O primeiro trabalho expõe um sistema de identificação multimodal utilizando dados de face e voz baseados em uma fusão de característica de baixo nível com dados de vídeos (JIANG; SADKA; CROOKES, 2010). Dentre os principais desafios do estado da arte na época, destaca-se a ação de creditar regiões de face semelhantes em quadros sucessivos de vídeos. Em relação a detecção e extração da face, os colaboradores utilizaram o algoritmo já consagrado *AdaBoost*, proposto por Viola e Jones (2001). O classificador foi treinado para detectar olhos em uma região de face e, através de um modelo gaussiano, as faces sucessivas foram ponderadas para creditar as imagens frontais com maior peso em relação as faces mal posicionadas. Como a maioria dos demais trabalhos, a técnica utilizada para extração de caraterísticas de voz foi a MFCC. Esse método utiliza bandas de frequência igualmente espaçadas, o que o torna mais similar com a resposta do sistema auditivo humano. Para a etapa de fusão, os autores utilizaram a técnica Laplaciano *Eigenmap,* que é uma abordagem não linear capaz de preservar a geometria intrínseca dos dados e superar a Análise Semântica Latente, LSA (do inglês *Latent Semantic Analysis*), um modelo tradicional linear convencional. Os autores utilizaram como métrica de avaliação a taxa de erro comparando cenários diferentes: modalidade simples versus multimodal, com e sem estimação da posição da face e fusão com subespaço LSA versus subespaço laplaciano. Verificou-se que, em todos os cenários, o sistema proposto obteve melhor acurácia e uma redução na taxa de erro ao aumentar a quantidade de trechos de vídeos. O trabalho se destaca, por dois fatores: a proposta multimodal gerou melhores resultados frente as modalidades que utilizaram somente face ou voz e, a fusão com método não linear apresentou melhor desempenho que o estado da arte da fusão baseada em LSA.

Kumar e Swamy (2010) inovaram apresentando um sistema biométrico com modos de face e voz que utiliza imagens de face de indivíduos antes e depois de cirurgias plásticas. De acordo com os autores, não havia até aquela data, experimentos científicos que identificassem faces com rostos que foram submetidos, previamente, a uma cirurgia plástica. A heterogeneidade da base dados foi um desafio, pois os procedimentos estéticos alteraram a forma e a textura das características faciais, dificultando uma correlação entre a face antes e após a cirurgia. Pode-se destacar como item chave desse trabalho, na etapa de extração de característica da face, o uso da Análise Diagonal do Componente Principal (DPCA). Os autores descreveram de forma clara a limitação da técnica 2DPCA, qual seja, somente reflete as informações entre as linhas, o que implica que algumas estruturas, como regiões de faces, não

sejam bem representadas. Porém, no DPCA, a imagem de face é transformada em sua correspondente imagem de face diagonal, contendo simultaneamente informação de linhas e colunas, assim gerando informações e estruturas para reconhecimento. Menciona-se ainda a escolha da técnica de quantização vetorial para reduzir recursos computacionais na etapa de combinação. Para o sistema multimodal, a acurácia foi medida através da métrica Taxa de Falsos Positivos (FAR do inglês *False Acceptance Rate*) atingindo 6,26%. Baseado nos resultados, os autores justificaram a necessidade de otimizar os algoritmos de reconhecimento de face, pois os procedimentos alteravam de forma significante as regiões da face e os principais algoritmos existentes na época eram dependentes das características faciais e texturas, e qualquer variação afetaria o desempenho do sistema biométrico multimodal.

O trabalho de Aronowitz e colaboradores (2014), além dos modos de face e voz, também utilizou dados de quirografia compondo um sistema de autenticação com três diferentes dados biométricos. Para a obtenção dos dados de voz, apesar da maioria dos trabalhos utilizarem, à época, vozes gravadas em condições bem controladas (gravações com interlocutor falando perto do microfone em ambientes silenciosos), optou-se por coletar os dados de voz com o distanciamento do microfone, ou seja, com o interlocutor segurando o microfone com braço estendido. Tal condição diminuiu a qualidade do áudio e provocou ruídos, gerando desafios ao processamento do áudio. O método de tratamento proposto para voz foi um modelo de mistura Gaussiana, mais especificamente a técnica GMM-NAP (do inglês *Gaussian Mixture Model with Nuisance Attribute Projection*). Ela se divide em 3 etapas: obtenção das características de baixo nível, alto nível e compensação de variabilidade entre sessões. A fusão em nível de pontuação foi implementada utilizando a razão de verossimilhança, LLR *(do Inglês Log-Likelihood Ratio*), que realiza um mapeamento dos dados providos de cada canal de tratamento e utiliza a soma ponderada para realizar a fusão. Este processo foi acrescido de um fator de qualidade que visava reduzir o impacto de dados ruidosos e enfatizar os canais com melhores dados. Os autores realizaram experimentos medindo o fator de qualidade através da relação sinal-ruído do sinal de entrada. Os resultados obtidos medidos através do EER para a fusão baseada em qualidade foram de 0,49%.

Diferente dos trabalhos já abordados, Zhang, Dai e Xu (2017) escolheram utilizar o mesmo extrator de característica para face e voz: transformada *wavelet* de Haar. A ideia consistiu em gerar uma única representação de característica, tornando as entradas em um domínio comum. O processo de transformação não gera perda de informação. Um ponto chave do sistema proposto é a capacidade de superar defeitos em algum canal. Em casos de relação

sinal ruído baixa, o usuário pode ser autenticado baseado no reconhecimento da face. Ou, em casos de baixa iluminação, o indivíduo pode ser identificado pela voz. Os autores utilizaram Máquina de Vetores de Suporte, SVM (do inglês *Support Vector Machine*), para realizar a classificação binária. Foi escolhida a função base radial como núcleo. Assim, foi possível mapear as amostras para um espaço de dimensão maior, relacionando melhor as classes e as características não lineares. A acurácia do sistema de identificação com método de fusão foi de 93,6%, com um tempo de teste de 0,16 segundos.

Abozaid et al. (2019) utilizaram uma abordagem comparativa em seu sistema de autenticação biométrica. Para a extração da face, contrapôs as técnicas de *Eigenface* e PCA. Em relação a voz, confrontou coeficientes Mel-Cepstrais e estatísticos. O processo de fusão foi realizado em duas abordagens, em nível de característica, ocorrendo antes do processo de combinação e em nível de pontuação após a etapa de combinação. Na classificação, as técnicas Rede Neural Artificial, ANN (do inglês *Artificial Neural Network*), Máquina de Vetores de Suporte, SVM e Modelo de Mistura Gaussiana, GMM (do inglês *Gaussian Mixture Model*) foram avaliadas. Para testar as metodologias propostas, os autores utilizaram uma base dados com 100 indivíduos, contendo 500 imagens de face e 500 trechos de áudios. Os melhores resultados foram obtidos pela fusão em nível de pontuação considerando diferentes traços biométricos, baseados nos pontos fracos e fortes do indivíduo. Também se destaca a técnica LLR, já citada anteriormente, para realizar a autenticação entre o indivíduo genuíno e o impostor, reduzindo a probabilidade de erro. O EER foi de 0,62%.

O trabalho de Olazabal e colaboradores (2019) aplicou a autenticação multimodal biométrica em dispositivos IoT (do inglês *Internet of Things*), expandindo as aplicações de autenticação para além de dispositivos *mobiles* e *desktops*. Seu objetivo, ao utilizar dois modos de dados, foi melhorar a segurança em dispositivos IoT, aumentando a acurácia do reconhecimento. Destaca-se a escolha do *Raspberry Pi* como *hardware* para exemplificar a possibilidade dos sistemas multimodais em dispositivos de baixo recurso computacional. Uma segunda abordagem dos autores foi a de buscar otimizar o sistema proposto implementando-o em um processador FPGA, que tem execuções mais rápidas, baixo consumo de energia e pode ser integrado em dispositivos IoT. Os autores justificaram que a escolha de fusão por característica foi pelo conhecimento de que essa técnica apresentava maior precisão quando comparada aos sistemas de pontuação ou decisão. Para essa etapa foi utilizada Análise de Correlação Discriminante, DCA (do inglês *Discriminant Correlation Analysis*). Esse algoritmo combina dois ou mais vetores de características em um único vetor, abrangendo a variação

independente em cada característica que é mais preditiva para classe final. O sistema multimodal fundido com DCA, utilizando face e voz apresentou EER de 8,04%, com o tempo de autenticação de 0,91 segundos.

Zhang e colaboradores (2020) implementaram uma solução em *smartphones Android*. Sua arquitetura também contempla técnicas clássicas de extração de características com LBP e MFCC. O sistema extrai as características de face e voz para realizar o treinamento e armazena na base de dados. No estágio de autenticação, a combinação é feita individualmente para cada canal. Utilizou-se a distância euclidiana para a face e máximo *a posteriori* para voz. Devido às diferenças entre as duas combinações, para a fusão, realizou-se uma normalização usando o método adaptativo mínimo-máximo. A decisão final da autenticação é feita por uma estratégia adaptativa de fusão ponderada. A implementação dessa técnica preserva as duas características biométricas e considera a influência do ambiente, melhorando a acurácia da autenticação. A extração das características de face foi realizada com o algoritmo LBP invariante a rotação. Dessa forma, impulsionaram a robustez das características de face, melhorando o processo de combinação. Todo o esquema proposto foi validado em ambiente *Desktop* antes de ser introduzido em terminal *Android*. Para realizar a validação do sistema multimodal de autenticação, utilizou-se a base de dados XJTU que contêm faces e vozes de 102 voluntários (SHANG et al., 2017). O método empregado atingiu um excelente desempenho, com 100% de Taxa de Verdadeiros Positivos (TAR do inglês *True Acceptance Rate*) e tempo médio de 0,341 segundos.

Dinesh e Rao (2020) utilizaram características de voz e face para propor um sistema multimodal biométrico. Seu sistema contém etapas de aquisição de sinal, pré-processamento, extração de características e classificação. No módulo de aquisição inseriu-se uma câmera embutida em um *Raspeberry Pi* para captura da face e um microfone para coleta do sinal de voz. No pré-processamento utilizou-se nas imagens: conversão para escala de cinza, redimensionamento, filtragem e equalização do histograma. Para o sinal da voz, optou-se por utilizar um filtro mediano para remover trechos ruidosos. Na extração e seleção de característica da face também foi utilizado o algoritmo Viola-Jones em conjunto com a técnica LBP e durante o treinamento através de PCA, as imagens com maior autovalor foram escolhidas. MFFC foi a técnica escolhida para extrair características da voz. O algoritmo KNN foi utilizado como classificador na tarefa de identificação do indivíduo com base na medida de similaridade. Este algoritmo é implementado com base na distância euclidiana entre vizinhos. O sistema proposto foi avaliado em 4 cenários, com 30, 60, 90 e 120 indivíduos. Calculou-se as métricas FAR, FRR

e acurácia. O melhor resultado obteve 98,4% de acurácia, 0,25% de FAR e 0,75%. Os autores acreditam que em trabalhos futuros é possível a implementação do mesmo método em ambiente FPGA.

O estudo de Singh, Khanna e Garg (2020) apresenta um sistema biométrico multimodal que realiza o reconhecimento baseado em detecção de face e impressão digital. Inicialmente os autores usaram câmeras para construir um banco de imagens de face e usaram o banco FVC-2004 de imagens de impressão digital, compondo um base de dados final com 1000 imagens. O algoritmo desenvolvido por Viola e Jones (2001) e o método PCA foram utilizados para detectar e extrair as características da face, respectivamente. A fusão foi realizada a nível de característica utilizando a normalização mínimo-máximo. O SVM e a distância euclidiana foram utilizados nas etapas de classificação e combinação. Os autores também implementaram uma segunda abordagem utilizando uma fusão por pontuação para avaliar qual a melhor maneira de fundir os dados. As avaliações realizadas para a base de dados mostraram que a fusão a nível de características obteve melhor desempenho em todos os experimentos, tempo de verificação e identificação, EER e uma acurácia de 95,38%.

Mehraj e Mir (2021) implementaram um novo *framework* que combina uma fusão de características híbridas: face, orelha e marcha. Diferente das abordagens anteriores, os autores optaram por utilizar técnicas de aprendizado profundo para obtenção das características utilizando modelos de Redes Neurais Convolucional, CNN (do inglês *Convolutional Neural Network*) pré-treinados, baseados em transferência de aprendizagem múltipla. Ressalta-se que o processo de fusão é feito em dois estágios, diferentemente dos trabalhos já citados: Análise de Correlação Canônica, CCA (do inglês *Canonical Correlation Analysis*) e DCA. A vantagem de empregar essa abordagem foi utilizar a força máxima de diferentes técnicas de fusão em cada nível de fusão. O sistema atingiu 97,73% de acurácia e, comparado com estado da arte de reconhecimento biométricos multimodais que utilizaram modos de orelha, face ou voz é o que atingiu melhor resultado.

Quadro 1: Resumo da análise realizada nos trabalhos relacionados

Autores	***Modalidade***	***Tipo do sistema***	***Conjunto de dados***	***Características***	***Fusão***	***Decisão***	***Resultados***
(JIANG; SADKA; CROOKES, 2010)	Bimodal: Face e Voz	Identificação	Base de dados proprietária; 10 vídeos de 10 indivíduos; 30 segundos de duração	Voz: MFCC + Laplaciano Eigenmap Face:AdaBoost + laplaciano Eigenmap	Nível de característica: combinação de baixo nível	Distância euclidiana	ER:35%
(KUMAR; SWAMY, 2010)	Bimodal: Face e Voz	Identificação	Base de dados proprietária de cirurgia plástica; Imagens de antes e após a cirurgia	Face: Diagonal PCA Voz: MFCC	Nível de pontuação	Distância euclidiana	VR: 93,74% FAR: 6,26%
(ARONOWITZ et al., 2014)	Multimodal: Face, Voz e Quirografia	Autenticação	Base de dados proprietária e face: FERET; Imagens de dispositivos celulares; 100 usuários; 250 gravações	Face: HOG, EBIF, LBP e (PCA+LDA) Voz: MFCC e GMM	Nível de pontuação: LLR	DTW	EER:0.5%
(ZHANG; DAI; XU, 2017)	Bimodal: Face e Voz	Autenticação	Base de dados proprietária; 100 pessoas, 10 imagens e 5 amostras de	Face e voz: transformada *wavelet* de Haar	Nível de característica	RBFSVM	TAR: 93,6% FAR: 0% FRR: 1,4%

Autores	*Modalidade*	*Tipo do sistema*	*Conjunto de dados*	*Características*	*Fusão*	*Decisão*	*Resultados*
			áudio por pessoa				
(ABOZAID et al., 2019)	Bimodal: Face e Voz	Autenticação	Base de dados proprietária; 100 indivíduos; 5 imagens e 5 amostras de áudios por pessoa	Face: PCA Voz: MFCC	Nível de pontuação: LLR	Voz: GMM Face: ANN	EER: 0,62%
(OLAZABAL et al., 2019a)	Bimodal: Face e Voz	Autenticação	CSUF-SG5; Imagens de face e voz adquiridos de Samsung S5	Face: HOG e LBP Voz: MFCC	Nível de característica: DCA	KNN	EER: 8,04%
(ZHANG et al., 2020)	Bimodal: Face e Voz	Autenticação	XJTU multimodal; Imagens e trechos de voz de 102 indivíduos; 10 imagens e áudios por indivíduo	Face: rotação-invariante LBP Voz: MFCC	Nível de pontuação: produto	Face: Distância euclidiana Voz: MAP	TAR:100%, FRR:0%, FAR:0%, MT:0,341s
(DINESH; RAO, 2020)	Bimodal: Face e Voz	Identificação	Base de dados proprietária; 30 indivíduos	Face: LBP + PCA Voz: MFCC	Nível de característica	KNN	ACC: 98,4% FRR:0,25%, FAR:0,75%

Autores	*Modalidade*	*Tipo do sistema*	*Conjunto de dados*	*Características*	*Fusão*	*Decisão*	*Resultados*
(SINGH; KHANNA; GARG, 2020)	Bimodal: Face e Impressão Digital	Identificação	Face: Base de dados proprietária; Impressão digital: FVC-2004; 200 indivíduos; 1000 imagens	Face: PCA Impressão digital: Raymond Thai	Nível de característica: normalização mínimo-máximo	SVM + Distância euclidiana	ACC: 95,39% FAR:0% FRR:9,125% EER: 4,61% ET:0,189s
(MEHRAJ; MIR, 2021)	Multimodal: Orelha, Face e Marcha	Identificação	CASIA, EarVN e VidTI-MIT; Múltiplos ângulos de macha; imagens da orelha; faces coletadas em diferentes seções	Pre-trained CNN + HOG	Nível de característica: CCA e DCA	SVM+ Algoritmo de otimização bayesiana	ACC: 99,54% F1-SCORE: 99,54% MCC: 99,52% Kappa: 95,19%

2.2 DISCUSSÃO DOS TRABALHOS

Os trabalhos analisados fortalecem a necessidade de pesquisas na área, pois, com o advento dos *smartphones*, progressivamente mais informações estão associadas aos dispositivos, como cartão de crédito, senhas, informações bancárias e informações pessoais. Com o passar dos anos, percebeu-se que a criminalidade sofisticou seus métodos, compreendendo os mecanismos de autenticação, quebrando senhas e até mesmo produzindo traços biométricos. Diante do exposto, os autores, cujos trabalhos foram analisados, identificaram a carência de sistemas mais fortes e seguros e abordaram metodologias multimodais, agregando mais informações, variando instruções biométricas providas de múltiplos sensores, reduzindo as taxas de erro e aumentando a segurança.

Ao observar os trabalhos mencionados no Quadro 1, nota-se que os sistemas propostos são do tipo autenticação, em que se busca verificar se o indivíduo em questão é de fato ele mesmo, obtendo as características e padrões e comparando-as com um modelo que já foi salvo previamente em um banco de dados. Os sistemas de identificação, diferentemente dos de autenticação, tem por objetivo descobrir a identidade de um indivíduo. Assim, as características biométricas são testadas com outras treinadas pelo sistema. Não há certeza de que o indivíduo testado irá corresponder com algum outro indivíduo modelado pelo sistema.

Quando se analisam as bases de dados utilizadas, verifica-se que grande parte dos autores optaram por desenvolver suas próprias bases de dados, e a ausência de utilização de bases públicas dificulta a comparação dos resultados, pois cada base tem sua peculiaridade, níveis de dificuldade e critérios que foram adotados na coleta dos dados. Além disso, alguns trabalhos não detalharam se foi feito alguma divisão para treinamento e teste, ou mesmo como foi feita essa divisão, qual percentual de treinamento e teste, se houve ou não validação cruzada ou se foi aplicado algum método para a seleção das amostras.

A maioria dos sistemas podem ser divididos nas etapas de pré-processamento, extração das características, fusão e decisão. Observou-se que os trabalhos de Aronowitz et al. (2014), Kumar; Rao V (2019), Olazabal et al. (2019b), Singh, Khanna e Garg (2020) e de Zhang et al. (2020) utilizaram as técnicas clássicas de pré-processamento para os canais de face e voz. Pode-se citar as principais técnicas de processamento digital de imagens (PDI): equalização de histograma, conversão para níveis cinza, ajustes de intensidades, alinhamento da face, filtragem e extração da face através do algoritmo Viola-Jones. Assim como o canal de voz, destacaram-se: redução de ruído, filtro mediano detector de atividade de voz.

Com relação a extração das características, especialmente para os trabalhos que utilizaram a modalidade de face e voz, os autores optaram por utilizar técnicas clássicas (PCA, HOG, LBP para a face e MFCC para voz) e amplamente abordadas na literatura, desde os trabalhos mais antigos como o de Jiang, Sadka e Crookes (2010) e o de Kumar e Swamy (2010), aos mais recentes Dinesh; Rao (2020) e Zhang et al. (2020).

Diversamente, no trabalho de Zhang, Dai e Xu (2017) foi utilizada a transformada *wavelet* de Haar para extração tanto da face quanto para voz. Apenas o artigo de Zhang et al. (2020) aborda o interesse em desenvolver trabalhos futuros baseado em *framework* de aprendizado profundo para otimizar a acurácia da autenticação e reduzir os dados do treinamento. Dos trabalhos selecionados, somente o estudo de Mehraj e Mir (2021) utilizou aprendizado profundo através de redes CNN pré-treinadas, com transferência de aprendizado, para produzir vetores de características robustos. As redes testadas foram: AlexNet, Inceptionv2, Densenet201, Resnet101 e Resnet-Inceptionv2. Esse sistema biométrico utilizou os modos orelha, face e marcha.

A etapa de fusão foi realizada por dois tipos de métodos: em nível de característica e em nível de pontuação. Não foi possível identificar unanimidade entre os autores do melhor método de fusão. Os autores em geral apenas abordaram as características e qualidades dos métodos escolhidos. Nos trabalhos de Abozaid et al. (2019) e Singh, Khanna, Garg (2020) compararam-se os dois métodos. No primeiro trabalho conclui-se que a fusão por pontuação resulta nos melhores resultados para o modelo proposto, com menor valor de EER, considerando uma abordagem promissora para fusão. Em contrapartida, no segundo trabalho, a fusão das características atingiu o melhor desempenho de todos os experimentos. Dentre os métodos de fusão por pontuação destacam-se: LLR e fusão por produto. Para fusão por característica: DCA, CCA e normalização mínimo-máximo.

Os classificadores escolhidos para realizar a decisão foram os mais variados e conhecidos na literatura. Observou-se que a escolha não se deu pelas ferramentas mais recentes do estado da arte, mas sim pela opção do mais adequado para solução do sistema, buscando classificadores de simples implementação como SVM, KNN, quantização vetorial e distância euclidiana.

As métricas de avaliação dos sistemas multimodais utilizadas foram similares, o que facilitou a comparação dos trabalhos escolhidos. A principal métrica é a EER, que predetermina os valores de limiar para FAR e FRR. Quando esses dois valores são iguais, é obtido o EER. Assim, quanto menor o valor do EER, maior é a precisão do sistema biométrico. Outras métricas

utilizadas foram acurácia, F1-Score e tempo médio. Por fim, do Quadro 2, nota-se que, no decorrer dos anos, os desempenhos foram otimizados, evoluindo o estado da arte.

3 REFERENCIAL TEÓRICO

Neste capítulo são abordados conceitos que servirão de base para compreensão do trabalho. A Seção 3.1 exibe os conceitos sobre redes *autoencoder*. Os fundamentos teóricos sobre aprendizado por quantização vetorial são apresentados no Tópico 3.2. Posteriormente, serão apresentados conceitos da técnica de extração de características de voz MFFC. Por fim, na Seção 3.4 apresentam-se conceitos relacionados a redes neurais convolucionais em tarefas de classificação.

3.1 REDES AUTOENCODER

O *autoencoder* é uma técnica que tem como objetivo reconstruir os dados de entrada para a saída com a menor distorção possível. Possui ideia simples, porém, desempenha papel importante em aprendizagem de máquina. As ideias iniciais do *autoencoder* foram introduzidas no trabalho de Rumelhart, Hinton e Williams (1986) os quais desenvolveram modelos de aprendizagem, que através de propagação de erro, obtinham-se representações internas da entrada e, a partir disso, os *autoencoders* tornaram-se paradigmas fundamentais para aprendizagem não-supervisionada.

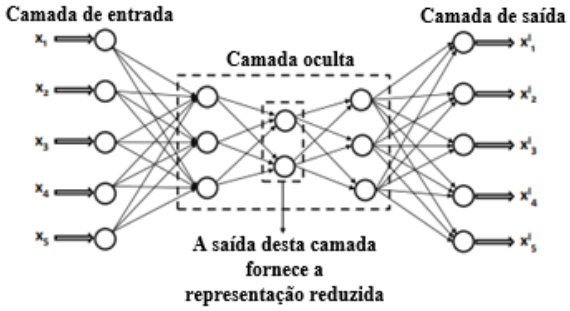
Posteriormente, através de Hinton e Salakhutdinov (2006), os *autoencoders* entraram em foco novamente. Nesse trabalho foi demonstrado que dados de dimensões elevadas poderiam ser codificados em uma dimensão menor utilizando o treinamento de uma rede neural profunda, sendo a camada central responsável por reconstruir os vetores de entrada de alta dimensão. A técnica proposta mostrou-se melhor do que PCA como ferramenta de redução de dimensionalidade, sendo sua principal aplicação atualmente.

O *autoencoder* mais conhecido é o do tipo padrão ou também chamado de Subcompleto. Sua função é copiar a entrada na saída, mantendo a quantidade de neurônios nas camadas de entrada e saída iguais. Seu nome deriva do fato da camada escondida ser de dimensão menor em relação a da entrada. Realizar essa tarefa em primeiro momento parece ser inútil, porém o que de fato interessa é obter uma camada interna de representação com propriedades relevantes para sua reprodução. Utilizar uma camada menor força o *autoencoder* a obter os dados mais relevantes para a representação da entrada (GOODFELLOW; BENGIO; COURVILLE, 2016).

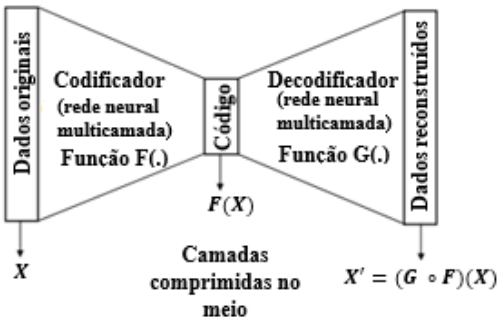
Ainda de acordo com Goodfellow, Bengio e Courville (2016), uma outra abordagem são os *autoencoders* regularizados. Esta variação surgiu devido ao problema dos subcompletos

não estarem aptos a trabalhar com uma grande quantidade de neurônios. Esse tipo de *autoencoder* pode ser não linear e sobrecompleto, mantendo, ainda assim, a capacidade de aprender dados úteis, uma vez que sua função de perda estimula a aprender propriedades adicionais. Pode-se citar o *autoencoder* esparso como um modelo regularizado que utilizada um fator de penalidade de esparsidade na camada interna de representação. Sua aplicação é voltada para classificação. Ressalta-se, também, que há os *autoencoders* do tipo *Denoising,* que tem por finalidade inserir uma penalidade na função de perda que permita o *Autoencoder* aprender uma informação útil alterando o critério de reconstrução.

A Figura 1(a) exibe a representação do *autoencoder* contendo três camadas ocultas. Evidencia-se que a camada mais interna está relacionada com as mais externas de forma hierárquica, o que possibilita uma redução graduada dos dados. De forma geral, um *autoencoder* possui uma arquitetura simétrica entre a sua entrada e saída. Na Figura 1(b), é possível visualizar a camada código, que é responsável pela representação reduzida dos dados de entrada, sendo a dimensionalidade da redução definida através do número de unidades dessa camada. A primeira parte da arquitetura caracteriza-se como codificação, sendo o agente que cria uma codificação reduzidas dos dados. Já a segunda parte, o decodificador, realiza a reconstrução dos dados a partir da camada código (AGGARWAL et al, 2018).



(a)



(b)

Figura 1: Esquema básico do *autoencoder* (a) esquema com três camadas escondidas; (b) esquema geral

Fonte: Adaptado de Aggarwal et al. (2018)

Aggarwal et al. (2018) descrevem a utilização do *autoencoder* para fatoração de matrizes. Essa versão é mais simples e contém apenas uma camada escondida. Sendo D a matriz de entrada, deseja-se fatorá-la em duas matrizes U e V, em que U contém a representação reduzida dos dados e V os valores de base, conforme a Equação 1. Basicamente, o problema consiste em aprender as matrizes U e V que minimizem a Equação 2, descrita como a norma de Frobenius da matriz residual, J:

$$D \approx UV^T \tag{1}$$

$$J = \|D - UV^T\|_F^2 \tag{2}$$

Deve-se observar que as linhas de D são entradas do *autoencoder*, as linhas de U são as ativações da camada escondida e V^T é o decodificador. A saída reconstruída contém as linhas de UV^T. O *autoencoder* busca reconstruir D através do produto das matrizes DW^TV^T, sendo W^T o codificador. Em outras palavras, o algoritmo do gradiente descendente tenta otimizar $\|D - DW^TV^T\|^2$.

3.2 APRENDIZADO POR QUANTIZAÇÃO VETORIAL

Aprendizado por Quantização Vetorial, LVQ (do inglês *Learning Vector Quantization*) pode ser definido como um método de classificação de padrões em que cada unidade de saída é representada por uma classe particular ou uma categoria. A técnica foi introduzida nos trabalhos de Kohonen (1989, 1990), o qual propôs as primeiras versões e melhorias sobre LVQ. Essa rede é do tipo aprendizado supervisionado e é possível utilizá-la em problemas de classificação binária e multiclasse. Sua estratégia de aprendizado é baseada em medidas de similaridades e o vencedor leva tudo.

A Figura 2 representa a arquitetura clássica do LVQ. A entrada e a saída possuem n e m neurônios, respectivamente. As camadas são totalmente interconectadas através dos pesos representados por W. A arquitetura basicamente possui *codebooks,* que são compostos por vetores de pesos. O objetivo da rede, após realizar o treinamento, é atribuir para um vetor de entrada a mesma classe da saída que tem o *codebook* mais próximo do vetor de entrada.

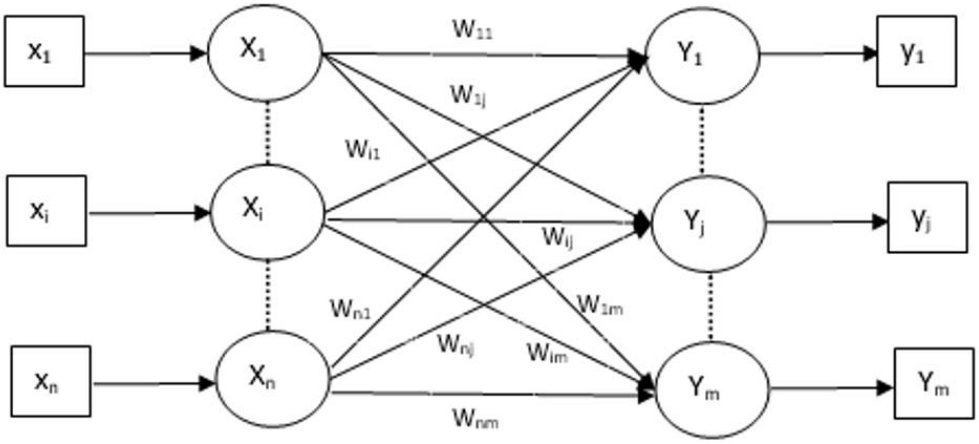


Figura 2: Arquitetura de uma rede neural LVQ

Fonte: Adaptado de Fausett (1994)

Resumidamente, o algoritmo pode ser descrito com os seguintes passos: primeiro, inicializam-se os vetores *codebook* e a taxa de aprendizagem. Em seguida, calcula-se a distância euclidiana $D(j)$ conforme a Equação 3 e, dessa forma, obtém-se a unidade J vencedora em que $D(j)$ é mínimo. Atualiza-se o vetor de pesos w_j conforme a regra da Equação 4. Reduz-se a taxa de aprendizagem e repete-se o processo até que a condição de parada seja atendida, podendo essa ser o número máximo de época atingido ou taxa de aprendizagem reduzida à um valor insignificante. Métodos como *K-means clustering* ou *the self-organizing map* podem ser utilizados para inicialização dos pesos do *codebook* (FAUSETT, 1994).

$$D(j) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_i - w_{ij}\right)^2 \tag{3}$$

$$\begin{cases} se\ T = C_j\ , & então \quad w_j(novo) = w_j(antigo) + \ \alpha[x - w_j(antigo)\ ] \\ se\ T \neq C_j\ , & então \quad w_j(novo) = w_j(antigo) - \ \alpha[x - w_j(antigo)\ ] \end{cases} \tag{4}$$

Sendo x o vetor de treinamento, T a classe do vetor x, w_j o vetor de pesos para a j-ésima unidade de saída e C_j a classe associada com a j-ésima unidade de saída.

O método LVQ permite dividir o espaço de dados em regiões distintas, definindo em cada região um *codebook*. Observa-se na Figura 3 uma representação do LVQ com três classes predefinidas para mapeamento, exibindo a localização final dos *codebook* após o treinamento, bem como os dados de entrada. Nesse exemplo, cada região é representada apenas por um único *codebook*, porém isso não é uma regra, visto que há possibilidade de alguma classe possuir mais *codebooks* do que outras.

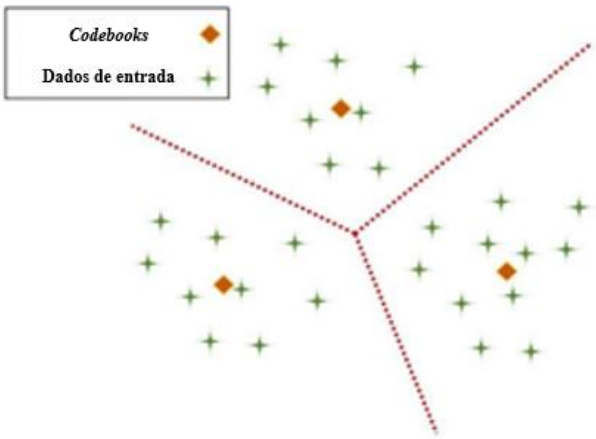


Figura 3: Representação do LVQ com 3 classes predefinidas.

Fonte: Adaptado de Maity (2013)

Conforme já mencionado, Kohonen foi responsável pela criação de variantes do LVQ, como por exemplo LVQ2.1 e LVQ3. Essas são consideradas de maior complexidade em relação ao LVQ1, uma vez que possuem conceitos de unidade vencedora e vice-campeã aprendem. Dentre outras modificações que foram exploradas pela comunidade científica, menciona-se o trabalho de Sato e Yamada (1995), que propuseram um *Generalized* LVQ (GLVQ) e sua contribuição baseia-se na atualização do *codebook*, utilizando o método do gradiente descente para minimizar a função de custo.

3.3 COEFICIENTES MEL-CEPSTRAIS

Os coeficientes Mel-Cepstrais habitualmente conhecidos como *Mel-frequency Cepstral Coefficients* (MFCC) caracterizam-se como uma das técnicas mais exploradas para extração de características em sinal de voz. Essa técnica é baseada no domínio da frequência utilizando a escala Mel, que foi desenvolvida no trabalho pioneiro de Steve, Volkman e Newman (1937), baseada na escala do ouvido humano. A técnica MFCC foi apresentada no trabalho de Bridle e Brown (1974) e posteriormente utilizada em reconhecimento de fala por Mermelstein (1976) e Davis e Mermelstein (1980). A partir de então, o MFCC foi amplamente difundido e utilizado pela comunidade cientifica e, até os dias atuais, a técnica é empregada para tarefas que visam o reconhecimento da fala e obtenção de características de um sinal de voz.

A descrição a seguir do processo de aplicação do MFFC foi baseada nos artigos de Martinez et al (2012) e Gupta et al. (2013). A transformação é dividida, basicamente, em sete etapas. Dado um sinal de entrada, realizam-se as filtragens pré-ênfase com objetivo de enfatizar

as altas frequências, aumentando a energia do sinal. Em seguida, aplica-se o *Framing* que segmenta a amostra da fala em N amostras. Essas janelas temporais normalmente são de 20 a 40ms ou 10 a 20 ms. O *Windowing* é aplicado com propósito de gerar um comprimento limitado dos frames, ou seja, a janela aplicada busca minimizar a distorção espectral. Há diversos tipos de janelas na literatura, como a retangular, topo de plano ou *Harmming*. De forma geral, a janela de *Harmming* (Equação 5) é a mais utilizada para essa tarefa.

$$w(n) = 0{,}54 - 0{,}46\ cos\left(2\pi\frac{n}{N}\right), \quad 0 \leq n\ \leq N \tag{5}$$

Em que N representa o número de amostras em cada *frame*.

Após o janelamento, efetua-se a *Fast Fourier Transform* (FFT), convertendo cada *frame* do domínio do tempo para o domínio da frequência. Em suma, a FFT é um algoritmo eficiente e mais rápido que aplica a *Discrete Fourier Transform* (DFT). A transformação via DFT ou FFT possui a mesma saída, mas suas diferenças estão na complexidade computacional. A DFT pode ser representada pela Equação 6:

$$X_k\ = \sum_{n=0}^{N-1} x_n\, e^{-2\pi i k\frac{n}{N}}, \qquad para\ k = 0, 1, 2\ , \cdots, N-1 \tag{6}$$

Sendo i denotado como a unidade imaginária $i^2 = -1$; x_n é o sinal no domínio do tempo; k é a frequência na direção n; N é o número de amostras.

O espectro obtido da FFT é largo. Considera-se que um sinal de voz segue uma escala linear apenas para frequências de 0 a 1000Hz e a partir dessa frequência segue uma escala logarítmica. Conforme já citado, no trabalho de Steve, Volkman e Newman (1937), a percepção subjetiva do som em relação a sua frequência foi estimada. A frequência Mel pode ser calculada pela Equação 7:

$$F(Mel) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

Em que f é a frequência em Hertz.

A seguir, aplicam-se filtros triangulares de banda passante variáveis espaçadas pelo escalamento de Mel. Essa filtragem tem o intuito de eliminar excesso de informações e obter uma estimativa na energia acumulada nas bandas críticas.

Por fim, executa-se a transformada discreta de cossenos, DCT (do inglês *Discrete Cosine Transform*), que leva o Espectro Mel para o domínio do tempo, obtendo os coeficientes Mel-Cepstrais. A DCT, por característica, possui seus primeiros coeficientes com uma maior concentração de energia. Normalmente, são utilizados os treze primeiros coeficientes. Há a possibilidade de calcular a primeira e segunda derivadas dos MFFCs para obtenção de

informação dinâmicas. Esses coeficientes são conhecidos como delta Cepstrais e delta-delta Cepstrais, respectivamente.

A Figura 4 exemplifica os passos citados para obtenção dos coeficientes Mel-Cepstrais:

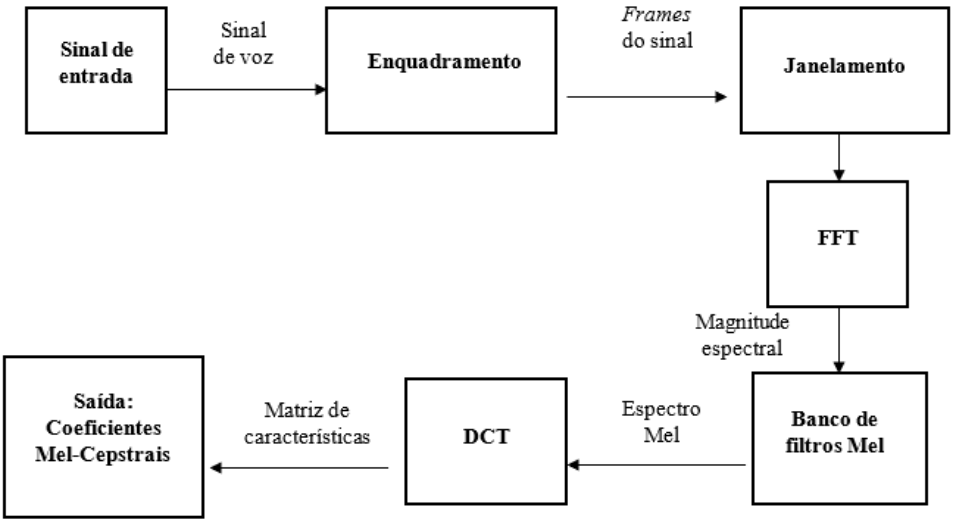


Figura 4: Arquitetura do MFCC.

Fonte: Traduzida de Chauhan e Desai (2014)

3.4 REDES NEURAIS CONVOLUCIONAIS

As Redes Neurais Convolucionais, CNNs (do inglês *Convolutional Neural Network*) são redes neurais amplamente difundidas e exploradas em diversos trabalhos científicos. Suas principais aplicações são no campo de visão computacional, reconhecimento facial, identificação de objetos ou até mesmo no processamento da fala. Em relação as redes anteriores à CNN, ela possui a capacidade de identificar as características mais relevantes de forma automática. Diferente das redes totalmente conectadas, a CNN foi desenvolvida para ser empregada com dados de entrada de duas dimensões. Esta foi inspirada em como as informações são processadas nas células do córtex visual. Em suma, os neurônios são compostos por campos receptivos locais, e cada um desses campos reage a um estímulo localizado em região limitada do campo visual. Esse modelo foi aplicado nas CNN através de pesos compartilhados e conexões locais, o que na prática cria operações com um número pequeno de parâmetros, simplificando o processo de treinamento da rede.

A primeira versão de CNN foi introduzida por Lecun et al. (1998), a rede *LeNet-5*. Essa arquitetura foi utilizada para reconhecimento de caracteres manuscritos e impressos por máquinas (Figura 5). No decorrer dos anos, surgiram diversas variações de arquiteturas para serem aplicadas em problemas distintos. Todavia, as principais camadas de uma CNN, são: a

camada de Convolução, *Batch Normalization*, ReLu, *Pooling*, *Dropout* e Classificação. Cada uma destas camadas será detalhada a seguir.

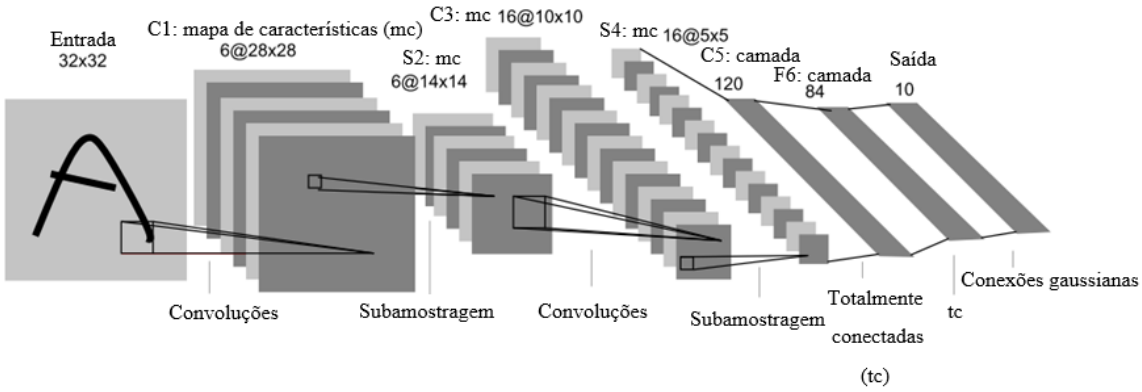


Figura 5: LeNet-5 (A primeira versão de CNN)

Fonte: Lecun et al. (1998)

3.4.1 Camada de convolução

A convolução é uma operação entre duas funções reais que combinadas formam uma terceira função expressando como a forma de uma é modificada pela outra. No contexto de aprendizado de máquinas, esta camada é responsável por criar o mapa de características dos dados de entrada, tornando-a mais importante dentre as demais camadas das CNNs. Pode-se representar a operação de convolução através da Equação 8, sendo x o sinal a entrada, w o *kernel* e $s(t)$ o mapa de características. Nestes tipos de aplicações, é comum que as CNNs utilizem as imagens da entrada e do *kernel* como matrizes multidimensionais (GOODFELLOW; BENGIO; COURVILLE, 2016). Imagina-se uma imagem *I* bidimensional como entrada, K kernel bidimensional. A convolução discreta bidimensional $S(i,j)$ é descrita na Equação 9:

$$s(t) = \int x(a)w(t-a)da = (x * w)(t) \tag{8}$$

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n) \tag{9}$$

x é o sinal de entrada; w é o *kernel*; $s(t)$ é o mapa de características; I é a imagem bidimensional.

A Figura 6 exibe uma exemplificação gráfica da operação de convolução bidimensional. Tem-se a entrada, o kernel e a saída com dimensão $3x4$, $2x2$ e $3x4$, respectivamente. Neste caso, a saída está limitada às regiões em que o *kernel* está contido na imagem de entrada.

Percebe-se que a dimensão de saída pode ser diferente da dimensão entrada, pois depende do tamanho *kernel* utilizado. Cada elemento de saída é a soma dos produtos de cada elemento da entrada pelos elementos do *kernel* em suas posições correspondentes.

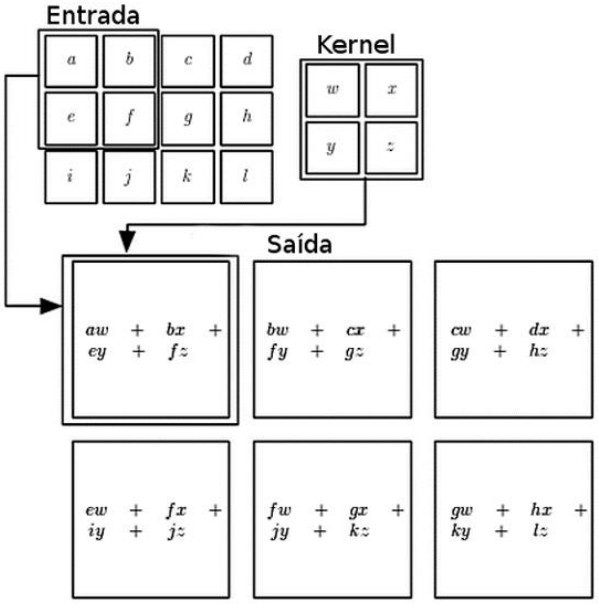


Figura 6: exemplo da convolução 2-D.

Fonte: adaptado de Goodfellow, Bengio e Courville (2016)

3.4.2 Camada de normalização em lotes

Proposto no trabalho de Ioffe e Szegedy (2015), a *Batch Normalization* é uma das mais recentes contribuições para as CNNs. O método surgiu para atuar em dois problemas. O primeiro problema ocorre quando os parâmetros da rede mudam durante o treinamento, o que implica na mudança das ativações das camadas ocultas (problema conhecido como mudança de covariância interna). Como resultado, a convergência ocorre mais lentamente devido à instabilidade dos dados de treinamento para camadas posteriores. Este efeito pode ser minimizado utilizando uma camada de normalização em lotes *(batch normalization)*. Sua outra atuação é no problema de explosão do gradiente, que ocorre devido ao acúmulo do erro do gradiente que é atualizado durante o treinamento. Dessa forma, a atualização de peso se torna muito grande, o que torna a rede instável e sem a capacidade de aprender com os dados do treinamento (AGGARWAL et al, 2018).

De acordo com Ioffe e Szegedy (2015), a normalização em lote adiciona apenas dois parâmetros extras por ativação, que são aprendidos no treinamento (γ e β). Considera-se um minilote $B = \{x_1, x_2, x_3, \cdots, x_m, \}$, μ_B a média do minilote, σ_B^2 a variância do minilote, $\hat{x}_i$ os

valores normalizados e y os valores escalados e deslocados pelos fatores γ e β e ε uma constante numérica para estabilidade numérica. Tem-se:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{10}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{11}$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_i^2 + \varepsilon}} \tag{12}$$

$$y = \gamma \hat{x}_i + \beta \tag{13}$$

3.4.3 Camada ReLU

A Unidade Retificadora Linear, *ReLU* (do inglês *Rectified Linear Unit*) realiza uma transformação que só ativa o elemento da entrada se estiver acima de uma certa quantidade. Dessa forma, se entrada estiver abaixo de zero, a saída é transformada para zero, porém, se a entrada estiver acima de um certo limiar, tem-se uma relação linear com a entrada. A utilização da ativação ReLU não altera a dimensão da camada pois é realiza um mapeamento 1:1 para os valores de ativação (PATTERSON; GIBSON, 2017). Pode-se defini-la pela Equação 14, demonstrada pela Figura 7:

$$y = f(x) = max\ (0, x) \tag{14}$$

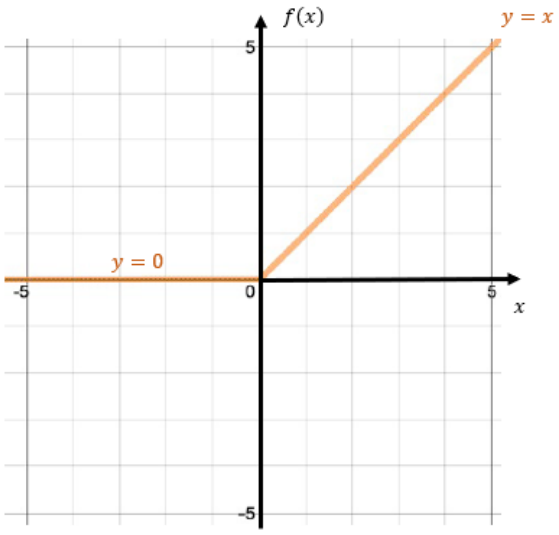


Figura 7: função de ativação ReLU.

Fonte: Patterson e Gibson (2017).

Destaca-se que usar a função de ativação ReLU foi uma evolução dentro das redes neurais em comparação com as funções sigmoide e tangente hiperbólica. Suas principais vantagens são a velocidade e a precisão. Isso ocorre devido a sua derivada ser 1 para x diferente de zero e 0 nos demais casos, reduzindo e simplificando de forma relevante a quantidade de cálculos para treinamento da rede. Este aumento de velocidade permitiu criar modelos mais profundos e treiná-los por mais tempo. A ReLU também possui a capacidade de evitar o problema da "explosão do gradiente", visto que seu gradiente é zero ou uma constante, evitando a saturação da função para valores positivos. Há algumas variantes da ReLU, como LeakyReLU, que ao invés de usar 0 para valores onde $x < 0$, possui uma pequena inclinação $0.01x$. Ou também a *Softplus*, que é considera uma versão com uma curva mais suave da ReLU (PATTERSON; GIBSON, 2017).

3.4.4 Camada de Pooling

Devido à associação de sucessivas camadas de convolução em uma CNN, necessita-se de uma camada para reduzir a dimensão da entrada, obtendo uma representação menor dos dados que serão propagados pela rede e reduzindo ainda o *overfitting*. Assim, o estágio de subamostragem substitui a saída da rede em uma certa localização, por uma medida estatística dos valores em sua vizinhança.

A camada de *pooling* opera sobre uma vizinhança retangular ($2x2$, $3x3$, $4x4$, dentre outras). Há diferentes técnicas para realizar essa filtragem. Por exemplo, o *max pooling* realiza a substituição pelo valor máximo da vizinhança. Já o *average pooling* substitui pelo valor médio e o L^2*pooling* troca pela norma L^2 da vizinhança. A Figura 8 exibe o exemplo do *max pooling* aplicado para um mapa de características inicial de dimensão $4x4$ com profundidade 1, com os parâmetros passo 2 e filtro $2x2$, gerando um mapa final $2x2$.

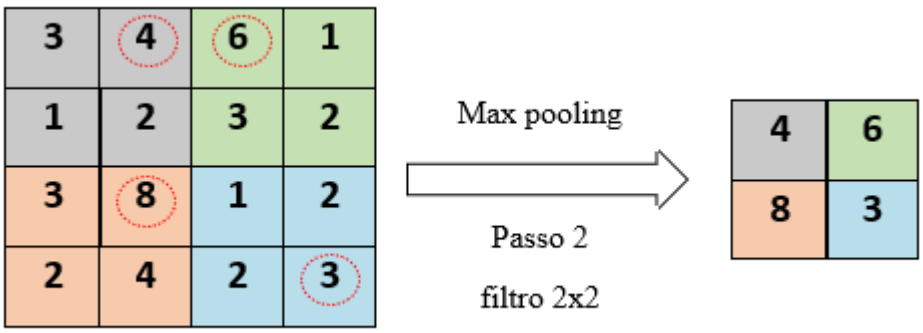


Figura 8: Aplicação do max pooling

É importante destacar que o *pooling* pode ser considerado como invariante para pequenas translações na entrada. Essa propriedade é importante para situações em que se está mais interessando em saber se uma determinada característica está presente na imagem e não exatamente onde está localizada. Assim, em diferentes imagens com localização e formas distintas, a invariância à translação permite classificar as imagens de forma similar (AGGARWAL et al, 2018).

Uma outra abordagem que vem ganhando notoriedade é *Global Average Pooling* (GAP). Essa camada foi proposta no trabalho de Lin, Chen e Yan (2013) e foi projetada para substituir as camadas *flatten* e as totalmente conectadas. A GAP está posicionada entre os mapas de características e a camada de classificação. A camada *flatten* realiza um "achatamento", convertendo os dados em uma matriz unidimensional, tornando a saída das camadas convolucionais em um longo vetor de características. Em contrapartida, a GAP gera um mapa de características para cada categoria de classificação na última camada, obtendo a média de cada mapa de característica e o vetor resultante é inserido na camada de classificação. A Figura 9 ilustra as diferenças entre as camadas *flatten* e GAP.

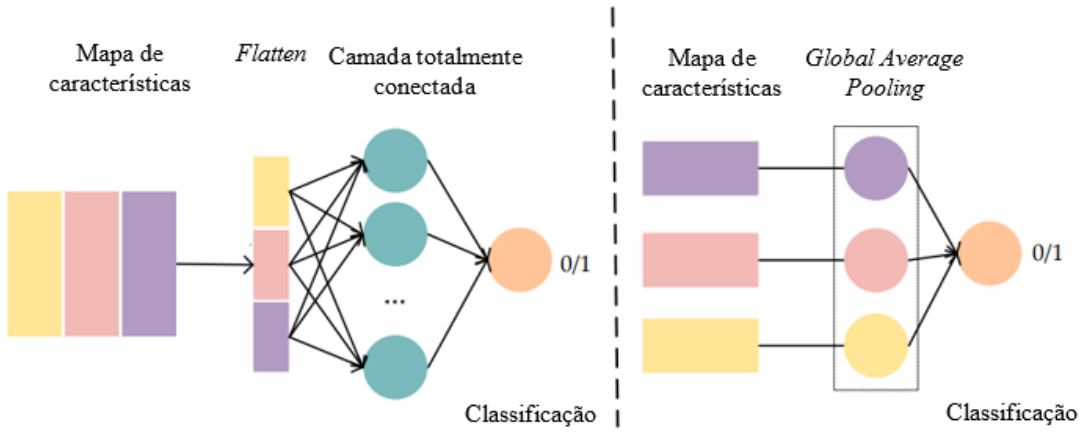


Figura 9: Diferenças entre as camadas *Flatten* e GAP.

Fonte: Li et al. (2020).

Uma vantagem do GAP sobre as camadas totalmente conectadas é que o GAP realiza um processo mais nativo à estrutura de convolução, impondo correspondências entre mapas de característica e as categorias. Assim, os mapas de características podem ser interpretados como mapas de confiança de categorias. Outra vantagem é que não há parâmetro para otimizar no GAP, reduzindo o *overfitting* nesta camada (LIN; CHEN; YAN, 2013).

3.4.5 Dropout

Desenvolvido por Srivastava et al. (2014), a técnica *dropout* é um mecanismo usado para melhorar o treinamento de redes neurais omitindo alguns neurônios ocultos. É considerada uma grande contribuição para prevenir *overfitting* em redes CNNs. Basicamente, ela atua desligando de forma aleatória um neurônio para que este não contribua nem para o avanço e nem para a retropropagação da rede. O termo "*Dropout*" não se refere a eliminar esses neurônios ocultos e sim desativá-los de forma temporária da rede. Para realizar essa tarefa, durante o treinamento, a *Dropout* gera uma nova rede composta, utilizando apenas os neurônios que não foram desativados. Ao final, obtém-se uma média da predição das diversas redes criadas. A escolha de quais neurônios descartar é aleatória. Sua parametrização é feita em porcentagem, escolhendo o percentual de neurônios que será desativado da camada anterior. Na Figura 10, ilustram-se dois modelos, um com e o outro sem utilização de *Dropout.*

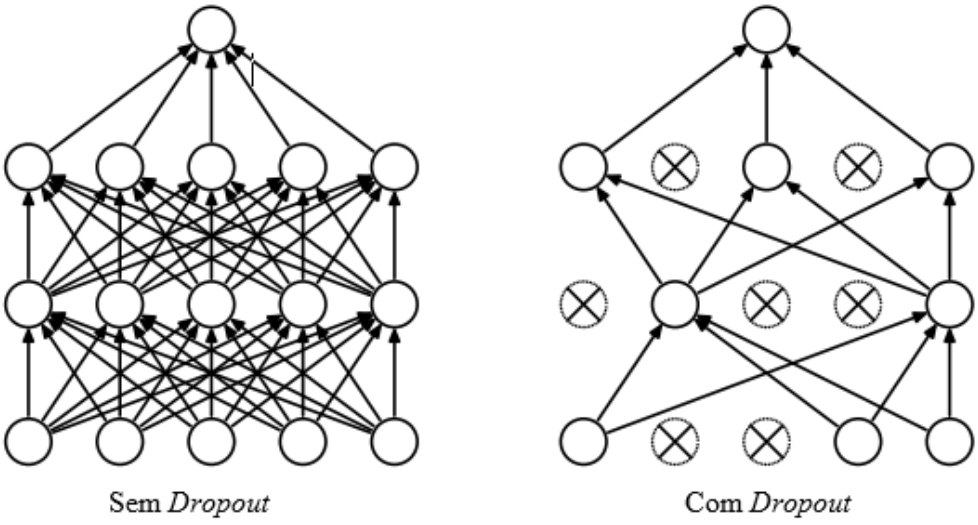


Figura 10: No lado esquerdo, tem-se a rede neural padrão com duas camadas escondida e no lado direito, a rede produzida ao aplicar o *Dropout* na rede da esquerda.

Fonte: Srivastava et al. (2014).

3.4.6 Classificação

As CNNs podem ser utilizadas com variações em sua arquitetura para atender diferentes problemas, como por exemplo, classificação de imagens ou segmentação de imagens. Para problemas de classificação, a CNN é dividida em duas partes: convolucional e densa. A primeira etapa é responsável por aprender a extrair as melhores características da imagem e a segunda aprende a classificar essas características em diferentes classes. A Figura 11

exemplifica as duas etapas presentes em um CNN classificadora de imagens. A etapa de classificação é composta de uma rede neural artificial que é conhecida como *Perceptron* Multicamada, em que sua arquitetura contém uma entrada, camadas ocultas e a camada de saída.

O vetor *Flatten* ou GAP é usado como entrada e contém as informações geradas na etapa de aprendizagem de características. A camada oculta é composta pelos neurônios que serão aprendidos na etapa de treinamento. Essa etapa pode conter uma ou mais camadas ocultas, compondo uma rede densa. A camada de saída também é uma sequência de neurônios, porém, contém uma função de ativação diferente. Pode-se citar algumas funções de ativação já consolidadas em aprendizado de máquina: *Sigmoide*, *ReLU*, *Hardlim*, *Tanh* e *Softmax*. Neste trabalho, mais especificamente, será utilizada a *Softmax*, que é uma função de ativação voltada para aplicações multiclasse.

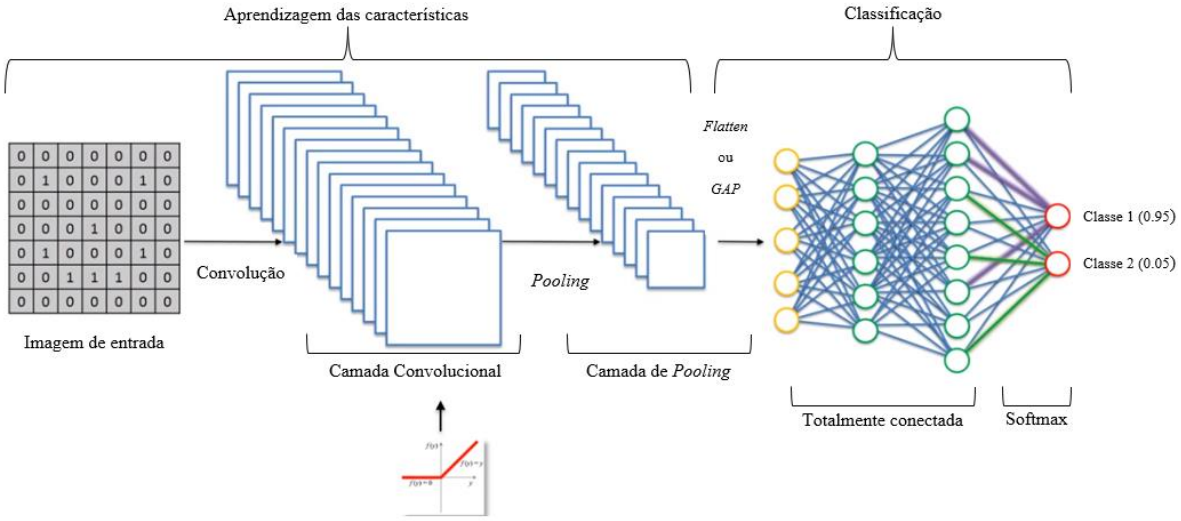


Figura 11: Etapas de uma rede CNN classificadora de imagem.

Fonte: adaptado de Kplanoglou (2017)

A *Softmax* possui como entrada valores positivos, negativos ou mesmo zero, que foram providos da penúltima camada e esses não estão dimensionados adequadamente. Assim, como resultado, a *Softmax* gera valores entre 0 e 1, que podem ser interpretados como probabilidades. Sua aplicação é voltada para problemas com classes que são mutualmente exclusivas. A ideia básica da *Softmax* é converter os valores para distribuição de probabilidades normalizada. De acordo com Goodfellow, Bengio e Courville (2016), pode-se definir matematicamente a função Softmax como:

$$softmax(z)_i = \frac{e^{z_i}}{\sum_{J=1}^{K} e^{z_j}} \tag{15}$$

Sendo z o vetor de entrada e K o número total de classes.

O princípio de funcionamento da *Softmax*, exemplificado na Figura 12, baseia-se em um vetor de saída compondo 5 classes e, após aplicação do bloco de ativação, encontram-se as probabilidades de cada classe, que somadas totalizam 1. Neste exemplo, a classe 2 teve a maior probabilidade de ser a classe correta com 90%.

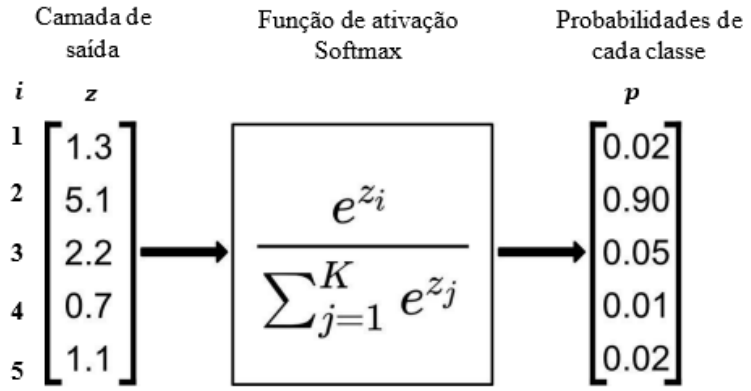


Figura 12: princípio de funcionamento da função *Softmax*.

Fonte: adaptado de Nabiyev e Malekzadeh (2021).

3.4.7 Treinamento de uma CNN

O processo de treinamento de uma rede neural convolucional é supervisionado e utiliza o algoritmo do *backpropagation*. Este consiste em obter a predição de saída quando uma entrada é apresentada a rede. Se a saída corresponder ao *target* em questão, nada é realizado. Caso não corresponda, é realizada a atualização dos parâmetros internos da rede com objetivo de minimizar o erro.

Devido à grande quantidade de pesos a serem atualizados, o *backpropagation* é considero lento, porém a utilização de unidades de processamento gráfico (GPU, do inglês *Graphics Processing Unit*), que estão mais sofisticadas e acessíveis, contribuem para viabilização do treinamento de uma CNN (PATTERSON; GIBSON, 2017). Além das GPUs, outra ferramenta de *hardware* utilizada no desenvolvimento de aplicações de redes neurais e aprendizado de máquina é a TPU (do inglês *Tensor Processing Unit*). A TPU foi criada pelo google em 2015 e introduzida ao público em 2018. A solução *Cloud* TPUs é considerada incrivelmente rápida na execução de cálculos densos de vetores e matrizes, permitindo o aprendizado de máquina de uma rede neural no *software TensorFlow*.

O algoritmo *backpropagation* contém três passos: inicialização dos pesos da rede de forma aleatória; a etapa de *forward*, em que a entrada é propagada pela rede até a camada de saída, obtendo o erro resultante; e a etapa *backward*, em que os pesos e polarizações são

atualizados através do gradiente descendente. Repete-se este processo até que o erro seja minimizado. A atualização dos pesos e polarizações é dada através das equações 16 e 17. A função de custo mais utilizada é o erro quadrático médio, conforme a equação 18:

$$W(t+1) = W(t) - \alpha \frac{\partial L}{\partial W} \tag{16}$$

$$b(t+1) = b(t) - \alpha \frac{\partial L}{\partial b} \tag{17}$$

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{18}$$

Sendo α o fator de aprendizagem.

Os pesos iniciais são definidos com valores aleatórios e são atualizados através da regra do gradiente descendente. A derivada do erro em relação a cada peso é calculada através da regra da cadeia, propagando o erro para trás. Dessa forma, o algoritmo distribui a contribuição do erro para cada peso na rede. A derivada parcial do gradiente da última camada (camada de saída) é usada para calcular a camada imediatamente anterior até que o erro seja propagado até a camada de entrada.

Na Figura 13(a) é ilustrado o processo aplicado para uma CNN, em que o erro entre a saída e o *target* é passado para trás e então é calculado o gradiente do erro em relação aos pesos W_1 e W_2, para posteriormente serem atualizados através da regra do gradiente descendente. Na Figura 13(b) exemplifica-se o passo para frente em que é computado a ativação do neurônio e o passo para trás que aplica a regra da cadeia.

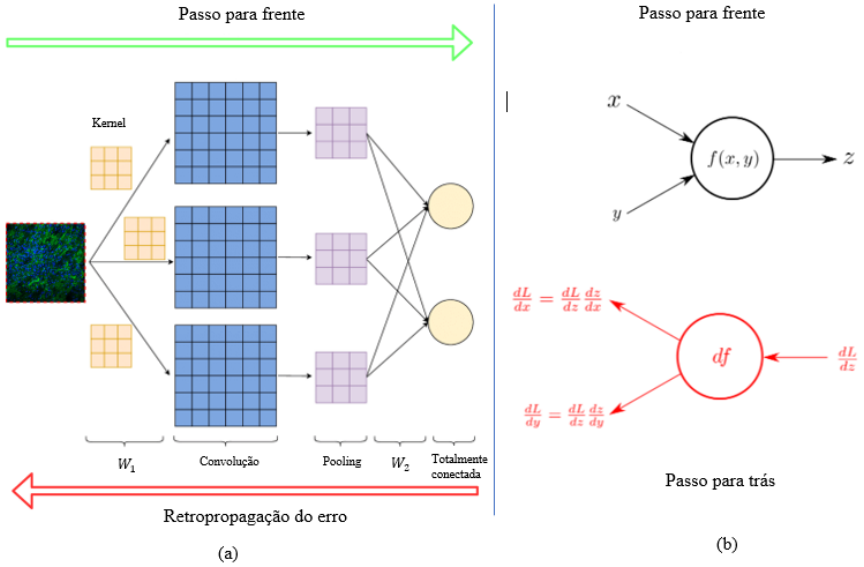


Figura 13: Processo de treinamento de uma CNN usando *backpropagation*.

Fonte: adaptado de Bazaga et al. (2019).

3.4.8 Métodos de otimização

Os métodos de otimização são responsáveis por permitir continuar atualizando os parâmetros do modelo e minimizar a função de perda, fornecendo resultados mais precisos possíveis. Há diversos otimizadores utilizados para treinamento da rede disponíveis na literatura, porém, neste trabalho, optou-se por realizar simulações com os métodos Gradiente Descendente Estocástico, SGD (do inglês *Stochastic Gradient Descent*), Propagação da Raiz Média Quadrática, RMSProp (do inglês *Root Mean Square Propagation*) e Estimativa de Dinâmica Adaptativa, ADAM (do inglês *Adaptive Moment Estimation)*. A variação desses três otimizadores nas redes CNNs, teve o intuito de identificar qual teve melhor desempenho nas redes propostas. A seguir, detalha-se cada um dos otimizadores.

3.4.8.1 Gradiente descente estocástico

O SGD pode ser considerado um dos algoritmos de otimização mais utilizados no campo de aprendizado de máquinas. Este é uma extensão do algoritmo do gradiente descendente que surgiu devido a problemas de se treinar grandes partições e obter uma boa generalização, visto que o treinamento de grandes partições possui um custo computacional maior (GOODFELLOW; BENGIO; COURVILLE, 2016). Seja um conjunto de treinamento com n amostras, sendo L_i a função de perda em relação ao index i, θ o parâmetro da função e $J(\theta)$ a média das funções de perda (equação 19), então:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} L_i(\theta) \tag{19}$$

O gradiente da função $\mathrm{J}(\theta)$ é calculado pela equação 20:

$$\nabla_\theta J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta L_i(\theta) \tag{20}$$

Ao se analisar a equação 20, verifica-se que o custo computacional de cada variável independente é $O(m)$, crescendo linearmente com m. Assim em partições muito grandes, com milhões ou bilhões de amostras, o tempo para cada iteração do gradiente será muito longo. O ponto chave do SGD é utilizar o gradiente como uma expectativa, estimando para um conjunto pequeno de amostras. Seja B um mini *batch* contendo m' amostras, compondo um conjunto relativamente menor que o conjunto total de tamanho m. Dessa forma a estimativa do gradiente é definida pela equação 21 e atualização de θ pela equação 22:

$$g = \frac{1}{m'} \nabla_\theta \sum_{i=1}^{m'} L_i(\theta) \tag{21}$$

$$\theta \leftarrow \theta - \eta g \tag{22}$$

em que, η é taxa de aprendizagem.

O SGD reduz o custo computacional em cada interação para $O(1)$, tornando uma ferramenta crucial em aprendizado de máquina, visto que cada atualização do SGD não depende mais do tamanho da partição de treinamento m.

3.4.8.2 Propagação da raiz média quadrática

O algoritmo RMSProp proposto por Tieleman e Hinton (2012), possui motivação similar ao algoritmo AdaGrad, utilizando uma normalização com magnitude $\sqrt{r}$. Nele, entretanto, foi realizada uma modificação para melhorar o desempenho em uma configuração não convexa, visto que o mesmo foi projetado para convergir rapidamente para funções convexas. Todavia, pode-se deparar com uma região que seja localmente convexa. A ideia por trás do AdaGrad é reduzir a taxa de aprendizagem de acordo com todo o histórico do gradiente ao quadrado, já o RMSProp utiliza uma média exponencialmente decrescente para descartar os valores mais antigos (GOODFELLOW; BENGIO; COURVILLE, 2016).

Seja g o gradiente da função (equação 26), η a taxa de aprendizagem, r o valor da média exponencial ponderada (equação 24), ρ o fator de decaimento para ponderação dos dados e δ uma pequena constante para estabilizar a divisão. Tem-se a atualização do parâmetro θ pela equação 25:

$$g = \frac{1}{m} \nabla_\theta \sum_{i=1}^{m} L_i(\theta) \tag{23}$$

$$r \leftarrow \rho r + (1 - \rho) \left(\frac{\partial L}{\partial \theta}\right)^2 \tag{24}$$

$$\theta \leftarrow \theta - \frac{\eta}{\sqrt{r} + \delta} \left(\frac{\partial g}{\partial \theta}\right) \tag{25}$$

3.4.8.3 Estimativa de dinâmica adaptativa

O ADAM é um algoritmo de otimização de taxa de aprendizado adaptativo que requer somente o gradiente de primeira ordem, permitindo a utilização de pouca memória. O método foi desenvolvido para combinar vantagens do Adagrad, que trabalha bem com gradientes esparsos e o RMSProp, que funciona bem em configurações não estacionarias (KINGMA; BA, 2015). Um ponto chave do ADAM é que ele usa médias móveis ponderadas exponenciais para obter uma estimativa do momento e do segundo momento do gradiente.

Seja g o gradiente da função (equação 26), ρ_1 e ρ_2 taxas de decaimento exponencial com valores entre [0,1), η a taxa de aprendizagem, δ uma pequena constante para estabilizar a divisão, s a estimativa do primeiro momento do bias, r estimativa do segundo momento do bias, $\hat{s}$ é o bias correto do primeiro momento, $\hat{r}$ é o *bias* correto do segundo momento. A atualização do parâmetro θ é feita pela equação 31:

$$g = \frac{1}{m}\nabla_\theta \sum_{i=1}^{m} L_i(\theta) \tag{26}$$

$$s \leftarrow \rho_1 s + (1-\rho_1)\left(\frac{\partial g}{\partial \theta}\right) \tag{27}$$

$$r \leftarrow \rho_2 r + (1-\rho_2)\left(\frac{\partial g}{\partial \theta}\right)^2 \tag{28}$$

$$\hat{s} \leftarrow \frac{s}{1-p_1^t} \tag{29}$$

$$\hat{r} \leftarrow \frac{r}{1-p_2^t} \tag{30}$$

$$\theta \leftarrow \theta - \frac{\hat{s}}{\sqrt{\hat{r}} + \delta} \tag{31}$$

4 MATERIAIS E MÉTODOS

A metodologia proposta neste trabalho tem por objetivo comparar os métodos multimodais para reconhecimento de indivíduos. A primeira abordagem, implementa três sistemas com reconhecimento utilizando LVQ. Um sistema unimodal para face, um sistema unimodal para voz e, o terceiro, um sistema multimodal face-voz. A segunda abordagem utiliza aprendizado profundo através de CNNs e, de forma análoga, desenvolveram-se dois sistemas unimodais e um multimodal. A ideia chave é utilizar o mesmo conjunto de dados, MOBIO, que contém amostras de face e voz adquiridas através de dispositivos celulares.

No Capítulo 3, os conceitos fundamentais para implementação desse trabalho foram apresentados. Segue-se com a apresentação dos materiais necessários para a implementação, os métodos aplicados e o detalhamento das arquiteturas propostas.

4.1 MATERIAIS

4.1.1 Definição do ambiente de trabalho

Para a elaboração deste trabalho, utilizou-se uma conta *Google Colaboratory PRO,* uma ferramenta que permite a alocação e a utilização de *Hardware* na nuvem acessível através de qualquer navegador. A máquina padrão disponível possui processador Intel® Xeon® CPU @ 2.30GHz *Cache* 56 MB, 16 GB de memória RAM e GPU 16 GB de Memória NVIDIA Tesla P100PCIe.

Para a implementação, utilizou-se a linguagem *Python* com as bibliotecas *TensorFlow, Keras, Sklearn, Numpy, Matplotlib, Pillow, MTCNN, python_speech_features e Seaborn.*

4.1.2 Conjunto de dados

A base de dados utilizada no trabalho foi a MOBIO, apresentada em Mccool et al. (2012). Essa é uma base de dados bimodal que consiste em dados de áudio e vídeo de 158 indivíduos, contendo uma relação homem-mulher 2:1. Para cada indivíduo, geraram-se 192 amostras de áudio e vídeo distintas. A construção dessa base foi realizada em agosto de 2008 até julho de 2010, em 6 lugares distintos de 5 diferentes países. Esta diversidade de países é

composta por falantes nativos e não nativos de Inglês. Utilizou-se 50 indivíduos, sendo 37 homens e 13 mulheres.

Um ponto chave da MOBIO é que a aquisição de dados foi realizada de forma não controlada. O microfone e a câmera não foram fixados e o dispositivo de aquisição foi fornecido para o usuário, o qual interagiu de diversas maneiras. De acordo com os autores, a escolha desse método gerou diversos desafios: alta variabilidade de poses e condições de iluminação, alta variabilidade de qualidades de voz e variações de ambientes de aquisições como iluminação, plano de fundo e ambientes acústicos. A Figura 14 exemplifica essas diferentes condições para dois indivíduos.

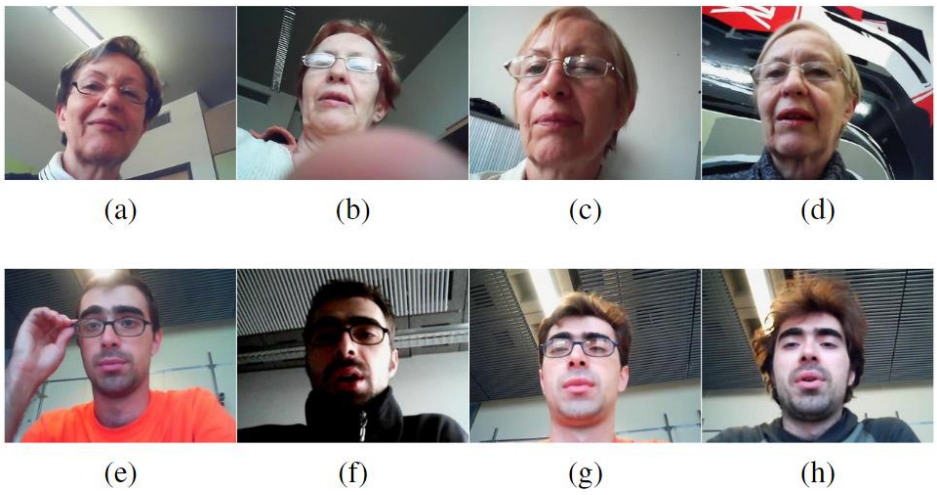


Figura 14: Exemplo de imagens de dois indivíduos. Visualiza-se as diferenças de poses, iluminação, estilos de cabelo, maquiagem e utilização de óculos.

Fonte: MCCOOL et al. (2012)

A aquisição de voz foi coletada com respostas para perguntas do tipo: resposta curta, discurso livre de resposta curta, discurso definido e discurso livre. As repostas para questões curtas foram relacionadas a 5 questões pré-definidas. As repostas curtas livres foram obtidas através de questões escolhidas aleatoriamente em um conjunto de 40 questões, no qual a resposta do usuário foi gravada por aproximadamente 5 segundos. A fala livre foi capturada através de leituras de textos pré-definidos em voz alta. A duração das leituras foi de, aproximadamente, 10 segundos. Por fim, as falas livres foram providas de 10 questões aleatórias de um total de 30, com respostas para cada questão em torno de 10 segundos.

Para aquisição dos dados foram utilizados dois dispositivos: um celular NOKIA N93i e um *Laptop* MacBook 2008. A utilização do *Laptop* foi apenas na primeira parte da aquisição, de um total de 12 sessões realizadas, compondo um conjunto minoritário.

4.2 MÉTODOS

Nesta seção serão abordados todos os aspectos metodológicos da pesquisa realizada, descrevendo-se os procedimentos necessários para o desenvolvimento de métodos multimodais (face e voz) de reconhecimento biométrico de indivíduos que utilizam uma abordagem de aprendizado de máquina clássica e uma abordagem com aprendizado de máquina profundo para comparar os desempenhos das arquiteturas em diferentes cenários de generalização e otimização. Trata-se de um trabalho de natureza aplicada que utilizou tanto uma abordagem qualitativa quanto quantitativa. A qualitativa envolveu o delineamento do contexto do ambiente da pesquisa mas, primordialmente, o volume de experimentos realizados denota que a natureza primordial desse trabalho é quantitativa. Adicionalmente, com intuito de contextualizar e evidenciar o estado da arte no tema foi realizada uma pesquisa bibliográfica prévia.

Para atingir os objetivos gerais e específicos deste trabalho, foram realizadas as etapas apresentadas no diagrama de blocos da Figura 15. Cada uma das etapas expostas neste diagrama está descrita nas próximas seções:

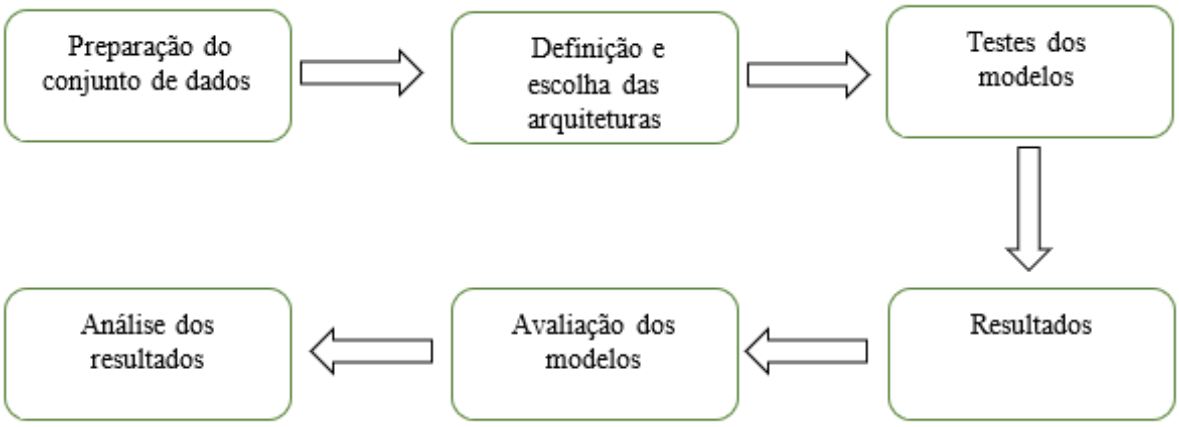


Figura 15: Etapas da metodologia utilizada

4.2.1 Preparação do conjunto de dados

Para este trabalho, utilizaram-se duas partições, denominadas de LIA e UNIS, do conjunto total disponível da base de dados, compondo um total 50 indivíduos, sendo 37 homens e 13 mulheres. Cada indivíduo é possuidor de 192 amostras de face-voz. Assim, foram utilizadas 9600 amostras de voz e 9600 amostras de face.

Realizou-se um pré-processamento nas amostras iniciais de face buscando obter somente a face e remover informações desnecessárias do plano de fundo da imagem. Para esta tarefa, utilizou-se o pacote MTCNN que consiste na implementação de detector de face proposto no artigo Zhang et al. (2016). Como resposta, foram obtidas as coordenadas que compõem a caixa de seleção da face. Através da biblioteca *Pillow* com a função *resize,* realizou-se uma interpolação bicúbica, padronizando cada imagem de face na dimensão 64x80. Por fim, converteram-se as imagens em monocromática, a fim de reduzir a dimensão da base de dados de face. O pré-processamento de uma amostra da face levou em torno de 1 segundo. A Figura 16 exemplifica o processo de pré-processamento realizado para o canal de face.

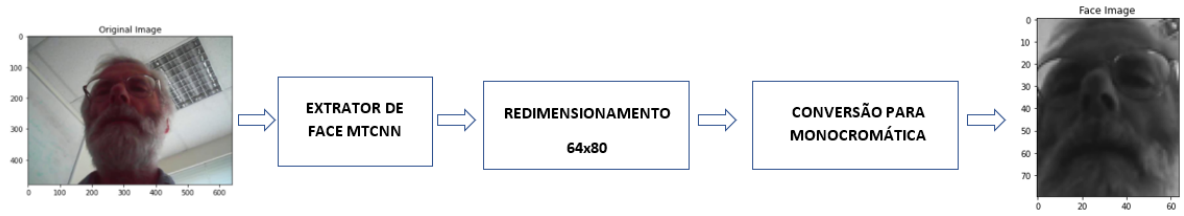


Figura 16: Pré-processamento para o canal de face

Para as amostras originais de voz, inicialmente, os sinais de vozes foram processados por um detector de atividade de voz (DAV), desenvolvido pelo Google para o projeto *Web real time comunication*, que fornece o pacote *webrtcvad* em linguagem Python. A função do detector é classificar os trechos de áudios com ou sem voz, removendo trechos ruidosos ou em silêncio. Quanto a configuração do DAV, escolheu-se um fator de agressividade 3, dentre a escala de 0 a 3. O tamanho dos frames foi de 512 amostras (32ms), com espaçamento entre as amostras de 8ms. Assim, de cada amostra de voz foram extraídos 192 frames. Somente os frames classificados como voz foram armazenados. O pré-processamento de uma amostra da voz levou em torno de 2 segundos. A Figura 17 exemplifica o processo de pré-processamento realizado para o canal de voz.

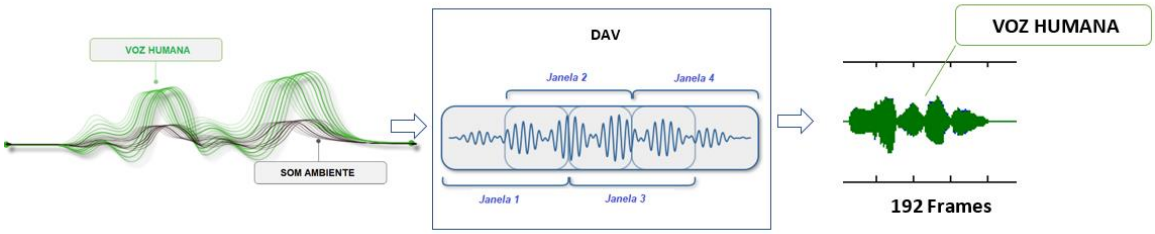


Figura 17: Pré-processamento para o canal de face

O conjunto de dados foi subdividido, de forma aleatória, nos conjuntos de treinamento e de teste, com as porcentagens: 50% para treinamento e 50% para teste. A divisão realizada foi balanceada em cada uma das 50 classes, ou seja, de cada indivíduo, separaram-se 96 amostras para treinamento e 96 amostras de testes de forma aleatória. Para realizar essa divisão, utilizou-se a função *train_test_split*, ajustando-a para realizar *shuffle* nas amostras e utilizando a semente aleatória 42, para facilitar a repetibilidade dos testes. A Figura 18 ilustra, para as imagens de face, o processo de divisão do conjunto de dados. Os mesmos procedimentos foram realizados nas amostras de voz.

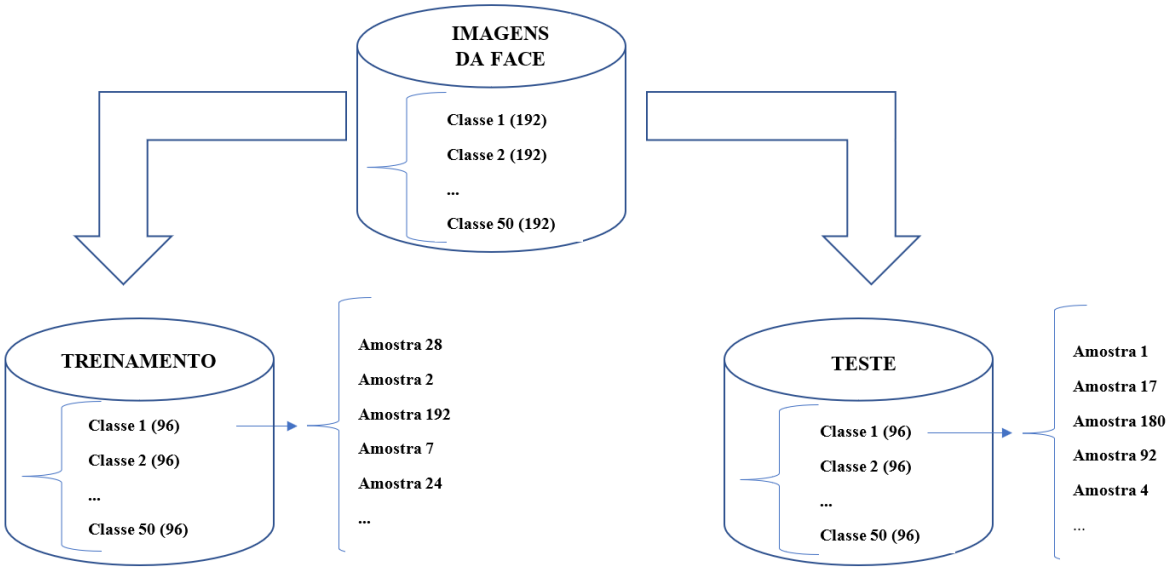


Figura 18: Ilustração da divisão do conjunto de dados das imagens de Face

4.2.2 Abordagens

Com respeito à primeira abordagem, as arquiteturas desenvolvidas foram baseadas nos trabalhos de Olazabal et al. (2019a) e Zhang et al. (2020), analisados anteriormente na revisão da literatura. O primeiro trabalho teve como objetivo propor uma abordagem de fortalecimento da segurança de sistemas biométricos de IoT, utilizando reconhecimento baseado em fusão de características com múltiplas modalidades biométricas. Seu sistema é composto por 3 blocos principais. O canal de face, responsável por extrair e obter características da face, gerou dois vetores de características 1x1296HOG e 1x1000LBP para cada amostra de face. O canal da voz, que utiliza redução de ruído, detector de atividades e extração de características através de MFCC, obteve um vetor de características 1x13500, contendo 15 coeficientes Cepstrais e os coeficientes diferenciais delta e delta-delta MFCCs. A etapa de fusão utiliza a técnica DCA para realizar a fusão a nível de característica. Destaca-se que os autores optaram por fundir os

3 vetores de características em pares, totalizando $\binom{3}{2}$ fusões. A classificação foi realizada pelo método KNN, que recebeu como entrada o vetor fundido, respondendo de forma binária, identificando se amostra era genuína (do indivíduo) ou não. Zhang et al. (2020) utilizou arquitetura para o sistema de verificação similar ao primeiro trabalho, porém optou por utilizar a fusão em nível de pontuação, utilizando um *min-max* para normalizar os dados de face e voz, devido a diferenças biométricas. Na estratégia de fusão utilizou como regra, a soma ponderada e o produto das pontuações da face e voz normalizadas. Com objetivo de evitar que um sinal unimodal prevalecesse sobre o outro, uma ponderação por um fator a, calculado a partir da relação SNR, foi realizada.

Apesar de termos nos baseado nos trabalhos de Olazabal et al. (2019a) e Zhang et al. (2020), para implementar a arquitetura sem abordagem de aprendizado profundo, para fins de comparação com a abordagem utilizando aprendizado profundo, salienta-se que a proposta deste trabalho de dissertação é de realizar a tarefa de identificação dos indivíduos (dentro de um conjunto N indivíduos), através de modos de face e voz é diferente dos trabalhos mencionados, uma vez que os referidos autores optaram apenas por realizar uma classificação binária, entre genuíno e impostor.

Adicionalmente, em cada abordagem serão implementados: uma arquitetura unimodal de face, uma de voz e o sistema multimodal face-voz.

4.2.3 Abordagem sem aprendizado profundo

No que se refere a abordagens não profundas, conforme apresenta-se na Figura 19, as 3 primeiras arquiteturas possuem estruturas similares, compondo etapas de entrada de dados, extração de características, máquinas de quantização vetorial, normalização e identificação. A primeira arquitetura (Figura 19a) é o sistema unimodal da face, que tem como entrada a face já pré-processada, a qual, antes de ser apresentada para o *autoencoder*, foi linearizada para um vetor de dimensão 5120. O *autoencoder* contém apenas 3 camadas: a camada de entrada de dimensão 5120, a camada de codificada e a camada de saída, com a mesma dimensão da entrada. As dimensões da codificação v_{face} são 256, 512 e 1024, representando uma compressão na ordem de 20, 10 e 5 vezes, respectivamente. Após isso, o vetor v_{face} é inserido na máquina LVQ, possuindo três possibilidades de quantidades de *codebooks* para treinamento (16, 32 e 48). Ao apresentar uma imagem de teste na entrada, a máquina LVQ computa a distância euclidiana entre a amostra e os centros de *clusters* que representam um indivíduo.

Devido às diferenças de sinais de entrada, utilizou-se um bloco de normalização vetorial para garantir uma escala comum entre os elementos do vetor de saída que é representado por v_{fn}. A amostra de face é identificada verificando o *codebooks* que possuir a menor distância euclidiana.

A arquitetura unimodal de voz (Figura 19b), a segunda arquitetura, possui um bloco de extração de coeficientes Mels. Foram extraídos 13 MFCCs e, a partir disso, calculados 13 delta MFCCs e 13 delta-delta MFCCs. Para esta etapa, utilizou-se a função *mfcc* pertencente a biblioteca *python_speech_features*. Parametrizou-se da seguinte maneira: frequência de amostragem em 16kHz, FFT com 512 pontos, frequência da borda inferior e superior da banda dos filtros mel, respectivamente, 0Hz e 8kHz, filtro pré-ênfase 0,97, janela aplicada em cada frame *hamming*, duração da janela 32ms, espaçamento entre as amostras 8ms, quantidade de filtros no banco de filtros mel 26 e quantidade de coeficientes Cepstrais preservados 13. Para obtenção dos coeficientes dinâmicos, configurou-se a função *delta*. De forma análoga ao canal de face, os vetores gerados v_{MFCC} e v_{d-MFCC} são inseridos nas máquinas LVQs, com 3 possibilidades de *codebooks* 16, 32 e 48 para treinamento. Ao inserir uma amostra de voz de teste na entrada, as máquinas LVQs computam as distâncias euclidianas entre a amostra e os centros de *clusters* que representam um indivíduo. Também foram inseridos dois blocos de normalização vetorial, localizados logo após as máquinas LVQs, e suas saídas foram representadas por v_{Mn} e v_{dMn}. O canal de voz já possui uma etapa de fusão para combinar a contribuição individual de cada distância euclidiana dos dois canais da voz. A identificação é realizada obtendo o sujeito que tiver a menor pontuação. O mesmo processo de fusão é realizado para o sistema multimodal (Figura 19c), porém, acrescenta-se o canal de face. As pontuações dos três vetores v_{fn}, v_{Mn} e v_{dMn} são somadas e a decisão de escolher a qual indivíduo a amostra face-voz pertence, é feita através da verificação de quem possui a menor pontuação.

4.2.4 Abordagem com aprendizado profundo

As arquiteturas profundas contêm entrada de dados, extração de características, treinamento da CNN e classificação das amostras. A quarta arquitetura (Figura 20a) foi elaborada para identificação de indivíduos utilizando traços da face através de uma rede CNN. De forma similar ao que foi realizado na abordagem não profunda, utilizou-se o *autoencoder* para obter características fundamentais da face e realizar uma representação reduzida dela. O *autoencoder* possibilita uma filtragem eficiente de ruído no sinal da face, descartando objetos

indesejados e reduzindo o impacto da baixa iluminação. O vetor v_{face} composto pela camada de codificação foi inserido na entrada da rede CNN. A arquitetura do modelo de face foi composta por quatro sequências de camada convolutiva 3x3, camada *batch normalization* e ReLU, seguida de uma camada de subamostragem 2x2 (*maxpooling*) e uma camada de *dropout*. A camada de GAP foi utilizada para diminuir o mapa de características e sua saída foi inserida na etapa de classificação. A etapa de classificação, consistiu em uma rede totalmente conectada composta pela sequência de camadas: *dense*, *reLU*, *dropout*, *dense* e *Softmax*. A primeira camada *dense* foi ajustada para 128 neurônios e a segunda camada para 50 neurônios com função de ativação *Softmax* responsável pela classificação final das amostras de face.

A quinta arquitetura (Figura 20b) foi idealizada para identificação de indivíduos utilizando os traços da voz através de uma rede CNN. O sinal de voz é levado ao extrator de MFCCs, em que são extraídos 13 coeficientes mel e calculados os coeficientes mel 13 delta e 13 delta-delta. As características da voz extraídas são inseridas na rede CNN. Sua arquitetura é composta por uma camada convolutiva 3x3, uma camada *batch normalization* e uma camada ReLU. No segundo bloco realiza-se a etapa de subamostragem (*maxpooling*), com a sequência de camadas: convolutiva 3x3, *batch normalization*, *ReLU*, *maxpooling 2x2* e *dropout*. No terceiro bloco, realiza-se a última sequência de convolutiva 3x3, *batch normalization* e *ReLU*. Similarmente a quarta arquitetura, a camada GAP foi inserida para diminuição do mapa de característica e sua saída foi colocada na etapa de classificação. A camada de classificação foi projetada com uma camada *dense* com 50 neurônios seguida de uma *Softmax* para classificação final das amostras de voz.

A sexta e última arquitetura (Figura 21) é o sistema multimodal de identificação de indivíduos utilizando os modos de face e voz através de uma rede CNN. A ideia chave dessa arquitetura é utilizar as camadas unimodais pré-treinadas. Assim, a transferência de aprendizado é realizada removendo as camadas de classificação de cada sistema unimodal e congelando os pesos e polarizações das camadas treinadas na etapa unimodal. A fusão é realizada a nível de característica através de uma camada *concatenate* que concatena cada camada GAP dos modelos unimodais. Essa camada recebe como entrada uma lista de 2 tensores com 64 valores cada (saída das redes CNN da face e da voz) e retorna um único tensor com 128 valores, representando a concatenação dos mesmos. A classificação final é realizada através da rede totalmente conectada: *dense* (128), *dense* (64), *dropout*, *dense* (50) e *Softmax*. No sistema multimodal apenas os parâmetros da rede final de classificação são treinados.

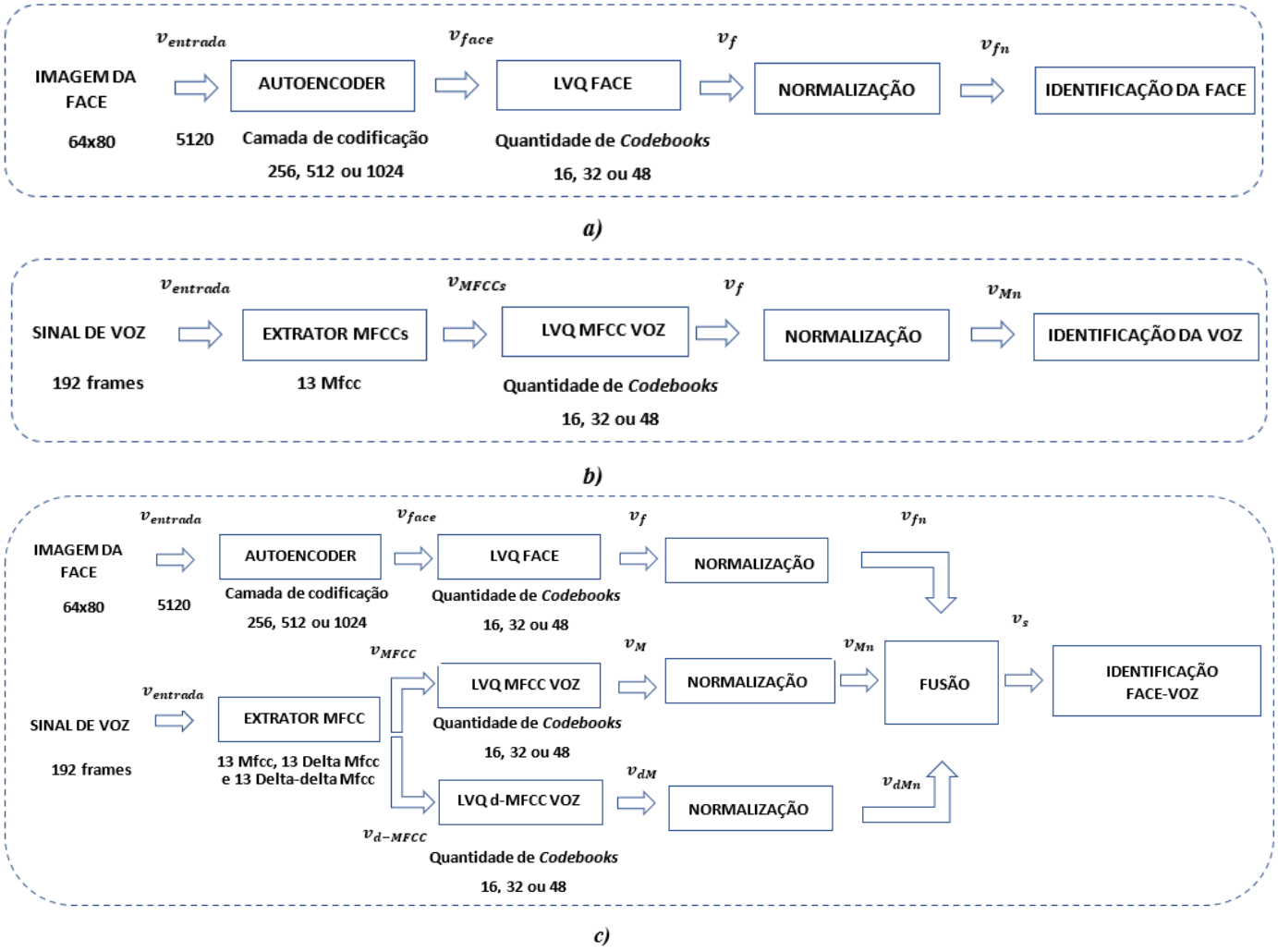


Figura 19: Abordagem sem aprendizado profundo.(a) Arquitetura 1 – Unimodal da Face com LVQ. (b) Arquitetura 2 – Unimodal da Voz com LVQ. (c) Arquitetura 3 – Multimodal Face-Voz com LVQ

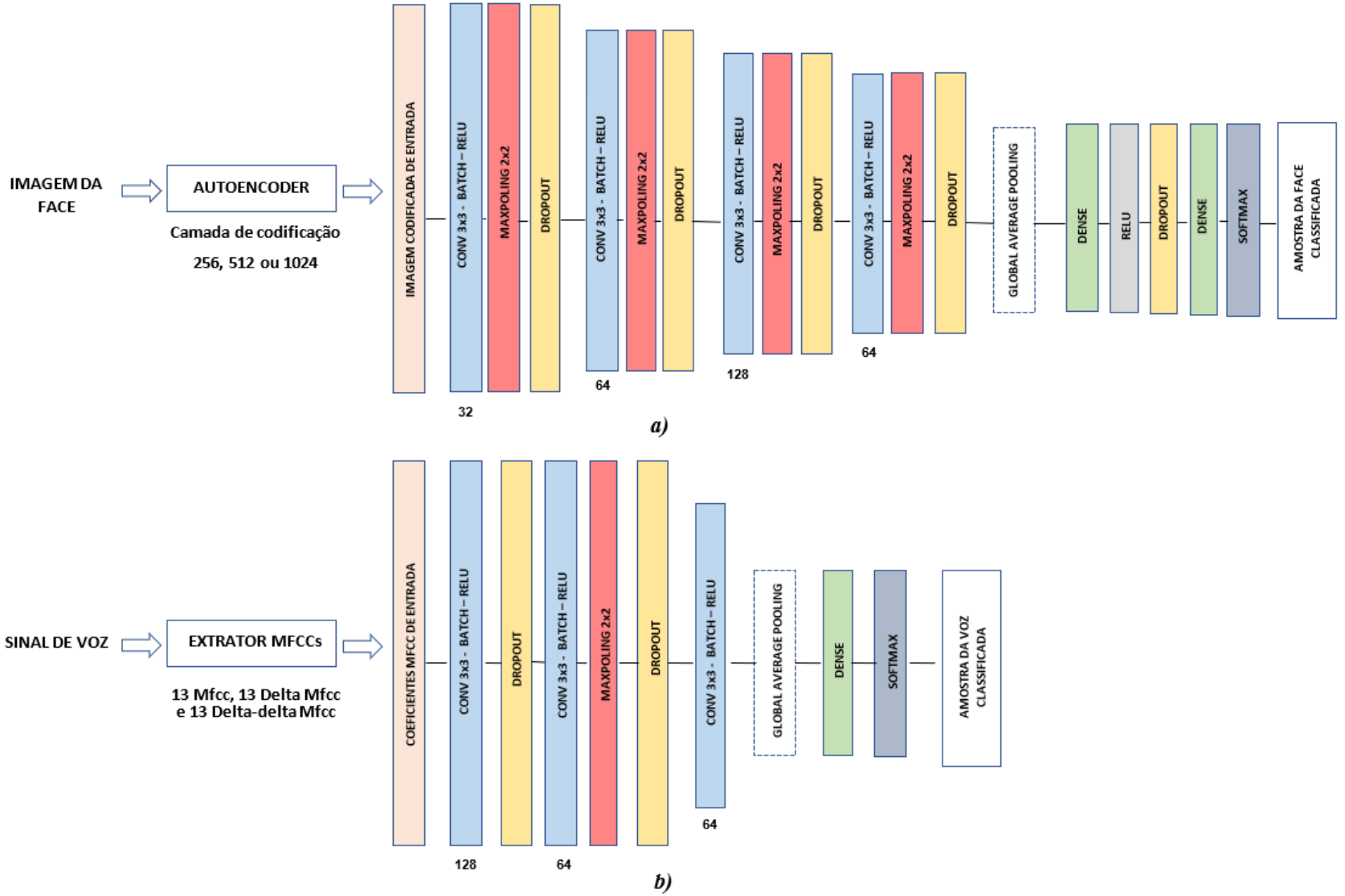


Figura 20: Abordagens que utiliza o aprendizado profundo: (a) Arquitetura 4 – Unimodal da Face com CNN. (b) Arquitetura 5 – Unimodal da Voz com CNN

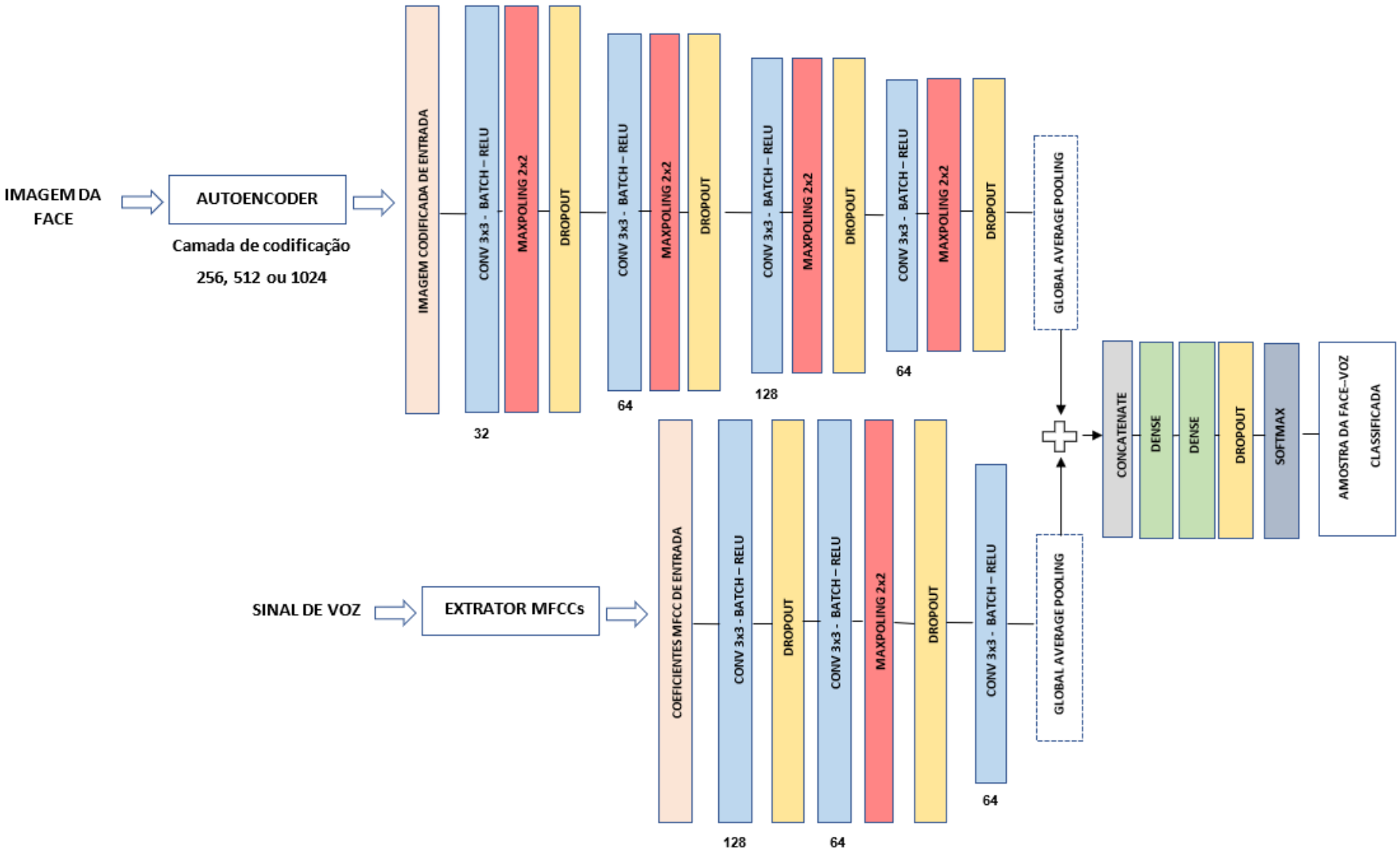


Figura 21: Abordagem que utiliza o aprendizado profundo (continuação): Arquitetura 6 – Multimodal Face-Voz com CNN.

4.2.5 Parâmetros do treinamento

Os treinamentos das arquiteturas propostas neste trabalho utilizaram parâmetros configurados experimentalmente durante as simulações. A Tabela 1 detalha os dados de entrada da face e voz, descrendo a composição dos vetores de entrada que foram utilizados inicialmente nas redes. Na Tabela 2, são exibidos os valores para o treinamento das arquiteturas 1, 2 e 3 que utilizam abordagem não profunda através da rede LVQ.

Tabela 1: Dados de entrada de face e voz

Parâmetros	Valor
Quantidade total de indivíduos	50
Quantidade total de imagens por indivíduo	192
Saída do *autoencoder* (SA)	$256, 512\ e\ 1024$
Dimensão do vetor da face	$(50, 192, SA)$
Quantidade total de conjunto de áudios por indivíduo	192
Quantidade total de frames de áudios por indivíduo	192
Quantidade total de coeficientes mel	39
Dimensão do vetor da voz	$(50, 192, 192, 39)$

Tabela 2: Parâmetros do treinamento das máquinas LVQs.

Parâmetros	Valor
Taxa de aprendizado inicial	0,3
Fator de queda de taxa de aprendizagem	Diminuição linearmente com o número de épocas
Número de épocas	1000
Clusters	$16, 32$ e 48
Saída do *autoencoder*	$256, 512$ e 1024
Quantidade de conjuntos de áudio por indivíduo	24 primeiros
Quantidade de *frames* por indivíduo	$32, 48$ e 64

Na Tabela 3, são mostrados os valores para ao treinamento das arquiteturas 4, 5 e 6 que utilizaram abordagem de aprendizado profundo através da rede CNN.

Tabela 3: Parâmetros do treinamento das redes CNNs.

Parâmetros	Valor
Taxa de aprendizado inicial	0,01
Fator de queda de taxa de aprendizagem	0,0001
Número de épocas	100
Tamanho do lote	64
Otimizadores	Adam, SGD e RMSProp
Função de perda	*Categorical Cross Entropy*
Saída do *autoencoder*	256, 512 e 1024
Quantidade de *frames* por indivíduo	32, 48, 64 e 192

Em simulações que utilizaram regularização, o fator da camada de *dropout* foi ajustado para $p = 0{,}1$.

4.2.6 Métricas de avaliação

A avaliação qualitativa das arquiteturas foi realizada através das métricas de desempenho: Curva ROC (*Receiver Operating Characteristic*) e AUC-ROC (*Area Under the ROC Curve*). Consideram-se as seguintes definições:

a) Verdadeiro positivo (VP): resultado positivo da identificação de um indivíduo que é o indivíduo procurado.
b) Verdadeiro negativo (VN): resultado negativo da identificação de um indivíduo que não é o indivíduo procurado.
c) Falso positivo (FP): resultado positivo da identificação de um indivíduo que não é o indivíduo procurado.
d) Falso negativo (FN): resultado negativo da identificação de um indivíduo que é o indivíduo procurado.

Antes de definir a curva ROC e AUC-ROC, é necessário definir as métricas $Sensibilidade$ e $Especificidade$. A $Sensibilidade$ (TAR, do inglês *True Acceptance Rate*) avalia a porcentagem de indivíduos que foram classificados corretamente e que pertencem a classe positiva. Em contrapartida, a $Especificidade$ calcula a porcentagem de indivíduos que

foram classificados corretamente da classe negativa. A FAR (do inglês *False Acceptance Rate*) é definida pela relação $FAR = 1 - Especificidade$ (GÉRON, 2019).

$$Sensibilidade\ (TAR) = \frac{VP}{VP + FN} \tag{32}$$

$$Especificidade = \frac{VN}{VN + FP} \tag{33}$$

$$FAR = 1 - Especificidade = \frac{FP}{FP + VN} \tag{34}$$

A curva ROC resume o desempenho do modelo para todos os limiares possíveis. A curva ROC é plotada através dos parâmetros TAR e FAR sobre diferentes limiares, normalmente variando entre 0 e 1. O limiar ótimo é definido através da menor distância em relação ao ponto (0, 1) em que a AUC-ROC é considerada como modelo perfeito com área 1. Classificadores aleatórios normalmente possuem uma relação crescente e linear entre TAR e FAR e AUC-ROC próxima de 0,5 (DAS; CAKMAK, 2018; GÉRON, 2019). A Figura 22 exemplifica a curva ROC, a métrica AUC-ROC e o melhor ponto de operação.

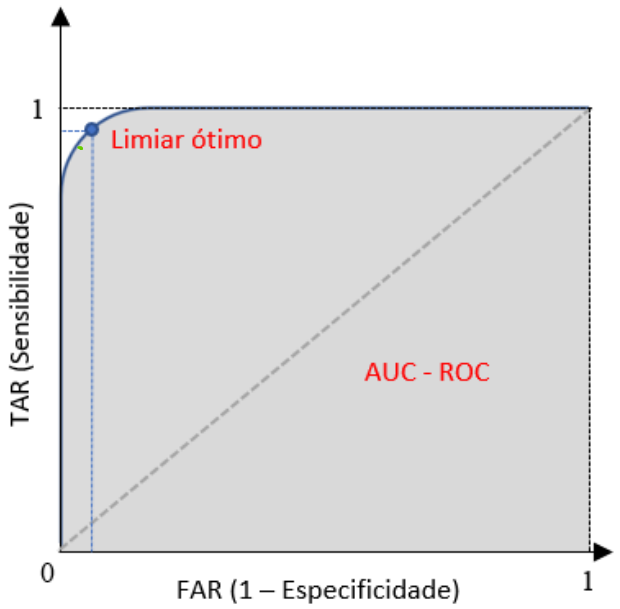


Figura 22: Ilustração de uma Curva ROC com indicação da área sob a curva ROC (AUC-ROC), em cinza, e do limiar ótimo.

5 RESULTADOS E DISCUSSÕES

Neste capítulo serão apresentados os resultados obtidos, quais sejam, os resultados das simulações das três primeiras arquiteturas que utilizam uma abordagem não profunda através das redes LVQ, totalizando 27 simulações. Também serão apresentados os resultados das simulações das três últimas arquiteturas que utilizam uma abordagem de aprendizado profundo, através das redes CNN. Um total de 258 simulações foram realizadas.

5.1 IMPLEMENTAÇÃO UTILIZANDO ABORDAGEM NÃO PROFUNDA – REDES LVQ

O desempenho da abordagem unimodal de identificação biométrica utilizando traços de voz primeiramente é apresentado, conforme mostrado na Tabela 4. Em seguida, o desempenho do modelo unimodal utilizando imagens da face e do sistema bimodal face-voz são apresentados, conforme mostrado nas Tabelas 5 e 6, respectivamente. Esses três métodos correspondem às arquiteturas 1, 2 e 3 apresentadas no capítulo anterior. Os referidos desempenhos são valorados através das métricas: Taxa de Verdadeiros Positivos (TAR, do inglês *True Acceptance Rate*) e Taxa de Falsos Positivos (FAR, do inglês *False Acceptance Rate*). Utilizando essas métricas, determina-se o ponto ótimo de operação da curva ROC e o limiar correspondente utilizado para obtê-lo. O limiar ótimo é obtido considerando o ponto da curva ROC mais próximo ao ponto (0,1). Também são obtidas as áreas sob a Curva ROC (AUC-ROC, do inglês *Area Under the Curve - Receiver Operating Characteristics*) para todas as três arquiteturas, em função do número de *clusters*, do tamanho dos *frames* utilizados para teste e do tamanho da camada de codificação. As métricas foram avaliadas utilizando-se o conjunto de teste. Avaliou-se também o tempo de treinamento e teste dos métodos LVQs em segundos.

Conforme visto na Tabela 4, para a abordagem não profunda do modelo unimodal de identificação da voz, a quantidade de *clusters* da máquina LVQ assumiu os valores de 16, 32 e 48, enquanto que o número de *frames* assumiu os valores de 32, 48 e 64.

Tabela 4: Resultados da abordagem não profunda do modelo unimodal de identificação biométrica utilizando o sinal de voz, no conjunto de teste

Número de *clusters*	*frames*/ indivíduo	Ponto ótimo de operação			AUC-ROC	Tempo de treinamento (s)	Tempo de teste (s)
		Limiar	TAR (%)	FAR (%)			
16	32	0,0193	77,8	19,1	0,869	1257	1,5
16	48	0,0194	75	23,3	0,833	1421	1,7

16	64	0,0193	78,4	18,8	0,871	1562	2,0
32	32	0,0193	80,2	18,9	0,872	1557	1,4
32	48	0,0193	72,9	20	0,836	1725	2,1
32	64	0,0193	80,1	18,8	0,875	1831	2,1
48	32	0,0193	78,3	18,3	0,874	1635	1,6
48	48	0,0194	77,3	22,1	0,845	1723	2,5
48	**64**	**0,0193**	**79,5**	**17,9**	**0,877**	**1822**	**2,7**

Conforme mostrado na Tabela 5, na abordagem não profunda do modelo unimodal de identificação biométrica utilizando os dados da face, o número de *clusters* assumiu os valores de 16, 32 e 48. Por outro lado, a camada codificada do *autoencoder* assumiu as seguintes dimensões: 256, 512 e 1024.

Tabela 5: Resultados da abordagem não profunda do modelo unimodal de identificação biométrica utilizando o sinal de face, no conjunto de teste

Número de *clusters*	Dimensão da camada de codificação	Melhor ponto de operação			AUC-ROC	Tempo de treinamento (s)	Tempo de teste (s)
		Limiar	TAR (%)	FAR (%)			
16	1024	0,0085	86,3	13,6	0,94	425	1,10
16	512	0,0084	87,6	13,1	0,94	399	0,87
16	256	0,0082	86	12,3	0,94	256	0,58
32	1024	0,0085	87	13,5	0,94	526	1,42
32	512	0,0084	87,5	13	0,94	429	0,92
32	256	0,0082	86,2	12,2	0,94	311	0,75
48	1024	0,0085	86,9	13,3	0,94	724	1,78
48	512	0,0083	86,3	11,5	0,94	628	0,96
48	256	0,0082	86,1	12,2	0,94	431	0,85

Na tabela 6 mostra-se os resultados obtidos com a fusão das informações de voz e face, utilizando o método LVQ. Observa-se que o número de *clusters* assumiu os valores de 16, 32 e 48, enquanto que as dimensões da camada codificada do *autoencoder* assumiu os valores de 256, 512 e 1024. Optou-se por fixar a quantidade de *frames* de áudio em 64, devido a este parâmetro ter atingido o melhor resultado em relação a área sob a curva ROC no sistema unimodal de voz.

Tabela 6: Resultados da abordagem não profunda do modelo multimodal de identificação biométrica (fusão da voz e face) com 64 *frames* por indivíduo, no conjunto de teste

Número de clusters	Dimensão da camada de codificação	Melhor ponto de operação			AUC-ROC	Tempo de treinamento (s)	Tempo de teste (s)
		Limiar	TAR (%)	FAR (%)			
16	1024	0,0275	90,5	6,4	0,977	1123	2,57
16	512	0,0272	90,2	5,1	0,979	754	1,84
16	256	0,0271	90,4	5,6	0,98	628	0,61
32	1024	0,0276	91,7	7,3	0,978	1200	2,95

32	512	0,0274	92,1	6,4	0,979	789	2,54
32	256	0,0274	92,9	7,8	0,98	689	0,89
48	1024	0,0277	92,3	8,1	0,978	1329	3,14
48	512	0,0274	91,9	6,4	0,98	852	2,73
48	**256**	**0,0274**	**92,6**	**7,7**	**0,98**	**724**	**1,07**

A Figura 23 mostra as curvas ROC da abordagem não profunda do modelo unimodal de identificação biométrica de traços de voz, com o número de *clusters* igual a 48 e diferentes valores de *frames* por indivíduo.

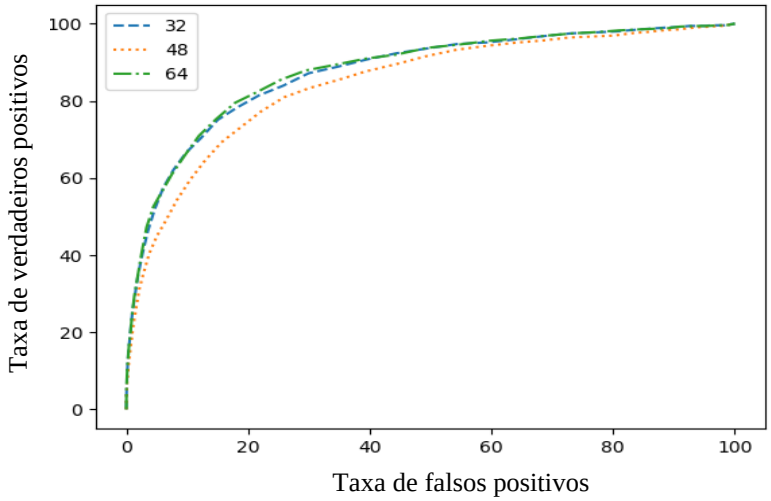


Figura 23: Curva ROC da abordagem não profunda do modelo unimodal de identificação biométrica da voz com 48 *clusters* e diferentes valores de números de *frames* de teste por indivíduo (32,48 e 64) no conjunto de teste

A Figura 24 exibe a curva ROC da abordagem não profunda do modelo unimodal de identificação biométrica através da face, com o tamanho da camada codificada do *autoencoder* igual a 1024 e diferente número de *clusters*.

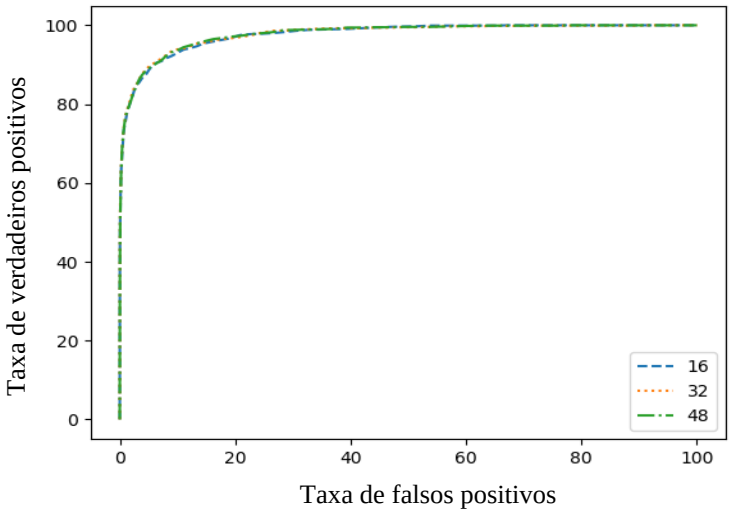


Figura 24: Curva ROC da abordagem não profunda do modelo unimodal de identificação biométrica através da face com dimensão da camada de codificação do *autoencoder* igual a 1024 e diferentes números de *clusters* (16, 32 e 48) no conjunto de teste.

A Figura 25 apresenta a curva ROC da abordagem não profunda do modelo de identificação biométrica bimodal (voz e face), com o tamanho da saída da camada codificada do *autoencoder* igual a 1024 e diferentes quantidades de *clusters* (16, 32, e 48).

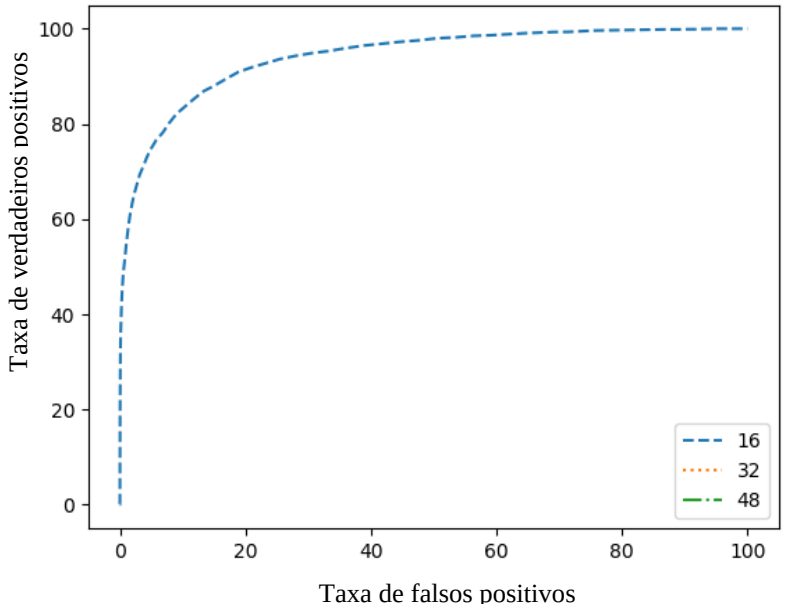


Figura 25: Curva ROC da abordagem não profunda do modelo de identificação biométrica bimodal (voz e face) com dimensão de saída do *autoencoder* igual a 1024 e diferentes quantidades de *clusters* (16, 32 e 48) no conjunto de teste.

5.2 IMPLEMENTAÇÃO UTILIZANDO ABORDAGEM DE APRENDIZADO PROFUNDO – REDES CNN

Nas Tabelas 7 e 8 são apresentados o desempenho do modelo unimodal de identificação biométrica com abordagem de aprendizado profundo utilizando imagens de face. Em seguida, nas Tabelas 9 e 10, são apresentados o desempenho do modelo unimodal, com abordagem de aprendizagem profunda utilizando traços da voz. O desempenho do modelo de identificação com aprendizado profundo bimodal face-voz é apresentado nas Tabelas 11 e 12. Esses três métodos correspondem às arquiteturas 4, 5 e 6 apresentadas no capítulo anterior. Os referidos desempenhos são valorados através das métricas: TAR e FAR. Utilizando essas métricas, são obtidos o ponto ótimo de operação da curva ROC e o limiar correspondente utilizado para obtê-lo. O melhor valor de limiar é obtido considerando o ponto da curva ROC mais próximo ao ponto (0,1). Também são obtidas as áreas sob a Curva ROC (AUC-ROC) para todas as três arquiteturas, em função do método de regularização, do método de otimização, do tamanho dos *frames* utilizados, número de coeficientes MFCCs e do tamanho da camada de codificação. Avaliou-se também o tempo de treinamento e teste das redes CNNs.

O desempenho foi avaliado combinando-se a presença da regularização (sim ou não), o método de otimização (ADAM, RMSprop e SGD), o número de *frames* (32, 48, 64 e 192), o número de coeficientes MFCCs (13, 26 e 39) e o tamanho da camada de codificação (1024, 512, e 256). Para as três arquiteturas, essas combinações resultaram em um total de 258 experimentos. Em virtude da enorme quantidade de simulações realizadas, optou-se por apresentar, para cada cenário, os dados de duas formas: a primeira, fixando-se o método de otimização como sendo o método ADAM; a segunda, fixando-se a dimensão da camada de codificação no maior valor, 1024 neurônios. Em ambos os casos se utilizou a maior quantidade de coeficientes MFCCs e o maior número de *frames* de teste, 39 e 192 respectivamente. Entretanto, no Apêndice deste documento pode-se encontrar os resultados dos 258 experimentos. As escolhas anteriormente citadas foram as que resultaram nos melhores desempenhos.

A Tabela 7 exibe o desempenho do modelo unimodal de identificação biométrica para o canal da face utilizando as redes CNN no cenário em que foi fixado o otimizador como o ADAM. Os experimentos foram realizados com e sem regularização e em 3 diferentes dimensões de camada de codificação.

Tabela 7: Resultado da abordagem de aprendizado profundo do modelo unimodal de identificação biométrica utilizando o sinal de face para o otimizador ADAM, no conjunto de teste

Regularização	Dimensão da camada de codificação	Melhor ponto de operação			AUC-ROC	Tempo de treinamento(s)	Tempo de teste (s)
		Limiar	TAR (%)	FAR (%)			
Não	1024	0,002	98,60	0,52	0,9791	132	0,4113
Não	512	0,0001	98,67	0,86	0,9788	71	0,3343
Não	256	0,0003	99,04	0,6	0,9793	70	0,3984
Sim	**1024**	**0,0008**	**99,23**	**0,85**	**0,9898**	**130**	**0,7343**
Sim	512	0,0005	99,08	0,76	0,9892	83	0,4009
Sim	256	0,0011	99,00	0,64	0,9891	78	0,4311

A Tabela 8, apresenta-se resultados com e sem a presença da regularização, variou-se os métodos de otimização e foi fixado a dimensão na camada de codificação em 1024 neurônios. As escolhas feitas para obtenção dos resultados nas Tabelas 7 e 8 foram as que resultaram nos melhores desempenhos.

Tabela 8: Resultado da abordagem de aprendizado profundo do modelo unimodal de identificação biométrica utilizando o sinal de face com camada de codificação do *autoencoder* 1024 neurônios, no conjunto de teste

Regularização	Método de Otimização	Melhor ponto de operação			AUC-ROC	Tempo de treinamento(s)	Tempo de teste (s)
		Limiar	TAR (%)	FAR (%)			
Não	ADAM	0,002	98,60	0,52	0,9791	132	0,4113
Não	SGD	0,0096	98,44	1,08	0,9682	129	0,4148
Não	RMSprop	0,00001	99,42	0,96	0,9793	134	0,4162
Sim	ADAM	0,0008	99,23	0,85	0,9898	130	0,7343
Sim	SGD	0,0259	98,56	1,33	0,9786	120	0,4159
Sim	RMSprop	0,0015	98,98	0,45	0,9895	130	0,4042

A Tabela 9 mostra o desempenho do modelo unimodal de identificação biométrica para o canal da voz utilizando as redes CNN para amostras do conjunto de teste. Nessa tabela são mostrados os resultados para o otimizador ADAM. Os seguintes parâmetros foram variados: com e sem regularização, três quantidades de coeficientes MFCCs e quatro quantidade distintas de *frames* de testes por indivíduo.

Tabela 9: Resultado da abordagem de aprendizado profundo do modelo unimodal de identificação biométrica utilizando o sinal de voz para o otimizador ADAM, no conjunto de teste

Regularização	Número de coeficientes	Número de *frames* de testes por indivíduo	Melhor ponto de operação			AUC-ROC	Tempo de treinamento (s)	Tempo de teste (s)
			Limiar	TAR (%)	FAR (%)			
Não	13	32	0,0093	78,79	33,86	0,8024	39	0,3135
Não	13	48	0,013	83,65	30,26	0,8455	46	0,3143
Não	13	64	0,0132	82,92	24,28	0,8926	53	0,3334
Não	26	32	0,0149	82,42	37,98	0,7994	45	0,3339
Não	26	48	0,0171	82,92	31,43	0,8383	62	0,3885
Não	26	64	0,012	83,00	27,48	0,8805	81	0,437
Não	39	32	0,0178	79,90	33,69	0,8068	57	0,3966
Não	39	48	0,0092	81,19	34,56	0,8287	95	0,4579
Não	39	64	0,0104	86,10	29,07	0,8781	112	0,5484
Não	39	192	0,0123	93,10	8,05	0,9833	443	1,19
Sim	13	32	0,0171	87,58	36,30	0,8202	140	0,3115
Sim	13	48	0,0192	86,94	30,06	0,8524	173	0,3078
Sim	13	64	0,019	89,06	24,05	0,9014	203	0,3309
Sim	26	32	0,0221	85,94	34,60	0,8189	209	0,3307
Sim	26	48	0,0216	84,38	27,72	0,8525	280	0,3934
Sim	26	64	0,0231	87,60	23,39	0,9013	357	0,4468
Sim	39	32	0,0185	85,62	34,64	0,8162	323	0,3851

Sim	39	48	0,0215	87,69	30,30	0,8531	415	0,462
Sim	39	64	0,0209	89,44	25,86	0,8969	563	0,5475
Sim	**39**	**192**	**0,0286**	**96,40**	**6,10**	**0,9803**	**756**	**1,2748**

Na Tabela 10 escolheu-se um número de coeficientes MFCC igual a 39 e o número de *frames* igual a 192, variando-se a presença ou não da regularização e os otimizadores. As escolhas feitas para obtenção dos resultados nas Tabelas 9 e 10 foram as que resultaram nos melhores desempenhos.

Tabela 10: Resultado da abordagem de aprendizado profundo do modelo unimodal de identificação biométrica utilizando o sinal de voz com 39 coeficientes e 192 número de *frames*, no conjunto de teste

		Melhor ponto de operação					
Regularização	Otimizador	Limiar	TAR (%)	FAR (%)	AUC-ROC	Tempo de treinamento (s)	Tempo de teste (s)
Não	ADAM	0,0123	93,10	8,05	0,9833	443	1,19
Não	SGD	0,0295	83,56	19,00	0,9092	862	2,1876
Não	RMSprop	0,0112	93,10	8,57	0,9851	435	1,2081
Sim	ADAM	0,0286	96,40	6,10	0,9803	756	1,2748
Sim	SGD	0,0269	78,67	23,37	0,8696	502	1,2148
Sim	RMSprop	0,0329	95,46	6,09	0,9886	470	1,2236

Por fim, seguindo o mesmo procedimento citado anteriormente, na Tabela 11 exibe-se o desempenho do modelo multimodal de identificação biométrica para os canais face-voz utilizando as redes CNN para amostras do conjunto de teste. Para obtenção dos resultados mostrados escolheu-se o otimizador ADAM e variou-se os seguintes parâmetros: com e sem regularização, três dimensões da camada de codificação do *autoencoder*, três quantidades de coeficientes MFCCs e quatro quantidade distintas de *frames* de testes por indivíduo. Na Tabela 12 escolheu-se um número de coeficientes MFCC igual a 39 e o número de frames igual a 192, variando-se a presença ou não da regularização e os otimizadores. As escolhas feitas para obtenção dos resultados nas Tabelas 11 e 12 foram as que resultaram nos melhores desempenhos.

Tabela 11: Resultado da abordagem de aprendizado profundo do modelo multimodal de identificação biométrica (voz e face) para o otimizador ADAM, no conjunto de teste

			Número de *frames* de testes por indivíduo	Melhor ponto de operação					
Regularização	Dimensão da camada de codificação	Número de coeficientes		Limiar	TAR (%)	FAR (%)	AUC-ROC	Tempo de treinamento (s)	Tempo de teste (s)
Não	1024	13	32	0,0075	98,77	0,79	0,9988	100	0,6524

Não	1024	13	48	0,0054	98,88	0,79	0,999	90	0,6275
Não	1024	13	64	0,0162	98,75	0,47	0,9993	143	0,6129
Não	1024	26	32	0,0091	98,58	0,59	0,999	94	0,6279
Não	1024	26	48	0,0081	98,71	0,73	0,999	103	0,8604
Não	1024	26	64	0,0058	98,92	0,77	0,9991	143	0,6736
Não	1024	39	32	0,0035	98,77	0,95	0,9987	101	0,6335
Não	1024	39	48	0,0094	98,79	0,68	0,9989	118	0,72
Não	1024	39	64	0,0069	98,96	0,75	0,9992	143	0,7827
Não	1024	39	192	0,01	99,19	0,46	0,9995	263	1,4324
Não	512	13	32	0,0078	98,75	0,81	0,9987	71	0,528
Não	512	13	48	0,0088	98,62	0,81	0,9989	80	0,537
Não	512	13	64	0,0127	98,62	0,59	0,999	89	0,5492
Não	512	26	32	0,0114	98,42	0,68	0,9981	89	0,549
Não	512	26	48	0,0124	98,60	0,61	0,9987	143	0,6406
Não	512	26	64	0,007	98,98	0,88	0,999	131	0,764
Não	512	39	32	0,0099	98,60	0,67	0,9986	110	0,638
Não	512	39	48	0,0068	98,92	0,88	0,9989	143	0,8033
Não	512	39	64	0,0068	98,92	0,87	0,9989	175	0,9779
Não	512	39	192	0,0235	99,02	0,37	0,9993	443	2,3001
Não	256	13	32	0,0088	95,73	2,57	0,9946	71	0,4718
Não	256	13	48	0,007	96,15	2,51	0,9947	79	0,5123
Não	256	13	64	0,0019	97,54	2,62	0,997	89	0,5392
Não	256	26	32	0,0026	96,73	2,39	0,9957	88	0,5489
Não	256	26	48	0,0072	96,63	2,63	0,9948	108	0,6134
Não	256	26	64	0,0071	96,71	2,34	0,9958	128	0,7244
Não	256	39	32	0,003	96,65	2,64	0,9953	143	0,6305
Não	256	39	48	0,0029	96,77	2,84	0,9953	141	0,7848
Não	256	39	64	0,0083	96,42	1,87	0,9963	203	0,9496
Não	256	39	192	0,0058	98,88	1,09	0,999	443	2,2609
Sim	1024	13	32	0,0101	98,81	0,50	0,9994	79	0,5039
Sim	1024	13	48	0,0173	98,83	0,37	0,9996	90	0,5326
Sim	1024	13	64	0,0134	99,10	0,39	0,9995	98	0,5608
Sim	1024	26	32	0,0072	98,96	0,56	0,9995	102	0,5736
Sim	1024	26	48	0,0088	0,88	0,47	0,9995	123	0,6645
Sim	1024	26	64	0,0064	99,19	0,59	0,9996	203	0,7638
Sim	1024	39	32	0,0056	99,08	0,69	0,9994	143	0,6601
Sim	1024	39	48	0,0047	99,21	0,75	0,9994	169	0,8982
Sim	1024	39	64	0,0113	99,04	0,46	0,9996	143	0,6818
Sim	**1024**	**39**	**192**	**0,0155**	**99,44**	**0,27**	**0,9997**	**261**	**1,3661**
Sim	512	13	32	0,0086	98,90	0,55	0,9992	69	0,457
Sim	512	13	48	0,0076	99,15	0,58	0,9993	73	0,4657
Sim	512	13	64	0,0138	99,00	0,40	0,9994	79	0,4782
Sim	512	26	32	0,0061	99,17	0,66	0,9994	79	0,4676
Sim	512	26	48	0,0058	99,15	0,68	0,9994	101	0,5385
Sim	512	26	64	0,013	98,98	0,40	0,9993	143	0,5902
Sim	512	39	32	0,0072	99,23	0,64	0,9994	106	0,6605
Sim	512	39	48	0,0079	99,23	0,58	0,9992	124	0,7072

Sim	512	39	64	0,0096	98,98	0,50	0,9993	140	0,7999
Sim	512	39	192	0,0073	99,62	0,48	0,9996	324	1,5624
Sim	256	13	32	0,0097	98,73	0,39	0,999	83	0,4842
Sim	256	13	48	0,0039	99,08	0,62	0,999	82	0,5082
Sim	256	13	64	0,0057	99,06	0,53	0,9991	143	0,5475
Sim	256	26	32	0,005	99,06	0,54	0,9992	94	0,549
Sim	256	26	48	0,0071	99,12	0,45	0,9992	113	0,6419
Sim	256	26	64	0,0037	99,25	0,65	0,9989	135	0,7328
Sim	256	39	32	0,0036	99,00	0,65	0,9992	115	0,6415
Sim	256	39	48	0,0038	99,17	0,68	0,9991	203	0,7935
Sim	256	39	64	0,0047	99,15	0,58	0,9993	183	0,9511
Sim	256	39	192	0,0123	99,33	0,26	0,9995	454	0,6783

Tabela 12: Resultado da abordagem de aprendizado profundo do modelo multimodal de identificação biométrica (voz e face) com 39 coeficientes e 192 número de *frames*, no conjunto de teste

Regularização	Otimizador	Dimensão da camada de codificação	Melhor ponto de operação			AUC-ROC	Tempo de treinamento (s)	Tempo de teste (s)
			Limiar	TAR (%)	FAR (%)			
Não	ADAM	1024	0,01	99,19	0,46	0,9995	263	1,4324
Não	ADAM	512	0,0235	99,02	0,37	0,9993	443	2,3001
Não	ADAM	256	0,0058	98,88	1,09	0,999	443	2,2609
Não	SGD	1024	0,0193	98,21	1,27	0,9982	263	1,3399
Não	SGD	512	0,0125	98,27	1,56	0,9983	232	1,3245
Não	SGD	256	0,0165	98,27	1,20	0,9981	263	1,2991
Não	RMSprop	1024	0,0078	99,65	0,31	0,9999	237	1,3673
Não	RMSprop	512	0,0069	99,62	0,38	0,9997	263	1,3116
Não	RMSprop	256	0,0074	99,50	0,34	0,9997	241	1,3636
Sim	ADAM	1024	0,0155	99,44	0,27	0,9997	261	1,3661
Sim	ADAM	512	0,0073	99,62	0,48	0,9996	324	1,5624
Sim	ADAM	256	0,0123	99,33	0,26	0,9995	454	2,2635
Sim	SGD	1024	0,0254	98,81	1,12	0,9991	263	1,4024
Sim	SGD	512	0,0168	99,10	1,05	0,9991	266	1,3831
Sim	SGD	256	0,0309	98,79	0,60	0,9992	250	1,3302
Sim	RMSprop	1024	0,0329	99,06	0,37	0,9995	255	1,3455
Sim	RMSprop	512	0,0162	99,00	0,60	0,9994	253	1,3389
Sim	RMSprop	256	0,0127	99,12	0,54	0,9989	252	1,3089

A Figura 26 mostra a curva ROC obtida com a abordagem de aprendizado profundo do modelo unimodal de identificação biométrica de traços de face, utilizando regularização, o otimizador ADAM e diferentes dimensões da camada de codificação do *autoencoder* (1024, 512 e 256).

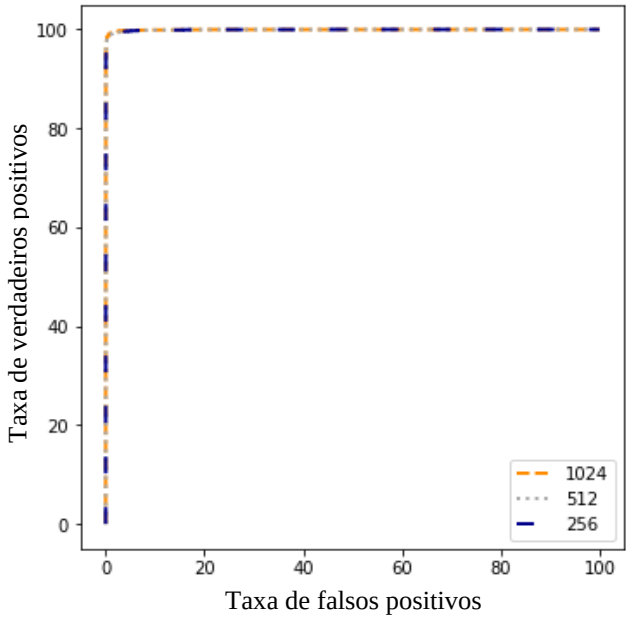


Figura 26: Curva ROC da abordagem de aprendizado profundo do modelo de identificação biométrica da face utilizando regularização e o otimizador ADAM e diferentes dimensões da camada de codificação do *autoencoder* (1024,512 e 256) no conjunto de teste

A Figura 27 mostra Curva ROC obtida com a abordagem de aprendizado profundo do modelo de identificação biométrica da voz, utilizando regularização, o otimizador ADAM, 39 coeficientes MFCCs e diferentes valores de frames (32,48, 64 e 192).

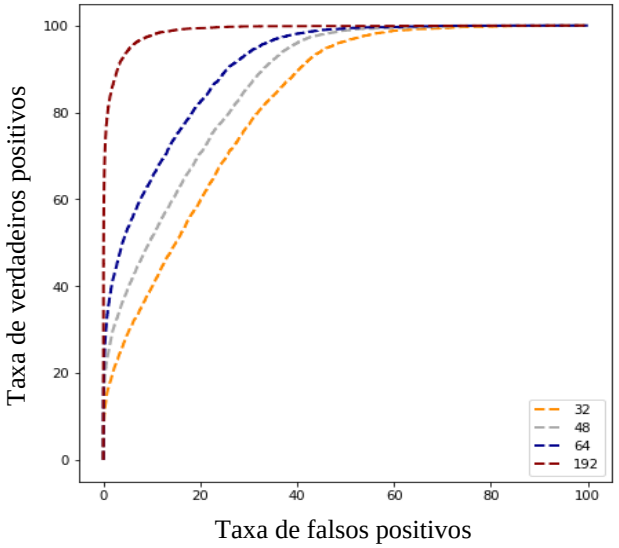


Figura 27: Curva ROC da abordagem de aprendizado profundo do modelo de identificação biométrica da voz utilizando regularização, otimizador ADAM, 39 coeficientes MFCCs e diferentes valores de frames (32,48, 64 e 192) no conjunto de teste

A Figura 28 mostra a Curva ROC obtida com a abordagem de aprendizado profundo do modelo multimodal de identificação biométrica face-voz, utilizando regularização, o otimizador ADAM, camada de codificação do *autoencoder* com 1024 neurônios, 39 coeficientes MFCCs e diferentes valores de *frames* (32, 48, 64 e 192) no conjunto de teste.

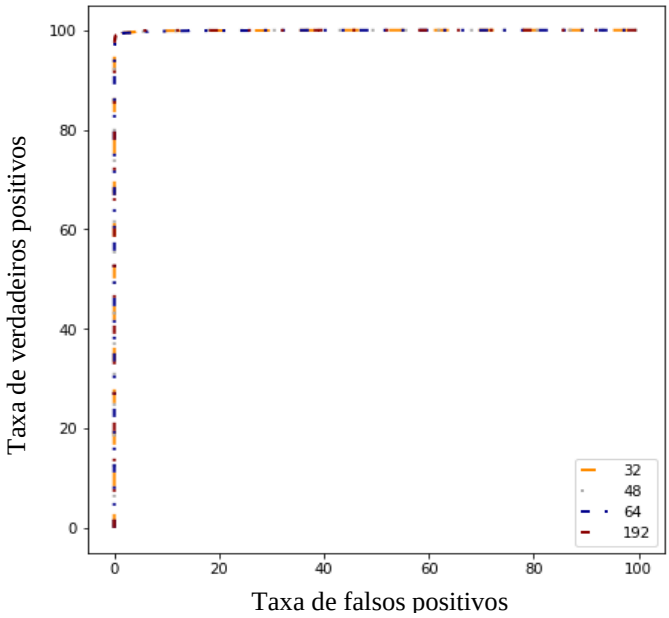


Figura 28: Curva ROC da abordagem de aprendizado profundo do modelo multimodal de identificação biométrica face-voz utilizando regularização, otimizador ADAM, *autoencoder* igual a 1024 neurônios,39 coeficientes MFCCs e diferentes valores de *frames* (32,48, 64 e 192) no conjunto de teste

5.3 DISCUSSÃO DOS RESULTADOS

Consideremos, inicialmente os resultados apresentados nas Tabelas 4, 5 e 6 e as curvas ROC apresentadas nas Figuras 23, 24 e 25, cenário do aprendizado com máquinas LVQ. A abordagem biométrica unimodal da face tem desempenho melhor do que a abordagem unimodal de voz. A maior AUC-ROC da primeira, 0,877, é obtida com 48 *clusters* e 64 *frames* por indivíduo, enquanto no segundo, a maior AUC-ROC, 0,94, é obtida com todas as configurações de *clusters* e com o maior tamanho da saída da camada de codificação do *autoencoder*. A abordagem bimodal tem desempenho superior à unimodal, com AUC-ROC igual a 0,98. Esse desempenho é alcançado com número de *frames* por indivíduo igual a 64 e com a dimensão da camada de codificação do *autoencoder* de 256 (16, 32 ou 48 *clusters*) e 512 (48 *clusters*) e independe do número de *clusters*, ou seja, 16, 32 ou 48.

Os melhores pontos de operação dos três sistemas de abordagem de aprendizado não profundo são obtidos com valores de limiar de 0,0082, para a abordagem unimodal utilizando voz, e 0,027, para a abordagem biométrica bimodal. O valor do limiar do sistema bimodal é a soma das pontuações dos sistemas unimodais de face e voz. Os valores de limiar obtidos neste trabalho para o ponto ótimo de operação na abordagem de identificação biométrica bimodal podem ser usados como guia para fixar o valor limite do sistema de identificação biométrica, desde que um procedimento de normalização semelhante ao deste trabalho seja aplicado.

O principal desafio de todas as etapas utilizando a abordagem não profunda é o treinamento das máquinas LVQ. Neste trabalho, as máquinas LVQ de face foram treinadas com 96 imagens e com 1000 épocas, resultando em um processo demorado, entre 15 e 20 minutos. Através de alguns experimentos, verificou-se que o desempenho da modalidade de identificação facial não diminui, se for utilizado um número menor de imagens e de épocas no processo de treinamento. Por exemplo, usando apenas 32 imagens, 16 *clusters* e 100 épocas, o desempenho do sistema de identificação da face é similar. Dessa forma, os resultados sugerem uma redução dos custos computacionais, diminuindo o tempo de treinamento.

Já para a máquina LVQ da voz, por exemplo, para 24 conjuntos de áudios e 32 *frames* de áudio, a máquina LVQ foi treinada com um total de 768 *frames* (24 x 32) de 32ms cada, resultando em 24,28s de gravação de voz. Verificou-se que mais amostras no conjunto de áudio não resultam em um melhor desempenho do classificador. Verificou-se também que o desempenho da abordagem de identificação por voz não diminui ao se utilizar apenas 384 *frames* (24 x 16) e 16 *clusters*. Sendo assim, os resultados sugerem utilizar o sistema biométrico da voz com um conjunto menor de amostras de áudios sem perda de desempenho, reduzindo o tempo de treinamento das máquinas LVQ.

Considerando os resultados apresentados nas Tabelas 7, 8, 9, 10, 11 e 12 e as curvas ROC das Figuras 26, 27 e 28, relativas ao cenário de abordagem de aprendizado profundo com redes CNN, observa-se que, com a combinação do otimizador ADAM, 1024 neurônios, 39 coeficientes MFCCs e 192 *frames*, os maiores valores da AUC-ROC dos modelos biométricos unimodais de face, 0,989, e voz, 0,980, foram similares. A abordagem bimodal teve desempenho superior à unimodal, com um valor de AUC-ROC igual a 0,9997.

Considerando somente os testes com as redes CNN, para o modelo de identificação facial, os resultados de AUC-ROC foram relativamente próximos, ficando no intervalo entre 0,96 a 0,98 de área sob a curva ROC. A redução do número de neurônios da camada de codificação do *autoencoder* não impactou de forma relevante nos resultados. Em contrapartida, o modelo de identificação da voz teve seu desempenho dependente do número de coeficientes MFCCs (39) e da quantidade *frames* (192), com intervalos de 0,76 a 0,98 nos valores de AUC-ROC. Devido a estrutura das redes CNN serem voltadas, em sua maioria, para aplicações de imagens, conseguiu-se extrair características relevantes das faces para distinção dos indivíduos, mesmo sobre cenários mais desafiadores de redução de dados. Porém, a escolha desse modelo para a característica dos dados de voz não é a mais adequada. Em trabalho futuros, sugere-se explorar uma arquitetura mais identificada com as características do sinal de voz, como por

exemplo as redes recorrentes, e para o modelo bimodal pode-se analisar uma arquitetura híbrida composta por uma rede neural convolucional para face e uma rede recorrente para voz.

Quando se compara o desempenho dos modelos de identificação multimodal de indivíduos utilizando redes CNN versus redes LVQ, as redes CNN atingiram desempenhos superiores ao método de aprendizado não profundo, em qualquer cenário, seja avaliando redes unimodal/unimodal, unimodal/bimodal e bimodal/bimodal, atingindo 0,99 de AUC-ROC no modelo de aprendizado profundo versus 0,98 no modelo sem aprendizado profundo. Dessa forma, a utilização de aprendizado profundo garante uma maior robustez aos modelos de identificação de indivíduos.

Ainda sobre as redes CNN, comumente, a utilização de regularização e os diferentes otimizadores produziram desempenhos com valores próximos, exceto o otimizador SGD para o cenário de unimodal da voz, que além de possuir uma convergência mais lenta durante o treinamento, o melhor resultado foi de 0,90 de AUC-ROC versus 0,98 com o otimizador ADAM e 0,98 com o otimizador RMSprop, para 39 coeficientes e 192 *frames*.

Em relação ao tempo de treinamento das redes LVQ e CNN, em geral, as redes CNN propostas tiveram um tempo de treinamento menor do que às redes LVQ, conforme visto nos resultados apresentados anteriormente. Isso ocorre devido à implementação das redes CNN utilizar camadas e bibliotecas otimizadas do *tensorflow*. Em uma aplicação com uma escala maior de indivíduos, a rede CNN é a opção mais recomendada. Em relação à rede CNN bimodal, os tempos foram relativamente menores do que as versões unimodais. Isso foi devido ao fato da rede bimodal ser treinada utilizando transferência de conhecimento (*tranfer learning*). No treinamento bimodal, os pesos das redes unimodais foram congelados, sendo apenas treinada a última camada de classificação, implicando numa redução do tempo de treinamento. O tempo de teste para as duas abordagens foi inferior a 3s. Na maioria dos casos, menor do que 1s. Também se observou que o tempo de treinamento foi proporcional à quantidade de dados inseridos nos experimentos.

Conforme já verificado, o sistema de identificação da face apresentou melhor desempenho em comparação ao sistema de identificação da voz. Uma justificativa para o desempenho melhor do sistema de reconhecimento de face pode vir do fato do *autoencoder* realizar uma filtragem eficiente de ruído do sinal da face, preservando somente os dados mais relevantes para sua representação, processo que descarta objetos indesejados na face e reduz o impacto da baixa iluminação, diferentemente do sinal de voz, que possui diversas interferências

indesejadas. A qualidade de gravação do dispositivo e o ruído ambiente também influenciam diretamente a taxa de identificação correta da voz.

Ressalta-se que o desempenho de cada modelo foi obtido com o uso da base de dados MOBIO, uma base adquirida em condições desfavoráveis como: alta variabilidade de poses, alta variabilidade de qualidade do sinal de voz e da face, com variações de ambientes de aquisições como iluminação, plano de fundo e ambientes acústicos. Apesar disso, os resultados dos sistemas unimodais de identificação da face e voz foram satisfatórios e dentro do estado da arte das duas modalidades para métodos que utilizam aprendizagem de máquina tradicionais. Melhores resultados seriam alcançados sob condições de ambientes controladas, com enquadramento da face (que evita a oclusão de alguma característica como olho ou sobrancelha), remoção de objetos indesejáveis, como óculos, escolhas de expressões faciais e uso de microfones de boa qualidade, com cancelamento de ruído, por exemplo.

5.4 COMPARAÇÃO DOS RESULTADOS COM A LITERATURA

Dos trabalhos apresentados na seção de revisão da literatura apenas um utilizou redes neurais convolucionais, porém dos três traços biométricos utilizados, apenas o traço da face foi empregado. Conforme visto no Quadro 1, o trabalho de Mehraj e Mir (2021) utiliza a métrica F1-score, com um valor de 99,54%. Para os trabalhos que utilizaram abordagens não profundas de classificação, o primeiro trabalho, Jiang, Sadka e Crookes (2010) apresntou uma taxa de erro de 35% no sistema de identificação multimodal. As principais métricas utilizadas nos demais trabalhos foram EER, TAR, FRR e FAR. Em Kumar e Swamy (2010), o sistema multimodal com modos de face e voz teve um desempenho medido através da métrica FAR de 6,26%. Aronowitz e colaboradores (2014), relataram, para o sistema multimodal face, voz e quirografia, um valor de EER de 0,49%. Zhang, Dai e Xu (2017), atingiram para o sistema multimodal fundido através de SVM, um valor para a métrica TAR igual à 93,6%, FAR igual à 0% e FRR igual à 1,4%. Abozaid et al. (2019) e Olazabal e colaboradores (2019), obtiveram, para a métrica EER, valores de 0,62% e 8,04%, respectivamente. Zhang e colaboradores (2020), avaliaram as métricas TAR, FRR e FAR para o seu sistema de autenticação biométrica com 102 indivíduos, atingindo um excelente resultado de 100% de TAR. Dinish e Rao (2020) atingiram resultados para um sistema de identificação biométrica multimodal face e voz, obtendo, para um classificador KNN, um valor de 0,25% para a métrica FRR de 0,25% e 0,75% para a métrica FAR. Por fim, Singh, Khanna e Garg (2020) desenvolveram um sistema para

identificação biométrica para sinais de face e impressão digital, tendo obtido para a métrica FAR um valor de 0%, para a métrica FRR, um valor de 9,125% e, para a métrica EER, um valor de 4,61%.

Deve-se ponderar que os trabalhos não utilizaram bases comuns e ainda possuem peculiaridade, níveis de dificuldade e critérios distintos de elaboração. Aliado a isto, os sistemas são do tipo identificação ou verificação e utilizaram modos biométricos além do face e voz, como impressão digital, quirografia e marcha. A utilização de diferentes métricas dificulta também a comparação entre os trabalhos apresentados. Em vista do exposto, a comparação dos sistemas de reconhecimento biométricos publicados não é trivial, não sendo viável afirmar qual proposta atingiu o melhor resultado.

Nas tabelas 13 e 14 foram consolidadas as métricas obtidas nos trabalhos citados para fins de comparação entre os resultados obtidos. Observa-se que não foi possível comparar todas as métricas. A Tabela 13 é relativa à comparação dos trabalhos que utilizaram técnicas clássicas de aprendizado de máquina, ou seja, técnicas não profundas.

Tabela 14, compara-se os trabalhos que utilizaram redes neurais convolucionais. Os valores apresentados de TAR e FAR para o trabalho ora apresentado, foram obtidos na condição do ponto ótimo de operação.

Tabela 13: Comparação entre os trabalhos publicados na literatura que utilizaram técnicas clássicas de aprendizagem de máquina e o sistema de identificação de multimodal utilizando redes LVQ proposto no trabalho ora apresentado.

Literatura	AUC-ROC	ER	EER	TAR	FAR	FRR
(JIANG; SADKA; CROOKES, 2010)	-	35%	-	-	-	-
(KUMAR; SWAMY, 2010)	-	-	-	-	6,26%	-
(ARONOWITZ et al., 2014)	-	-	0,5%	-	-	-
(ZHANG; DAI; XU, 2017)	-	-	-	93,6%	0%	1,4%
(ABOZAID et al., 2019)	-	-	0,62%	-	-	-

(OLAZABAL et al., 2019[a])	-	-	8,04%	-	-	-
(ZHANG et al., 2020)	-	-	-	100%	0%	0%
(DINESH; RAO, 2020)	-	-	-	-	0,75%	0,25%
(SINGH; KHANNA; GARG, 2020)	-	-	4,61%	-	0%	9,12%
Modelo utilizando redes LVQ desenvolvido	0,98	-	-	92,6%	7,7%	-

Tabela 14: Comparação entre os trabalhos publicados na literatura que utilizaram técnicas de aprendizagem profunda e o sistema de identificação de multimodal utilizando redes CNN proposto no trabalho ora apresentado

Literatura	AUC-ROC	F1-SCORE	TAR	FAR
(MEHRAJ; MIR, 2021)	-	99,54%	-	-
Modelo utilizando redes CNN desenvolvido	0,99	-	99,44%	0,27%

5.5 CENÁRIOS DE APLICAÇÃO

No Capítulo 1, abordou-se as principais aplicações para os sistemas biométricos. Quando se trata de reconhecimento biométrico para aplicações de acesso a dispositivos móveis, que exige a comparação de “n para 1”, o sistema proposto no trabalho ora desenvolvido, um “sistema de identificação”, teria que ser adaptado para um “sistema de verificação” biométrica. No sistema de verificação utiliza-se apenas as amostras de um indivíduo como verdadeiro positivo, o proprietário do dispositivo móvel, e amostras de vários outros indivíduos, como verdadeiro negativo. O ajuste do ponto ótimo, nesse caso, tem que ser feito no sentido de se

precaver-se contra um falso positivo. Para uma máquina LVQ, o custo computacional para o treinamento de um sistema de verificação é menor, na medida em que apenas uma comparação tem que ser feita, com os dados do indivíduo proprietário do dispositivo móvel. O trabalho vai residir no estabelecimento de um ponto ótimo de operação.

Para validar uma proposta de verificação, ainda que em ambiente *desktop*, seria necessário realizar ajustes para tornar o modelo em uma solução binária entre genuíno ou impostor. Também seria necessário adequar a disposição das amostras. Um exemplo de uma possível solução seria realizar experimentos para um conjunto de 5 indivíduos, para cada indivíduo da base dados identificá-lo como genuíno e os demais escolhidos de forma aleatória como impostor, separando as amostras de treinamento e teste. Em cada experimento, durante o treinamento, o sistema irá aprender as características do indivíduo e distingui-las das amostras impostoras. Já durante os testes, será avaliado o desempenho do sistema com as amostras não vistas. Seriam necessários, no mínimo, 50 experimentos para avaliar os 50 indivíduos utilizados neste trabalho.

Uma possível vantagem de se utilizar uma abordagem não profunda de aprendizagem é a simplificação em termos de desenvolvimento em sistemas móveis, pois, a criação de uma máquina LVQ não depende de bibliotecas complexas. Além disso, o sistema de fusão é a soma vetorial das contribuições de cada traço biométrico normalizado, técnica que não demanda grandes recursos computacionais para implementação e tempo de execução. Para o cenário dos dispositivos *Android*, o sistema operacional de dispositivos móveis mais difundido no mundo, é possível utilizar a linguagem de programação JAVA que contempla diversas bibliotecas matemáticas ou até mesmo bibliotecas específicas de aprendizado de máquina, como *Weka* ou *DeepLearning4J*.

O melhor cenário para aplicação das redes CNN desenvolvidas neste trabalho seria na identificação de indivíduos em aplicações *desktop*, criando uma aplicação responsável por treinar o modelo geral, e sempre que necessário incrementar/remover um indivíduo realizando um novo treinamento e uma segunda aplicação responsável por realizar a identificação se o indivíduo de interesse pertence ao grupo treinado. Para o cenário de dispositivos móveis, a utilização da biblioteca *tensorflow*, que foi utilizada neste trabalho, se torna mais complexa, visto que para o dispositivo móvel é necessário treinar o modelo localmente em dispositivo *desktop* e depois realizar o *deploy* do modelo treinado na aplicação do dispositivo móvel, o que inviabilizaria o processo de treinar um indivíduo totalmente no dispositivo móvel. Dessa forma,

seria necessário utilizar bibliotecas que permitam o desenvolvimento nativo no dispositivo móvel, *DeepLearning4J* é uma opção para o sistema operacional *Android*.

Embora os sistemas de reconhecimento biométrico estejam se tornando mais eficazes, esses não são infalíveis e possuem vulnerabilidade que devem ser consideradas em um sistema de autenticação de acesso. O principal ataque é chamado de *spoofing*, que consiste na tentativa de utilizar artefatos falsos ou réplicas de uma característica biométrica para enganar o sistema. Sendo assim, em uma aplicação comercial robusta, deve-se adicionar módulos de segurança para combater a ameaça de falsificação, como também um detector de vivacidade. É necessário, ainda, considerar os desafios de privacidade. No Brasil em 2018, foi sancionada a Lei Geral de Proteção dos Dados (LGPD), que classifica os dados biométricos como Dados Pessoais Sensíveis e esses devem ser tratados de forma especial, contendo camadas adicionais de segurança e controle. Uma biometria que esteja comprometida não tem a capacidade, como *tokens* ou *passwords*, de ser cancelada ou trocada. Assim, uma outra variação que pode ser explorada baseada neste trabalho é adicionar um módulo de biometria cancelável. Esse método busca proteger a informação biométrica por meio de aplicações de transformações nos dados biométricos. Caso os dados salvos sejam comprometidos, exclui-se a informação comprometida e realiza-se uma nova coleta e transformação.

6 CONCLUSÕES

O questionamento inicial que norteou esta pesquisa explorou se a utilização de redes neurais profundas seria uma técnica mais promissora na identificação biométrica de indivíduos em relação às técnicas tradicionais de aprendizagem de máquina. Assim, diante dos trabalhos revisados, com apenas um trabalho atuando com redes convolucionais, buscou-se analisar duas abordagens de aprendizagem de máquina para a tarefa de identificação de indivíduos. A primeira utilizou aprendizado por quantização vetorial e a segunda redes neurais convolucionais. Avaliou-se o desempenho dos modelos para sistema unimodais de face, voz e multimodal.

O trabalho investigou os desempenhos dos modelos sobre condições diferentes de quantidade de dados, variando o número de *clusters*, dimensão da camada de codificação do *autoencoder*, quantidade de *frames*, número de coeficientes MFFCs e diferentes técnicas de otimização. Explorou-se também a capacidade de identificação biométrica de indivíduos sobre maiores desafios de compressão de dados e concluiu-se que é possível reduzir a quantidade de dados para treinamento e obter resultados similares.

Para atingir resultados satisfatórios, foi fundamental o pré-processamento no banco de dados de face e voz. Nas amostras de imagens, foi realizado a detecção da face, redimensionamento e utilizando-se imagens em escala de cinza, em vez de RGB. Já nas amostras da voz, foi aplicado um detector de atividade de voz responsável por remover trechos ruidosos ou em silêncio. A contribuição desse trabalho para a base de dados MOBIO possibilita que outros pesquisadores possam explorar novas técnicas ou outras arquiteturas e compará-las com os resultados ora apresentados.

Em trabalhos futuros, é possível explorar arquiteturas que sejam mais adequadas à natureza unidimensional do sinal voz, utilizando, por exemplo, as redes recorrentes e, até mesmo, modelos híbridas, com redes convolucionais para o reconhecimento de face e com redes recorrentes para o sistema multimodal.

Para os dois cenários, com e sem aprendizagem profunda, obteve-se um desempenho satisfatório, acima de 0,98, para a métrica AUC-ROC do sistema multimodal. Além disso, a utilização de mais um modo biométrico para identificação de um indivíduo mostrou-se fundamental para aumentar o desempenho do modelo proposto. Ademais, a abordagem com redes profundas permitiu atingir um valor de 0,99 para a métrica AUC-ROC, com tempo de treinamento menor que a proposta utilizando LVQ. Assim, as arquiteturas propostas podem

constituir um bom ponto de partida para implementação de um sistema robusto de identificação automática de indivíduos.

REFERÊNCIAS

ABOZAID, A. et al. Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. **Multimedia Tools and Applications**, v. 78, n. 12, p. 16345–16361, 2019.

AGGARWAL; CHARU C. **Neural Networks and Deep Learning**. [s.l.] Springer, 2018. v. 10

ARONOWITZ, H. et al. Multi-modal biometrics for mobile authentication. **IJCB 2014 - 2014 IEEE/IAPR International Joint Conference on Biometrics**, p. 1–8, 2014.

BAZAGA, A. et al. A Convolutional Neural Network for the automatic diagnosis of collagen VI-related muscular dystrophies. **Applied Soft Computing**, v. 85, p. 105772, 2019.

COUNCIL, N. R.; COMMITTEE, W. B. **Biometric recognition: Challenges and opportunities**. [s.l.] National Academies Press, 2010.

DAS, S.; CAKMAK, U. M. **Hands-On Automated Machine Learning: A beginner's guide to building automated machine learning systems using AutoML and Python**. [s.l.] Packt Publishing, 2018.

DINESH, D. K.; RAO, P. V. Implementing and analysing FAR and FRR for face and voice recognition (multimodal) using KNN classifier. **International Journal of Intelligent Unmanned Systems**, v. 8, n. 1, p. 55–67, 2020.

GEARY, D. C. **Origin of mind: Evolution of brain, cognition, and intelligence**. [s.l: s.n.].

GÉRON, A. **Hands-on Machine Learning whith Scikit-Learing, Keras and Tensorfow**. [s.l: s.n.].

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

INTERPOL. **Cybercrime: Covid-19 Impact**. [s.l: s.n.]. Disponível em: <https://www.interpol.int/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19>.

IOFFE, S.; SZEGEDY, C. **Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift**. Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. **Anais**...: ICML'15.JMLR.org, 2015.

JAIN, A. K.; LI, S. Z. **Handbook of face recognition**. [s.l.] Springer, 2011. v. 1

JAIN, A. K.; ROSS, A. Multibiometric systems. **Communications of the ACM**, v. 47, n. 1, p. 34–40, 2004.

JAIN, A. K.; ROSS, A. A.; NANDAKUMAR, K. **Introduction to biometrics**. [s.l.] Springer Science \& Business Media, 2011.

JIANG, R. M.; SADKA, A. H.; CROOKES, D. Multimodal biometric human recognition for

perceptual humancomputer interaction. **IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews**, v. 40, n. 6, p. 676–681, 2010.

KINGMA, D. P.; BA, J. L. Adam: A method for stochastic optimization. **3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings**, p. 1–15, 2015.

KUMAR, M.; SWAMY, M. N. An iterative method for multimodal biometric face recognition using speech signal. **Communications in Computer and Information Science**, v. 131 CCIS, n. PART 1, p. 298–306, 2010.

LECUN, Y. et al. Gradient-Based Learning Applied to Document Recognition. **proc. OF THE IEEE**, 1998.

LI, F. et al. Classification of Heart Sounds Using Convolutional Neural Network. **Applied Sciences**, v. 10, n. 11, 2020.

LIN, M.; CHEN, Q.; YAN, S. Network in network. **arXiv preprint arXiv:1312.4400**, 2013.

MEHRAJ, H.; MIR, A. H. A Multi-Biometric System Based on Multi-Level Hybrid Feature Fusion. **HERALD OF THE RUSSIAN ACADEMY OF SCIENCES**, v. 91, n. 2, p. 176–196, mar. 2021.

NABIYEV, N.; MALEKZADEH, S. Anomalous Sound Localization Estimation. 2021.

OLAZABAL, O. et al. Multimodal biometrics for enhanced IoT security. **2019 IEEE 9th Annual Computing and Communication Workshop and Conference, CCWC 2019**, p. 886–893, 2019.

PATTERSON, J.; GIBSON, A. **Deep Learning: A Practitioner's Approach**. [s.l: s.n.].

PF. **Polícia Federal**. Disponível em: <https://www.gov.br/pf/pt-br/assuntos/noticias/2021/07/policia-federal-implementa-nova-solucao-automatizada-de-identificacao-biometrica>. Acesso em: 4 maio. 2022.

PRABHAKAR, S.; PANKANTI, S.; JAIN, A. K. Biometric recognition: Security and privacy concerns. **IEEE security \& privacy**, v. 1, n. 2, p. 33–42, 2003.

SHANG, D. et al. Multimodal-database-XJTU: An available database for biometrics recognition with its performance testing. **Proceedings of 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference, ITOEC 2017**, v. 2017- Janua, p. 521–526, 2017.

SINGH, L. K.; KHANNA, M.; GARG, H. Multimodal biometric based on fusion of ridge features with minutiae features and face features. **International Journal of Information System Modeling and Design**, v. 11, n. 1, p. 37–57, 2020.

SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, v. 15, n. 1, p. 1929–1958, 2014.

TIELEMAN, T.; HINTON, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. **COURSERA: Neural networks for machine learning**, v. 4, n. 2, p.

26–31, 2012.

TSE. **Biometria: identificação do eleitor pelas digitais garante mais segurança às eleições**. Disponível em: <https://www.tse.jus.br/imprensa/noticias-tse/2017/Marco/biometria-identificacao-do-eleitor-pelas-digitais-garante-mais-seguranca-as-eleicoes>. Acesso em: 4 maio. 2022.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, v. 1, n. February, 2001.

ZHANG, K. et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. **IEEE Signal Processing Letters**, v. 23, n. 10, p. 1499–1503, 1 out. 2016.

ZHANG, X. et al. An Efficient Android-Based Multimodal Biometric Authentication System with Face and Voice. **IEEE Access**, v. 8, p. 102757–102772, 2020.

ZHANG, X.; DAI, Y.; XU, X. Android-based multimodal biometric identification system using feature level fusion. v. 6, p. 120–124, 2017.

APÊNDICE A - TABELAS COMPLETAS DOS RESULTADOS

Tabela 15: Resultado da abordagem de aprendizado profundo do modelo unimodal de identificação biométrica utilizando o sinal de face no conjunto de teste

		Dimensão da camada de	Melhor ponto de operação					Tempo de
Regularização	Otimizador	codificação	Limiar	TAR (%)	FAR (%)	AUC	Tempo de treinamento(s)	teste (s)
Não	ADAM	1024	0,002	98,60	0,52	0,9791	132	0,4113
Não	ADAM	512	0,0001	98,67	0,86	0,9788	71	0,3343
Não	ADAM	256	0,0003	99,04	0,60	0,9793	70	0,3984
Não	SGD	1024	0,0096	98,44	1,08	0,9682	129	0,4148
Não	SGD	512	0,0037	98,44	1,32	0,9684	83	0,4075
Não	SGD	256	0,0081	98,10	0,81	0,9694	69	0,396
Não	RMSprop	1024	0,00001	99,42	0,96	0,9793	134	0,4162
Não	RMSprop	512	0,0002	98,77	0,53	0,9793	74	0,4075
Não	RMSprop	256	0,0003	98,77	0,52	0,9792	71	0,3536
Sim	ADAM	1024	0,0008	99,23	0,85	0,9898	130	0,7343
Sim	ADAM	512	0,0005	99,08	0,76	0,9892	83	0,4009
Sim	ADAM	256	0,0011	99,00	0,64	0,9891	78	0,4311
Sim	SGD	1024	0,0259	98,56	1,33	0,9786	120	0,4159
Sim	SGD	512	0,0329	98,67	0,84	0,9794	83	0,4001
Sim	SGD	256	0,022	98,92	0,87	0,9791	83	0,4037
Sim	RMSprop	1024	0,0015	98,98	0,45	0,9895	130	0,4042
Sim	RMSprop	512	0,0002	99,00	0,53	0,9895	84	0,4102
Sim	RMSprop	256	0,0002	99,15	0,57	0,9889	83	0,3553

Tabela 16: Resultado da abordagem de aprendizado profundo do modelo unimodal de identificação biométrica utilizando o sinal de voz no conjunto de teste

Regularização	Otimizadores	Número de coeficientes	Número de *frames* de testes por indivíduo	Melhor ponto de operação			AUC	Tempo de treinamento (s)	Tempo de teste (s)
				Limiar	TAR (%)	FAR (%)			
Não	ADAM	13	32	0,0093	78,79	33,86	0,8024	39	0,3135
Não	ADAM	13	48	0,013	83,65	30,26	0,8455	46	0,3143
Não	ADAM	13	64	0,0132	82,92	24,28	0,8926	53	0,3334
Não	ADAM	26	32	0,0149	82,42	37,98	0,7994	45	0,3339
Não	ADAM	26	48	0,0171	82,92	31,43	0,8383	62	0,3885
Não	ADAM	26	64	0,012	83,00	27,48	0,8805	81	0,437
Não	ADAM	39	32	0,0178	79,90	33,69	0,8068	57	0,3966
Não	ADAM	39	48	0,0092	81,19	34,56	0,8287	95	0,4579
Não	ADAM	39	64	0,0104	86,10	29,07	0,8781	112	0,5484
Não	ADAM	39	192	0,0123	93,10	8,05	0,9833	443	1,19
Não	SGD	13	32	0,0211	84,48	35,28	0,8136	82	0,3061
Não	SGD	13	48	0,0236	83,19	29,78	0,8402	73	0,3185
Não	SGD	13	64	0,0212	89,69	29,37	0,8811	82	0,3386
Não	SGD	26	32	0,022	84,19	38,39	0,7982	83	0,3362
Não	SGD	26	48	0,0242	83,13	33,97	0,8124	103	0,3887
Não	SGD	26	64	0,027	79,85	28,42	0,8338	142	0,434
Não	SGD	39	32	0,0249	81,67	36,59	0,7858	175	0,54
Não	SGD	39	48	0,0236	83,54	36,52	0,7994	222	0,6598
Não	SGD	39	64	0,0266	76,31	29,30	0,8149	322	0,8264
Não	SGD	39	192	0,0295	83,56	19,00	0,9092	862	2,1876
Não	RMSprop	13	32	0,009	80,77	36,12	0,8053	42	0,344
Não	RMSprop	13	48	0,013	87,21	33,89	0,84	53	0,3857
Não	RMSprop	13	64	0,0101	85,58	24,87	0,8902	86	0,4206
Não	RMSprop	26	32	0,0125	82,52	38,06	0,805	64	0,3641
Não	RMSprop	26	48	0,0173	83,17	31,89	0,8405	60	0,3861
Não	RMSprop	26	64	0,0161	84,33	25,75	0,8865	86	0,4314
Não	RMSprop	39	32	0,0079	79,65	35,92	0,7954	78	0,3914
Não	RMSprop	39	48	0,0135	82,56	33,38	0,8375	87	0,4543
Não	RMSprop	39	64	0,0112	83,27	26,46	0,8878	135	0,5538
Não	RMSprop	39	192	0,0112	93,10	8,57	0,9851	435	1,2081
Sim	ADAM	13	32	0,0171	87,58	36,30	0,8202	140	0,3115
Sim	ADAM	13	48	0,0192	86,94	30,06	0,8524	173	0,3078
Sim	ADAM	13	64	0,019	89,06	24,05	0,9014	203	0,3309
Sim	ADAM	26	32	0,0221	85,94	34,60	0,8189	209	0,3307
Sim	ADAM	26	48	0,0216	84,38	27,72	0,8525	280	0,3934
Sim	ADAM	26	64	0,0231	87,60	23,39	0,9013	357	0,4468
Sim	ADAM	39	32	0,0185	85,62	34,64	0,8162	323	0,3851
Sim	ADAM	39	48	0,0215	87,69	30,30	0,8531	415	0,462
Sim	ADAM	39	64	0,0209	89,44	25,86	0,8969	563	0,5475
Sim	ADAM	39	192	0,0286	96,40	6,10	0,9803	756	1,2748

Sim	SGD	13	32	0,0248	83,65	33,66	0,8146	82	0,3139
Sim	SGD	13	48	0,0239	86,23	33,09	0,8317	79	0,3175
Sim	SGD	13	64	0,0247	85,71	28,47	0,8643	90	0,3399
Sim	SGD	26	32	0,0246	80,79	36,57	0,7836	142	0,3512
Sim	SGD	26	48	0,024	79,75	35,19	0,7893	113	0,3995
Sim	SGD	26	64	0,0247	74,79	31,36	0,7943	142	0,4478
Sim	SGD	39	32	0,0238	77,04	36,82	0,7637	132	0,4098
Sim	SGD	39	48	0,023	76,46	36,85	0,7669	147	0,4703
Sim	SGD	39	64	0,0236	74,02	33,82	0,7738	202	0,5521
Sim	SGD	39	192	0,0269	78,67	23,37	0,8696	502	1,2148
Sim	RMSprop	13	32	0,0196	85,31	34,62	0,8191	37	0,3188
Sim	RMSprop	13	48	0,0167	87,21	31,34	0,8481	58	0,3124
Sim	RMSprop	13	64	0,0166	88,27	23,82	0,9014	87	0,3417
Sim	RMSprop	26	32	0,0176	83,44	34,15	0,8162	78	0,3442
Sim	RMSprop	26	48	0,016	85,69	30,36	0,8506	118	0,3967
Sim	RMSprop	26	64	0,0194	88,08	23,83	0,9033	140	0,4458
Sim	RMSprop	39	32	0,0207	85,21	34,83	0,8163	96	0,4072
Sim	RMSprop	39	48	0,0186	85,62	29,91	0,851	203	0,5594
Sim	RMSprop	39	64	0,014	91,83	27,62	0,9009	183	0,5487
Sim	RMSprop	39	192	0,0329	95,46	6,09	0,9886	470	1,2236

Tabela 17: Resultado da abordagem de aprendizado profundo do modelo multimodal de identificação biométrica (voz e face) no conjunto de teste

Regularização	Otimizador	Dimensão da camada de codificação	Número de coeficientes	Número de *frames* de testes por indivíduo	Melhor ponto de operação			AUC	Tempo de treinamento (s)	Tempo de teste (s)
					Limiar	TAR (%)	FAR (%)			
Não	ADAM	1024	13	32	0,0075	98,77	0,79	0,9988	100	0,6524
Não	ADAM	1024	13	48	0,0054	98,88	0,79	0,999	90	0,6275
Não	ADAM	1024	13	64	0,0162	98,75	0,47	0,9993	143	0,6129
Não	ADAM	1024	26	32	0,0091	98,58	0,59	0,999	94	0,6279
Não	ADAM	1024	26	48	0,0081	98,71	0,73	0,999	103	0,8604
Não	ADAM	1024	26	64	0,0058	98,92	0,77	0,9991	143	0,6736
Não	ADAM	1024	39	32	0,0035	98,77	0,95	0,9987	101	0,6335
Não	ADAM	1024	39	48	0,0094	98,79	0,68	0,9989	118	0,72
Não	ADAM	1024	39	64	0,0069	98,96	0,75	0,9992	143	0,7827
Não	ADAM	1024	39	192	0,01	99,19	0,46	0,9995	263	1,4324
Não	ADAM	512	13	32	0,0078	98,75	0,81	0,9987	71	0,528
Não	ADAM	512	13	48	0,0088	98,62	0,81	0,9989	80	0,537
Não	ADAM	512	13	64	0,0127	98,62	0,59	0,999	89	0,5492
Não	ADAM	512	26	32	0,0114	98,42	0,68	0,9981	89	0,549
Não	ADAM	512	26	48	0,0124	98,6	0,61	0,9987	143	0,6406
Não	ADAM	512	26	64	0,007	98,98	0,88	0,999	131	0,764
Não	ADAM	512	39	32	0,0099	98,6	0,67	0,9986	110	0,638
Não	ADAM	512	39	48	0,0068	98,92	0,88	0,9989	143	0,8033
Não	ADAM	512	39	64	0,0068	98,92	0,87	0,9989	175	0,9779
Não	ADAM	512	39	192	0,0235	99,02	0,37	0,9993	443	2,3001
Não	ADAM	256	13	32	0,0088	95,73	2,57	0,9946	71	0,4718
Não	ADAM	256	13	48	0,007	96,15	2,51	0,9947	79	0,5123
Não	ADAM	256	13	64	0,0019	97,54	2,62	0,997	89	0,5392
Não	ADAM	256	26	32	0,0026	96,73	2,39	0,9957	88	0,5489
Não	ADAM	256	26	48	0,0072	96,63	2,63	0,9948	108	0,6134
Não	ADAM	256	26	64	0,0071	96,71	2,34	0,9958	128	0,7244
Não	ADAM	256	39	32	0,003	96,65	2,64	0,9953	143	0,6305
Não	ADAM	256	39	48	0,0029	96,77	2,84	0,9953	141	0,7848
Não	ADAM	256	39	64	0,0083	96,42	1,87	0,9963	203	0,9496
Não	ADAM	256	39	192	0,0058	98,88	1,09	0,999	443	2,2609
Não	SGD	1024	13	32	0,0319	97,9	0,97	0,9982	143	0,5167
Não	SGD	1024	13	48	0,0206	98,35	1,3	0,9984	123	0,5156
Não	SGD	1024	13	64	0,0151	98,52	1,42	0,9982	166	0,4782
Não	SGD	1024	26	32	0,0235	98,15	1,17	0,9978	140	0,5702
Não	SGD	1024	26	48	0,0317	97,92	1,08	0,9979	143	0,59
Não	SGD	1024	26	64	0,0158	98,42	1,47	0,9979	98	0,5837
Não	SGD	1024	39	32	0,0151	98,52	1,58	0,9979	143	0,5266
Não	SGD	1024	39	48	0,0289	97,83	1,04	0,9979	103	0,6216
Não	SGD	1024	39	64	0,0333	97,77	1,04	0,9976	143	0,6911
Não	SGD	1024	39	192	0,0193	98,21	1,27	0,9982	263	1,3399

Não	SGD	512	13	32	0,0155	98,23	1,41	0,9985	65	0,4535
Não	SGD	512	13	48	0,0244	98,44	1,1	0,9987	70	0,4592
Não	SGD	512	13	64	0,0209	98,17	1,17	0,9984	83	0,4603
Não	SGD	512	26	32	0,0256	97,83	1,02	0,9983	75	0,4676
Não	SGD	512	26	48	0,0208	98,31	1,23	0,9984	88	0,5258
Não	SGD	512	26	64	0,0162	98,12	1,36	0,9984	94	0,5585
Não	SGD	512	39	32	0,0194	98,46	1,22	0,9987	84	0,4999
Não	SGD	512	39	48	0,0146	98,38	1,45	0,9986	98	0,5835
Não	SGD	512	39	64	0,0119	98,27	1,67	0,9983	143	0,6657
Não	SGD	512	39	192	0,0125	98,27	1,56	0,9983	232	1,3245
Não	SGD	256	13	32	0,0253	98,21	0,9	0,9983	64	0,4439
Não	SGD	256	13	48	0,0189	98,33	1,05	0,9983	69	0,4441
Não	SGD	256	13	64	0,0181	98,54	1,06	0,9985	83	0,4617
Não	SGD	256	26	32	0,0215	98,44	1,07	0,9983	72	0,4574
Não	SGD	256	26	48	0,0234	98,27	0,95	0,9981	143	0,5153
Não	SGD	256	26	64	0,026	98,02	0,84	0,9982	92	0,5607
Não	SGD	256	39	32	0,0291	98,15	0,92	0,9985	82	0,4983
Não	SGD	256	39	48	0,0134	98,58	1,27	0,9982	98	0,5713
Não	SGD	256	39	64	0,0244	98,1	1,04	0,9981	112	0,6594
Não	SGD	256	39	192	0,0165	98,27	1,2	0,9981	263	1,2991
Não	RMSprop	1024	13	32	0,0063	99,19	0,6	0,9995	71	0,4587
Não	RMSprop	1024	13	48	0,0074	99,12	0,55	0,9994	83	0,4899
Não	RMSprop	1024	13	64	0,0079	99,23	0,49	0,9996	80	0,4944
Não	RMSprop	1024	26	32	0,0066	99,12	0,57	0,9995	79	0,487
Não	RMSprop	1024	26	48	0,0058	99,06	0,56	0,9995	143	0,5488
Não	RMSprop	1024	26	64	0,0073	99,19	0,51	0,9995	99	0,5899
Não	RMSprop	1024	39	32	0,0076	99,15	0,53	0,9995	100	0,6176
Não	RMSprop	1024	39	48	0,0034	99,46	0,75	0,9996	102	0,6103
Não	RMSprop	1024	39	64	0,0131	99,08	0,37	0,9995	117	0,6957
Não	RMSprop	1024	39	192	0,0078	99,65	0,31	0,9999	237	1,3673
Não	RMSprop	512	13	32	0,0057	99,04	0,64	0,9994	70	0,454
Não	RMSprop	512	13	48	0,0071	99,08	0,53	0,9995	72	0,4633
Não	RMSprop	512	13	64	0,0057	99,17	0,63	0,9996	83	0,4672
Não	RMSprop	512	26	32	0,0038	99,29	0,71	0,9994	78	0,4609
Não	RMSprop	512	26	48	0,0075	99,12	0,53	0,9994	86	0,5088
Não	RMSprop	512	26	64	0,0062	99,33	0,57	0,9996	95	0,5573
Não	RMSprop	512	39	32	0,0046	99,21	0,74	0,9994	86	0,5587
Não	RMSprop	512	39	48	0,0031	99,17	0,71	0,9994	102	0,5928
Não	RMSprop	512	39	64	0,0037	99,35	0,8	0,9995	114	0,6676
Não	RMSprop	512	39	192	0,0069	99,62	0,38	0,9997	263	1,3116
Não	RMSprop	256	13	32	0,0079	98,98	0,49	0,9994	67	0,4483
Não	RMSprop	256	13	48	0,0032	99,08	0,7	0,9994	68	0,4426
Não	RMSprop	256	13	64	0,0051	99,25	0,62	0,9995	74	0,4649
Não	RMSprop	256	26	32	0,005	98,92	0,56	0,9993	104	0,6257
Não	RMSprop	256	26	48	0,0052	99,08	0,57	0,9993	143	0,5834
Não	RMSprop	256	26	64	0,0032	99,31	0,75	0,9994	106	0,6483
Não	RMSprop	256	39	32	0,0044	98,92	0,62	0,9992	97	0,5931

Não	RMSprop	256	39	48	0,0032	99,15	0,65	0,9994	143	0,6601
Não	RMSprop	256	39	64	0,0075	99,12	0,45	0,9995	125	0,7378
Não	RMSprop	256	39	192	0,0074	99,5	0,34	0,9997	241	1,3636
Sim	ADAM	1024	13	32	0,0101	98,81	0,5	0,9994	79	0,5039
Sim	ADAM	1024	13	48	0,0173	98,83	0,37	0,9996	90	0,5326
Sim	ADAM	1024	13	64	0,0134	99,1	0,39	0,9995	98	0,5608
Sim	ADAM	1024	26	32	0,0072	98,96	0,56	0,9995	102	0,5736
Sim	ADAM	1024	26	48	0,0088	0,88	0,47	0,9995	123	0,6645
Sim	ADAM	1024	26	64	0,0064	99,19	0,59	0,9996	203	0,7638
Sim	ADAM	1024	39	32	0,0056	99,08	0,69	0,9994	143	0,6601
Sim	ADAM	1024	39	48	0,0047	99,21	0,75	0,9994	169	0,8982
Sim	ADAM	1024	39	64	0,0113	99,04	0,46	0,9996	143	0,6818
Sim	ADAM	1024	39	192	0,0155	99,44	0,27	0,9997	261	1,3661
Sim	ADAM	512	13	32	0,0086	98,9	0,55	0,9992	69	0,457
Sim	ADAM	512	13	48	0,0076	99,15	0,58	0,9993	73	0,4657
Sim	ADAM	512	13	64	0,0138	99	0,4	0,9994	79	0,4782
Sim	ADAM	512	26	32	0,0061	99,17	0,66	0,9994	79	0,4676
Sim	ADAM	512	26	48	0,0058	99,15	0,68	0,9994	101	0,5385
Sim	ADAM	512	26	64	0,013	98,98	0,4	0,9993	143	0,5902
Sim	ADAM	512	39	32	0,0072	99,23	0,64	0,9994	106	0,6605
Sim	ADAM	512	39	48	0,0079	99,23	0,58	0,9992	124	0,7072
Sim	ADAM	512	39	64	0,0096	98,98	0,5	0,9993	140	0,7999
Sim	ADAM	512	39	192	0,0073	99,62	0,48	0,9996	324	1,5624
Sim	ADAM	256	13	32	0,0097	98,73	0,39	0,999	83	0,4842
Sim	ADAM	256	13	48	0,0039	99,08	0,62	0,999	82	0,5082
Sim	ADAM	256	13	64	0,0057	99,06	0,53	0,9991	143	0,5475
Sim	ADAM	256	26	32	0,005	99,06	0,54	0,9992	94	0,549
Sim	ADAM	256	26	48	0,0071	99,12	0,45	0,9992	113	0,6419
Sim	ADAM	256	26	64	0,0037	99,25	0,65	0,9989	135	0,7328
Sim	ADAM	256	39	32	0,0036	99	0,65	0,9992	115	0,6415
Sim	ADAM	256	39	48	0,0038	99,17	0,68	0,9991	203	0,7935
Sim	ADAM	256	39	64	0,0047	99,15	0,58	0,9993	183	0,9511
Sim	ADAM	256	39	192	0,0123	99,33	0,26	0,9995	454	22635
Sim	SGD	1024	13	32	0,0239	98,79	1,35	0,9989	74	0,4919
Sim	SGD	1024	13	48	0,0164	99,1	1,49	0,999	77	0,4993
Sim	SGD	1024	13	64	0,0366	98,56	0,89	0,9991	81	0,5292
Sim	SGD	1024	26	32	0,0193	98,85	1,37	0,9988	82	0,52
Sim	SGD	1024	26	48	0,0349	98,58	0,94	0,9989	94	0,5726
Sim	SGD	1024	26	64	0,0257	98,67	1,26	0,9987	105	0,6108
Sim	SGD	1024	39	32	0,0238	98,79	1,25	0,9989	143	0,5725
Sim	SGD	1024	39	48	0,0341	98,56	0,95	0,9988	111	0,6437
Sim	SGD	1024	39	64	0,0251	98,77	1,23	0,9988	127	0,728
Sim	SGD	1024	39	192	0,0254	98,81	1,12	0,9991	263	1,4024
Sim	SGD	512	13	32	0,0143	99,29	1,21	0,9991	101	0,6506
Sim	SGD	512	13	48	0,0184	99,12	0,98	0,9992	89	0,103
Sim	SGD	512	13	64	0,0403	98,69	0,65	0,999	93	0,8345
Sim	SGD	512	26	32	0,0232	99	0,91	0,9991	143	0,6192

Sim	SGD	512	26	48	0,0156	99,17	1,13	0,999	105	0,6301
Sim	SGD	512	26	64	0,0192	98,85	1	0,9989	115	0,6628
Sim	SGD	512	39	32	0,0164	99,06	1,1	0,9991	104	0,6219
Sim	SGD	512	39	48	0,0262	98,81	0,89	0,999	122	0,6889
Sim	SGD	512	39	64	0,0169	98,94	1,12	0,9991	137	0,7623
Sim	SGD	512	39	192	0,0168	99,1	1,05	0,9991	266	1,3831
Sim	SGD	256	13	32	0,0129	99,27	1,09	0,9991	83	0,4683
Sim	SGD	256	13	48	0,0242	98,98	0,7	0,9992	72	0,4613
Sim	SGD	256	13	64	0,0224	99,06	0,73	0,9993	83	0,4675
Sim	SGD	256	26	32	0,0213	98,79	0,8	0,9991	76	0,4784
Sim	SGD	256	26	48	0,0184	98,79	0,88	0,9992	89	0,5077
Sim	SGD	256	26	64	0,0105	99,12	1,2	0,9991	143	0,5629
Sim	SGD	256	39	32	0,0143	99,12	1,02	0,9989	143	0,5019
Sim	SGD	256	39	48	0,0154	99,04	0,94	0,9992	102	0,5837
Sim	SGD	256	39	64	0,0181	98,88	0,86	0,9991	143	0,666
Sim	SGD	256	39	192	0,0309	98,79	0,6	0,9992	250	1,3302
Sim	RMSprop	1024	13	32	0,0021	99,46	0,69	0,9997	73	0,4692
Sim	RMSprop	1024	13	48	0,0106	99,25	0,35	0,9997	83	0,4768
Sim	RMSprop	1024	13	64	0,0115	99,33	0,32	0,9996	81	0,4904
Sim	RMSprop	1024	26	32	0,0409	98,71	0,38	0,9992	92	0,5329
Sim	RMSprop	1024	26	48	0,0303	98,85	0,46	0,9993	93	0,5313
Sim	RMSprop	1024	26	64	0,0128	99,19	0,79	0,9994	143	0,5812
Sim	RMSprop	1024	39	32	0,029	98,73	0,46	0,9993	91	0,5328
Sim	RMSprop	1024	39	48	0,0179	98,96	0,62	0,9993	143	0,6168
Sim	RMSprop	1024	39	64	0,0231	98,88	0,52	0,9994	143	0,6916
Sim	RMSprop	1024	39	192	0,0329	99,06	0,37	0,9995	255	1,3455
Sim	RMSprop	512	13	32	0,0213	98,98	0,57	0,9993	69	0,4586
Sim	RMSprop	512	13	48	0,0347	98,96	0,36	0,9992	83	0,4633
Sim	RMSprop	512	13	64	0,0194	98,94	0,55	0,9992	90	0,5051
Sim	RMSprop	512	26	32	0,0167	99	0,61	0,9992	83	0,473
Sim	RMSprop	512	26	48	0,0123	98,98	0,76	0,9992	143	0,531
Sim	RMSprop	512	26	64	0,0344	98,77	0,4	0,9992	100	0,5624
Sim	RMSprop	512	39	32	0,0163	98,83	0,64	0,999	90	0,5148
Sim	RMSprop	512	39	48	0,0286	98,62	0,42	0,9991	105	0,5925
Sim	RMSprop	512	39	64	0,0198	98,94	0,58	0,9993	143	0,6676
Sim	RMSprop	512	39	192	0,0162	99	0,6	0,9994	253	1,3389
Sim	RMSprop	256	13	32	0,0158	99,08	0,55	0,9988	83	0,4463
Sim	RMSprop	256	13	48	0,0189	99,04	0,45	0,9987	83	0,4434
Sim	RMSprop	256	13	64	0,01	99,19	0,69	0,9987	77	0,4693
Sim	RMSprop	256	26	32	0,0177	98,85	0,49	0,9988	77	0,4659
Sim	RMSprop	256	26	48	0,0136	98,88	0,56	0,9986	88	0,5021
Sim	RMSprop	256	26	64	0,0198	98,83	0,45	0,9984	98	0,5579
Sim	RMSprop	256	39	32	0,0183	98,71	0,44	0,9986	86	0,5036
Sim	RMSprop	256	39	48	0,0313	98,81	0,35	0,9987	143	0,5877
Sim	RMSprop	256	39	64	0,0327	98,79	0,33	0,9988	143	0,6582
Sim	RMSprop	256	39	192	0,0127	99,12	0,54	0,9989	252	1,3089

APÊNDICE B – ARTIGO

Neste apêndice é apresentada a cópia do artigo originado deste trabalho. O artigo intitulado:" *Multimodal Biometric System Based on Autoencoders and Learning Vector Quantization*", foi apresentado no XXVII Congresso Brasileiro de Engenharia Biomédica (CBEB 2020), realizado na cidade de Vitória, Espírito Santo, de 26 a 30 de outubro de 2020.

Multimodal Biometric System Based on Autoencoders and Learning Vector Quantization

C. F. F. Costa Filho[1], J.V. Negreiro [1] and M. G. F. Costa[1]

[1]Centro de P&D de Tecnologia Eletrônica e da Informação – CETELI
Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Amazonas, Manaus, Brazil

Abstract— **This paper proposes a bimodal biometric verification system based on face and voice traits. The face characteristics are extracted using an autoencoder neural network. The voice characteristics are extracted using Mel-frequency cepstral coefficients. The matching procedure uses the Euclidean distance between one sample and the cluster centers obtained for each subject, through a learning vector quantization machine. The data fusion process is done through a simple normalization and sum of individual scores of the face-trait and the voice-trait. Several experiments are carried out varying the number of cluster centers, the size of the encoder output and the number of frames used for representing the voice trait of a subject. The performance of the biometric system is evaluated using the area under a receive operating characteristic (AUC of a ROC curve). The following performances are obtained: voice-trait biometric system: AUC =0.877; face-trait biometric system: AUC=0.94 and bimodal biometric system: AUC=0.98. The database used, the MOBIO, was collected from 50 individuals (37 male and 13 female) using mobile phones.**

Keywords— **Multimodal biometric system, face-trait, voice-trait, autoencoder, learning vector quantization.**

I. Introduction

Biometric Systems based on only one subject trait has been extensively used for personal recognition [1,2]. Although more secures than the association of ID cards and passwords, that can be stolen, these systems have limited resilience against noise due to the signal distortion during acquisition or other factors (low luminosity, fingerprint dirt, etc.) [3]. Recently, the development of multimodal biometric systems resulted in several advantages over unimodal systems, such as: better performance, security, and robustness [4].

Usually, the architecture of a multimodal biometric system is comprised of the following modules: signal acquisition, feature extraction, matching and decision. In Table 1 we summarize the information of seven systems, recently published, showing the characteristics of each one of these modules. As shown, these systems often use the combination of two biometric traits: face-voice [5,7,8], face-iris [6,9], fingerprint-voice [4], etc. Some authors argue that face recognition is more friendly and non-invasive, while iris recognition is more accurate [9].

All the sensors necessary to acquisition of the signals previously cited are now available in mobile phones, which has contributed to developments of multimodal biometric systems in these devices [4,7,10,11]. A challenge to implement multimodal systems in mobile devices is the extraction and size of multiple information data and the complexity of the recognition algorithm, especially when considering that. For security reasons, the training algorithm must be done on the device itself and not in the cloud. These aspects are constrained by the limited memory and power processing of mobile devices. In [10] the authors used a multilayer perceptron (MLP) for biometric recognition and pointed out a tradeoff between training and accuracy. To achieve an accuracy of 90%, with 45 training patterns, the MLP training took approximately 4 minutes. In [7] the authors used an FPGA to implement a KNN recognition algorithm to classify the features. For the training, the authors used 16 face and 16 voice samples from the genuine subject and the same amount from imposter subjects.

The feature extraction and data fusion are other key aspects of multimodal biometric systems. As shown in Table 1, for face feature extraction pyramidal Gabor wavelets [5], principal component analysis [6] and histogram-oriented gradient [7] have been used. For audio feature extraction, most of the papers used the Mel-frequency cepstral coefficients (MFCC).

For data fusion, most of the systems employed feature or score level fusion. According to [4] the performance of the recognition system using sensor fusion is not favorable as the raw features consist of many noise and redundant information. The same authors claim that the fusion in feature level may imply in obtaining a large feature vector, which will lead to the 'curse of dimensionality' problem.

In biometric systems, depending on the application, two classical systems are employed: identification or verification system. In the latter, the system verifies if the claimed identity, through a PIN card, for example, is true. In the former,

the system tries to discover one individual among several others.

In this work we propose a multimodal biometric system based on face and voice traits. Our main concern is obtaining simple processes for data extraction, data representation, fusion and classification, that accredit this system to be implemented on mobile devices. To store the face and voice trait information of each subject, we use a set of cluster centers, generated by three Learning Vector Quantization (LVQ) machines, two for the voice-trait and one for the face-trait. The scores of the matching process are generated through the Euclidean distance between a sample and a set of cluster centers representing a subject. In this paper we propose an autencoder network for face feature extraction. For audio feature extraction we used 12 MFCC and 12 delta MFCC. In this paper we propose a simple score level fusion based on normalization and sum of individual contributions of the traits. We seek to optimize the performance of the multimodal biometric system using a minimum size vector for face representation and a minimum number of clusters in the LVQ algorithm. We evaluate the performance of a verification system using a robust database, public available, collected using mobile devices sensors.

Table 1 Summary of published papers

Reference	Modality	Level of biometric fusion	Features	Matching	Database	Results
Huang et al. [5]	Bimodal: face and voice.	Score level: Golden Rate Algorithm (GRA)	Face: Pyramidal Gabor Wavelets (239 eigenvectors); Voice: 12 Mel-Frequency Cepstral Coefficients (MFCC)	Obtained through Probabilistic Neural Network	AT&T database: Training - 6 images, 12 MFCC; Testing: 4 images, 12 MFCC	Rank 1 identification rate Face: 96% Voice: 43.5% Bimodal: 100%
Chee et al. [4]	Bimodal: fingerprint and voice.	Feature level	Fingerprint: Minutiae Cilynder-Code (MCC); Voice: MFCC + PCA	Obtained through probability of collision between two hashed codes	Fingerprint: FVC2002 DB1 and FVC2002 DB2; Voice: NIST SRE	Equal error rate: bimodal 1%
Wang et al. [6]	Bimodal: face and iris	Feature level: normalization, using z-score model, and seralization	Face: Principal Component Analisys; Iris: 2D even Gabor filter	Matching score obtained using the Euclidian Distance in the series feature	Iris: Cassia database; Face: ORL database and Yale database	Equal rrror rate Face: 7.79% Iris: 3.11% Bimodal: 1.94%
Olazabal et al. [7]	Bimodal: face and voice	Feture level using discriminant correlation analysis (DCA)	Face: histogram oriented gradient (HOG) and LBP features; Voice: 45 MFCC	Obtained using k-Nearest Neighbor (KNN) classifier	own dataset of face photos and voice recordings captured using a Samsung Galaxy S5 device	Equal Error Rate Face: 16.91% Voice: 43.76% Bimodal: 8.04%
Chowdhury et al. [8]	Bimodal: face and voice.	Score level	Face: Encoder - Decoder + DR-GAN approach. Voice: 40 MFCC + 40 LBP coefficients + 1-D CNN.	Matching score obtained through sum rule, product rule, fusion rules 1,2,3 and 4.	MSU Audio-Visual Indoor Surveillance	Rank 1 identification rate: Face: 77% Voice: 40% Best bimodal 81%
Al-Waisy et al. [9]	Bimodal: face and iris	Score level	Face: Deep Belief Network (DBN); Iris: Convolutional Neural Network (CNN).	Matching score obtained through sum rule, product rule, weighted rule	Face: FERET; Iris: CASIA V1.0 and MMU1; bimodal: SDUMLA-HMT	Rank 1 identification rate: Face: 93.35% Iris: 100% bimoda:1 100%
Buriro et al. [10]	Bimodal: system profiles pressed screen points and the micro-movements of the phone during the signing process (sign & hold)	Feature level	Sign: location, pressure, size, orientation and velocity of the touch (80 features); Hold: data from accelerometer, the gravity sensor, and the magnetometer (13 features).	Obtained through Multi Layer Perceptron (MLP)	Proprietary database obtained with 30 individuals.	bimodal: true acceptance rate: 95%. false acceptance rate: 3.1%

The paper is organized as follows: in section II we describe the database, the bimodal biometric system architecture, the face and voice feature extraction, fusion and matching process. In section III we evaluate the performance of the biometric systems using the receiver operating characteristic (ROC) curves and the best operation point (true positive rate =1 and false positive rate =0), with the corresponding threshold. In section IV we evaluate the combination of parameters that result in the best recognition performance.

II. Methodology

A. Database

The database used in this paper, the MOBIO database, was presented in [11]. It is a face-voice database captured with the sensors of mobile phones. According to the authors, the main characteristics of this database are as follows: high variability of pose and illumination conditions, high variability in the quality of speech and variability in the acquisition environments in terms of acoustics as well as illumination and background. The database acquisition was inherently uncontrolled because the mobile phone was given to the user and so the recording device was no longer in a fixed position. We used the data partition collected in LIA and UNIS, with a total of 50 subjects (37 male and 13 female). For each subject, there are 192 separate audio-video samples. To record the audio samples, a dialogue manager was installed in the mobile phone and prompted the participants with short response questions, free speech questions, and to read a predefined text. The training set is formed by the first 96 face samples and the first 96 audio samples of each subject, while the testing set is formed by the remaining samples.

B. Architecture of the Multimodal Biometric System

Figure 1 shows the architecture of the proposed multimodal biometric verification system.

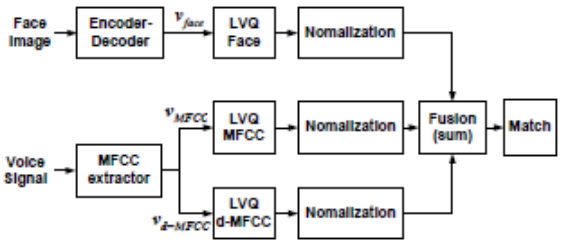


Fig. 1 Architecture of the multimodal biometric verification system

The dimensions of the face images at the input are 64x80 pixels. Before being presented to the autoencoder, they are linearized to a vector with length 5120, with a unit module. The vector at the encoder output, v_{face}, could have lenghts of 256, 512 or 1024.

C. Face Trait

The LVQ Face block shown in Figure 1 contains LVQ Machines with c clusters or codebooks [12]. The value of c can be equal to 16, 32 or 48. For each subject, a different LVQ machine is trained with 96 samples on the training set. For training, the number of epochs was fixed in 1000. The initial learning rate was fixed in 0.3 and decreased linearly with the number of epochs.

In Fig. 1, when a face test image is provided at the input, a matrix C_f with dimensions $(50\ x\ c)$ is generated at the LVQ Face block. Each line of matrix C_f contains the values of the Euclidean distances between the vector v_{face} and the codebooks of the c face clusters of a subject. At the output of the LVQ Face is generated a vector v_f of dimension $(50\ x\ 1)$ as follows:

$$v_{f_j} = \min\left(C_{f_{jk}}\right), for\ 1 \leq j \leq\ 50\ and\ 1 \leq k \leq c \quad (1)$$

The normalization block calculates a normalized vector v_{fn}: Each k coordinate of the vector v_{fn} is given by

$$v_{fn_k} = v_{f_k}/norm_f,\ 1 \leq k \leq\ 50 \quad (2)$$

$$norm_f = \left(\sum_{k=1}^{50} v_{f_k}\right)/50 \quad (3)$$

D. Voice Trait

The sample rate of the voice signal was 16kHz. Before sampling, the voice signal was processed with a voice activity detector (VAD); We used the one developed by Google to the WEB real time communication project - WebRTC [13], through the *webrtcvad* package available for the Python language. A VAD classifies a pi ece of audio data as being voiced or unvoice, discarding stretches of silence or noise. The aggressiveness of the VAD was set to 3, in a scale from 0 to 3. The voice frames size was 512 samples, or 32ms. The sample spacing was 8ms. Therefore, there is a superposition of 24ms in successive frames. From each voice sample of each subject we extracted 192 frames. A frame is only stored if is considered as voice by the VAD.

The Mel-frequency cepstral coefficients [14] are used in this work to represent the voice signal. Two groups of coefficients are employed, the 13 MFCCs and the 13 delta MFCCs. The steps to obtain the MFCCs can be stated as: 1. for each frame calculate the periodogram estimate of the

power spectrum; 2. apply the Mel filter bank to the power spectra, sum the energy in each filter; 3. take the logarithm of all filter bank energies; 4. take the DCT of the log filter bank energies and 5.keep DCT coefficients 2-13, discard the rest. In this work, the following parametrization of the python_speech_features's package was employed: an FFT with 512 points; a lower band edge of Mel filters of 0Hz and a higher band edge of Mel filters of 8kHz; the number of filters in the filter bank equal to 26 and 13 DCT coefficients were preserved.

The LVQ MFCC and LVQ d-MFCC blocks shown in Figure 1 are separated LVQ Machines for the MFCCs and delta MFCCs with c clusters, where c could be equal to 16, 32 or 48. For each subject, different LVQ machines (LVQ MFCC and LVQ d-MFCC) are trained with 768 samples. The number of epochs, initial learning rate and learning rate decrease policy used in the training of these machines were the same as the ones used in the LVQ face training.

In Figure 1, when a voice signal is presented at the input, n frames are extracted at the MFCC extractor block, where n could be equal to 32, 48 or 64. For each frame, at the output of the MFCC extractor block are generated two vectors, v_{MFCC_i} and v_{dMFCC_i} $(0 \leq i \leq n)$. At the LVQ MFCC and LVQ d-MFCC blocks are generated n matrices C_{M_i} and C_{dM_i}, respectively, with dimensions $(50\ x\ c)$. Each line of a matrix C_{M_i} and C_{dM_i} has the values of the Euclidean distances between the vectors v_{MFCC_i} and v_{dMFCC_i}, respectively, and the codebooks (cluster centers) of the c voice clusters of a subject. At the output of LVQ MFCC and LVQ d-MFCC are generated the vectors v_M and v_{dM}, respectively, with dimensions $(50\ x\ 1)$ as follows:

1. For each matrix C_{M_i} are generated vectors v_{M_i} and v_{dM_i} with dimensions $(50\ x\ 1)$ as follows:

$$v_{M_{i_j}} = \min\left(C_{M_{i_{jk}}}\right), for\ 1 \leq j \leq\ 50\ and\ 1 \leq k \leq c \quad (4)$$

$$v_{dM_{i_j}} = \min\left(C_{M_{i_{jk}}}\right), for\ 1 \leq j \leq\ 50\ and\ 1 \leq k \leq c \quad (5)$$

2. Vectors v_M and v_{dM} are generated with k coordinates, as follows:

$$v_{M_k} = \sum_{i=1}^{n} v_{M_{ik}}, for\ k = 1\ to\ 50 \quad (6)$$

$$v_{dM_k} = \sum_{i=1}^{n} v_{dM_{ik}}, for\ k = 1\ to\ 50 \quad (7)$$

The normalization block calculates normalized vectors v_{Mn} and v_{dMn} as follows:

$$v_{Mn_k} = v_{M_k}/norm_m \quad (8)$$

$$v_{dMn_k} = v_{dM_k}/norm_{dm} \quad (9)$$

Where:

$$norm_m = \left(\sum_{i=1}^{n}\sum_{j=1}^{50} v_{M_{ij}}\right)/(50\ x\ n) \quad (10)$$

$$norm_{dm} = \left(\sum_{i=1}^{n}\sum_{j=1}^{50} v_{dM_{ij}}\right)/(50\ x\ n) \quad (11)$$

E. Fusion and Matching

In the fusion process, a fusion score sum vector v_s is generated as follows:

$$v_s = v_{fn} + v_{Mn} + v_{dMn} \quad (12)$$

In the match block, to decide to which subject the face-voice traits belongs, we simple verify the row of v_s vector with the lowest value. The index of this row corresponds to the subject that the face-voice traits presented in the input of the biometric system belong.

III. Results

We will show the performance of biometric verification systems using voice and face traits, separated, and of the bimodal face-voice system. Tables 2, 3 and 4 show the percentages values of True Positives (TP), False Positives (FP) of the best operation point with the corresponding threshold used to obtain it. The best threshold is obtained considering the point of ROC curve nearly to the point (0,1). It is also shown the area under ROC Curve for all three systems, depending on the number of clusters and frame size used for testing.

Figure 2 shows the ROC curve of the voice trait biometric system, with the number of cluster equal to 48 and different values of n (number of test frames p/ subject). Figure 3 shows the ROC curve of the face trait biometric system, with the size of encoder output equal to 1024 and different number of clusters. Figure 4 shows the ROC curve of the face trait biometric system, with the size of encoder output equal to 1024 and different number of clusters.

IV. Analysis of the Results

From the analysis of the results presented in Tables 2, 3 and 4 and from the ROC curves presented in Figures 1, 2 and 3, some conclusions can be drawn. As shown, the face trait biometric system has a better performance than the voice trait biometric system. The best area under ROC curve of the former, 0.877, is obtained with 48 clusters and 64 test frames per subject, whilst in the latter, the best area under ROC curve, 0.94, is obtained with all configurations of clusters and size of encoder output. The bimodal biometric system has a superior performance, with an area under ROC curve equal

to 0.98. This performance is achieved with 16, 32 or 48 clusters, with number of frames per subject (voice trait) equal to 64 and with the size of encoder output of 256 (16, 32 or 48 clusters) and 512 (48 clusters). This best performance was expected, since it incorporates the recognition power of both traits, face and voice.

The best operating points of all three systems are obtained with threshold values between 0.82, for face-trait biometric system, to 2.7, for bimodal biometric system. The large values correspond to the bimodal biometric system. This was also expected, once the score value of the bimodal system is a sum of the scores of the face and voice scores.

Table 2 Results for the voice trait biometric system

Number of clusters	*n* - *n* umber of test frames p/ subject	Best operation point			Area under ROC curve
		Threshold	TP	FP	
16	32	1.93	77.83	19.12	0.869
16	48	1.94	75	23.3	0.833
16	64	1.93	78.41	18.82	0.871
32	32	1.93	80.16	18.91	0.872
32	48	1.93	72.91	19.98	0.836
32	64	1.93	80.08	18.8	0.875
48	32	1.93	78.33	18.27	0.874
48	48	1.94	77.25	22.13	0.845
48	**64**	**1.93**	**79.5**	**17.88**	**0.877**

Table 3 Results for the face trait biometric system

Number of clusters	Size of encoder-decoder output	Best operation point			Area under ROC curve
		Threshold	TP	FP	
16	1024	0.85	86.33	13.55	0.94
16	512	0.84	87.64	13.08	0.94
16	256	0.82	86	12.26	0.94
32	1024	0.85	86.95	13.5	0.94
32	512	0.84	87.52	12.96	0.94
32	256	0.82	86.17	12.23	0.94
48	1024	0.85	86.87	13.34	0.94
48	512	0.83	86.27	11.47	0.94
48	256	0.82	86.1	12.19	0.94

Table 4 Results for the bimodal biometric system

Number of clusters	n - Number of test frames p/ subject (voice trait)	Size of encoder-decoder output (face trait)	Best operation point			Area under ROC curve
			Threshold	TP	FP	
16	64	1024	2.75	90.5	6.4	0.977
16	64	512	2.72	90.16	5.11	0.979
16	64	256	2.71	90.41	5.59	0.98
32	64	1024	2.76	91.67	7.25	0.978
32	64	512	2.74	92.08	6.38	0.979
32	64	256	2.74	92.91	7.78	0.98
48	64	1024	2.77	92.25	8.05	0.978
48	64	512	2.74	91.92	6.42	0.98
48	64	256	2.74	92.58	7.69	0.98

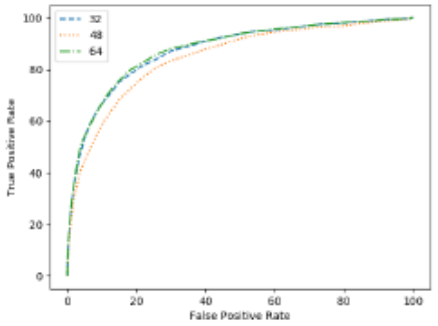


Fig. 2 ROC curve of the voice trait biometric system with the number of clusters equal to 48 and different values of *n* (number of test frames p/ subject)

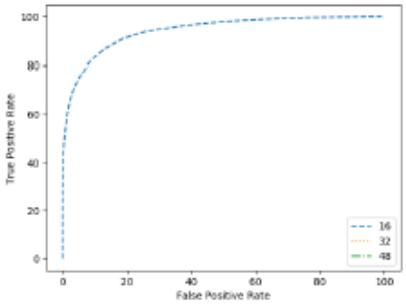


Fig. 3 ROC curve of the face trait biometric system with the size of the encoder output equal to 1024 and different number of clusters

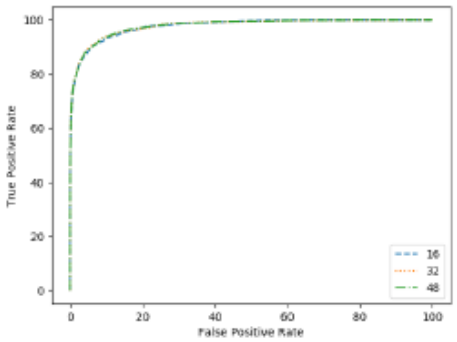


Fig. 4 ROC curve of the bimodal biometric system with the size of the encoder output equal to 1024 and different number of clusters

V. Conclusion

To apply this system to mobile phones, the main challenge is the training of the LVQ machines. In this paper the face trait LVQ machines were trained with 96 images and with 1000 epochs, resulting in a time-consuming process. Through some experiments, we verified that the performance of the face recognition modality thus not fall if we use a fewer number of images and number of epochs in the training process. For example, using only 32 images, 16 clusters and 100 epochs, the performance of the face-trait system is almost the same. The voice trait LVQ machines were trained with 768 frames of 32ms each one, resulting in 24.28s of voice recording. We verified that with more samples we did not obtain better performance of the classifier. We also verified that the performance of the voice-trait system does not fall we if we use only 384 frames and 16 clusters. In cell mobile phones, instead of having a verification biometric system, we have an identification biometric system. The threshold values obtained in this paper for the best operation point of the bimodal biometric verification system can be used as a guide to fix the threshold value for the biometric identification system, as long as a normalization procedure similar to the one applied here is used. In a future work we intend to show the results obtained with the implementation of the proposed method in mobile phones.

Acknowledgment

This research, carried out within the scope of the Samsung-UFAM Project for Education and Research (SUPER), according to Article 48 of Decree nº 6.008/2006(SUFRAMA), was funded by Samsung Electronics of Amazonia Ltda., under the terms of Federal Law nº 8.387/1991 through agreement 001/2020, signed with UFAM and FAEPI, Brazil.

Conflict of Interest

We declare that there are no conflicts of interest".

References

1. Macedo R, Costa, M and Costa Filho C (2013) Fingerprint verification using characteristic vectors based on planar graphics. Signal, Image and Video Processing, 9, 1121-1135.
2. Costa Filho C, Pinheiro, C, Costa, M and Pereira, W (2013) Applying a novelty filter as a matching criterion to iris recognition for binary and real-valued feature vectors. Signal, Image and Video Processing 7: 287-296.
3. Oloyede M and Hancke, G (2016) Unimodal and multimodal biometric sensing system: a review," IEEE Access 4: 7532 – 755.
4. Chee K-Y, Jin Z., Yap W-S. and Goi B-M (2018) Two-dimensional winner-takes-all hashing in template protection based on fingerprint and voice feature level fusion. In: Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference APSIPA, pp. 1411-1419.
5. Huang L, Yu C and Cao X (2018) Bimodal Biometric Person Recognition by Score Fusion. 5th International Conference on Information Science and Control Engineering, ICISCE.
6. Wang Z, Wang E, Wang S and Ding Q (2011) Multimodal biometric system using face-iris fusion feature. Journal of Computers 6(5): 1093-1097.
7. Olazabal O, Gofman M, Bai, Y et al (2019) Multimodal biometrics for enhanced IoT security. 2019 IEEE 9th Annual Computing and Communication Workshop and Conference, pp. 886-893.
8. Chowdhury A, Atoum Y, Tran L et al (2018) MSU-AVIS dataset: Fusing Face and Voice Modalities for Biometric Recognition in Indoor Surveillance Videos. International Conference on Pattern Recognition, pp. 3567-3573.
9. Al-Waisy A, Qahwaji , Ipson S et al (2017) A multimodal biometrie. 7th International Conference on Emerging Security Technologies, pp. 163-168.
10. Buriro A, Crispo B, DelFrari F et al (2016) Hold and Sign: A Novel Behavioral Biometrics for Smartphone User Authentication. IEEE Symposium on Security and Privacy Workshops, pp. 276-285.
11. McCool C, Marcel S, Hadid A et al (2012) Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data. 2012 IEEE International Conference on Multimedia and Expo Workshops, Melbourne, VIC, 2012, pp. 635-640.
12. Hagan M, Demuth H, Beale M and Jesús O (2019) Neural Network Design. 2nd Edition, eBook, Copyright by Martin T. Hagan and Howard B. Demuth.
13. WebRTC project: https://webrtc.org/. Accessed in April 2020.
14. Davis S. and Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28:357-366.

Corresponding author:

Author: Cícero Ferreira Fernandes Costa Filho
Institute: CETELI/ UFAM
Street: Av. Rodrigo Otávio Jordão Ramos, 3000
City: Manaus
Country: Brazil
Email: ccosta@ufam.edu.br