UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
ENG. DE CONTROLE E AUTOMAÇÃO

KAROLYNE PEREIRA DA SILVA - 243721

ANÁLISE DE APLICAÇÃO DE VISÃO COMPUTACIONAL E REDES NEURAIS, EM CONJUNTO COM O USO DE TÉCNICAS DE AUMENTO DE DADOS, NA TRADUÇÃO AUTOMÁTICA DE LIBRAS

Porto Alegre
2023

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
ENG. DE CONTROLE E AUTOMAÇÃO

KAROLYNE PEREIRA DA SILVA - 243721

ANÁLISE DE APLICAÇÃO DE VISÃO COMPUTACIONAL E REDES NEURAIS, EM CONJUNTO COM O USO DE TÉCNICAS DE AUMENTO DE DADOS, NA TRADUÇÃO AUTOMÁTICA DE LIBRAS

Trabalho de Conclusão de Curso (TCC-CCA) apresentado à COMGRAD-CCA da Universidade Federal do Rio Grande do Sul como parte dos requisitos para a obtenção do título de *Bacharel em Eng. de Controle e Automação* .

Orientador:
Prof.Dr. Heraldo José de Amorim

Co-orientador:
Me. Yachel Rogério Mileski

Porto Alegre
2023

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
ENG. DE CONTROLE E AUTOMAÇÃO

KAROLYNE PEREIRA DA SILVA - 243721

ANÁLISE DE APLICAÇÃO DE VISÃO COMPUTACIONAL E REDES NEURAIS, EM CONJUNTO COM O USO DE TÉCNICAS DE AUMENTO DE DADOS, NA TRADUÇÃO AUTOMÁTICA DE LIBRAS

Este Trabalho de Conclusão de Curso foi julgado adequado para a obtenção dos créditos da Disciplina de TCC do curso *Eng. de Controle e Automação* e aprovado em sua forma final pelo Orientador e pela Banca Examinadora.

Orientador: _____
Prof.Dr. Heraldo José de Amorim, UFRGS
Doutor pela (UFRGS – Porto Alegre, Brasil)

Banca Examinadora:

Prof.Dr. Heraldo José de Amorim, UFRGS
Doutor pela (UFRGS – Porto Alegre, Brasil)

Prof. Dr. Edson Cordeiro do Valle, UFRGS
Doutor pela (UFRGS – Porto Alegre, Brasil)

Prof. Dr. Herbert Martins Gomes, UFRGS
Doutor pela (UFRGS – Porto Alegre, Brasil)

_____
Alceu H. Frigeri
Coordenador de Curso
Eng. de Controle e Automação

Porto Alegre, Setembro 2023

DEDICATÓRIA

Dedico este trabalho à minha família, aos meus amigos que me ajudaram tanto em fazer o trabalho quanto ao não surtar fazendo esse trabalho e ao ministério de LIBRAS da IBFC.

AGRADECIMENTOS

Agradeço a Deus por ter me criado com um propósito e por ter me concedido inúmeras oportunidades, capacitando-me a aproveitar cada uma delas ao máximo. Expresso minha profunda gratidão à minha família, pela transmissão dos valores e princípios que moldaram minha jornada, assim como pelo constante apoio e amor que sempre me proporcionaram. Em particular, faço um reconhecimento especial à minha mãe e irmã, minha base sólida e maiores incentivadoras.

Também desejo expressar minha sincera gratidão à IBFC, que me proporcionou a valiosa oportunidade de aprender a língua tão especial que é a LIBRAS, enquanto convivia com a comunidade surda. Aos meus amigos, peço desculpas por ter recorrido tantas vezes ao "não posso"durante este período de dedicação ao TCC, e agradeço imensamente por não apenas compreenderem, mas também por me oferecerem apoio incondicional. Quero estender um agradecimento especial ao grupo do "cafézinho", amo vocês.

À orientação de Heraldo, meu orientador, e a Yachel, meu co-orientador, expresso minha profunda gratidão. Não apenas por apoiarem a minha escolha de tema para o TCC, mas também por suas revisões e incentivos constantes ao longo de todo o trabalho.

Não poderia deixar de mencionar o SENAI-RS, especialmente Martini e Solon, que me proporcionaram a oportunidade única de adquirir e aplicar conhecimentos práticos na área de visão computacional e aprendizado de máquinas.

RESUMO

A Língua Brasileira de Sinais (LIBRAS), uma língua de modalidade gestual-visual empregada pela comunidade surda no Brasil, enfrenta cotidianamente o desafio da barreira comunicacional entre surdos e ouvintes. Nesse contexto, este estudo busca desenvolver um sistema de visão computacional capaz de identificar sinais para auxiliar na tradução de LIBRAS para português, visando aumentar a inclusão da comunidade surda através da comunicação.

O escopo da pesquisa aborda fundamentos da LIBRAS, redes neurais com memória de longo e curto prazo (Long Short-Term Memory - LSTM), e a tecnologia Mediapipe. Além disso, o estudo compreendeu os treinamentos de 10 conjuntos, realizando modificações nos conjuntos de dados para avaliar o impacto nas métricas de desempenho. As modificações realizadas nos conjuntos de dados foram realizadas através do espelhamento horizontal, translação e aumento do brilho das sequências de imagens extraídas de vídeos contendo três sinais distintos.

A avaliação do sistema se deu por meio de métricas de desempenho, incluindo a taxa de acerto na tradução dos gestos. Os melhores resultados foram obtidos com o Conjunto 9. Esse conjunto utilizou um grupo de dados gerado a partir de vídeos de cinco indivíduos, cada um executando cada sinal por dez vezes, com subsequente aplicação das três técnicas de aumento de dados avaliadas. Isso resultou em uma acurácia de 100% tanto no treinamento quanto na validação, indicando um potencial promissor das ferramentas e metodologias empregadas neste trabalho na tradução de línguas de sinais.

Palavras-chave: Tradução de língua de sinais, visão computacional, redes neurais, aprendizado de máquina, inclusão social.

ABSTRACT

Brazilian Sign Language (LIBRAS) is a gestural-visual language used by the deaf community in Brazil that consistently faces the challenge of communicational barriers between deaf and hearing individuals. In this context, this study develops a computer vision system to translate LIBRAS into Portuguese, aiming to increase the inclusion of the deaf community through communication.

The research scope covers the fundamentals of LIBRAS, long short-term memory (LSTM) neural networks, and Mediapipe technology. Moreover, the study comprised ten distinct training sets, applying data augmentation techniques to the datasets to assess their impact on the performance metrics. The data augmentation techniques applied to the image sequences extracted from the videos in which three different signs were recorded were horizontal mirroring, offset, and brightness increase.

System evaluation was conducted through the analysis of performance metrics, including gesture translation accuracy rate. The best results were obtained by Set 9. This training used a dataset in comprised by videos of five individuals in whose each sign was performed ten times, followed by the use of the three data augmentation techniques evaluated. This resulted in a 100% accuracy rate for both training and validation, indicating good feasibility of translating sign languages with the tools and methodology employed in this work.

Keywords: Sign Language Translation, Computer Vision, Neural Networks, Machine Learning, Social Inclusion.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

LIBRAS	Língua Brasileira de Sinais
LSTM	Memória de Longo e Curto Prazo (do inglês, Long-Short Term Memory)
RN	Redes Neurais
ML	Aprendizado de Máquina (do inglês, Machine Learning)
CNN	Rede Neural Convolucional (do inglês, Convolution Neural Network)
RNN	Rede Neural Recorrente (do inglês, Recurrent Neural Network)
SVM	Máquina de Vetores de Suporte (do inglês, Support Vector Machine)
CSL	Língua Chinesa de Sinais (do inglês, Chinese Sign LAnguage)
TCC	Trabalho de Conclusão de Curso
CCA	Curso de Eng. em Controle e Automação

SUMÁRIO

1 INTRODUÇÃO

A Língua Brasileira de Sinais (LIBRAS) é uma língua visual-espacial utilizada pela comunidade surda no Brasil. Essa língua possui gramática e estrutura próprias, e que é amplamente difundida no território nacional. No entanto, a comunicação entre surdos e ouvintes que não conhecem a LIBRAS ainda pode ser um desafio. Nesse espaço, a tradução de LIBRAS para a língua falada permite uma comunicação mais eficiente entre surdos e ouvintes que não conhecem a língua de sinais. Essa capacidade de comunicação facilita a inclusão e promove a igualdade de oportunidades para pessoas surdas em diversas áreas da sociedade, como educação, trabalho e vida cotidiana. Devido ao custo elevado da tradução humana e à escassez de ferramentas de tradução, a maioria dos serviços públicos não é traduzida para a língua de sinais. Não existe uma forma escrita comum da linguagem de sinais, o que significa que toda comunicação escrita é realizada na língua falada local (COOPER; HOLT; BOWDEN, 2011). De acordo com a análise de Magno (2021), a promulgação da Lei Brasileira de Inclusão da Pessoa com Deficiência (LBI) ocorreu somente no ano de 2015. O autor observa, adicionalmente, que atualmente, pessoas surdas enfrentam obstáculos ao acessar serviços essenciais, como instituições bancárias, estabelecimentos de saúde e processos de seleção profissional, devido à ausência de indivíduos proficientes em LIBRAS. Esse cenário frequentemente os obriga a depender de amigos e familiares para a realização de interpretação ou, em alguns casos, a arcar com os custos de um intérprete profissional.

As dificuldades que os indivíduos surdos enfrentam não se limitam apenas ao cenário nacional, mas também têm alcance global, conforme destacado pela Federação Mundial dos Surdos (OMS - no inglês, WFD) a maioria das pessoas com perda auditiva carece de acesso a intervenções. A OMS também observa que atualmente, 1 em cada 5 pessoas em todo o mundo convive com algum grau de perda auditiva. Até o ano de 2050, estima-se que 1 em cada 4 pessoas poderá apresentar problemas relacionados à audição.

Tecnologia Assistiva (TA) é O termo usado para descrever a tecnologia empregada na reabilitação ou no aprimoramento da funcionalidade de indivíduos com deficiências (FEDERICI; SCHERER, 2018). Exemplos de TA voltados para a comunicação de pessoas com deficiência auditiva são aplicativos como HandTalk (HANDTALK, 2023), a ferramenta VLibras (DIGITAL, 2020), além do crescente uso de intérpretes em shows e eventos (VANINI, 2023). Nesse sentido, a visão computacional tem desempenhado um papel fundamental na busca por soluções para a tradução automática de LIBRAS em tempo real. A aplicação de técnicas de visão computacional aliada ao uso de redes neurais tem apresentado resultados promissores no campo da tradução e língua de sinais: algumas taxas de acerto reportadas por estudos na área são de 82.55% (QU; QIN; YIN, 2020), 92.19% (BINH; EJIMA, 2005) e 94.06% (KARAMI; ZANJ; SARKALEH, 2011).

Diante desse contexto, este trabalho tem como objetivo explorar e aplicar técnicas de visão computacional e redes neurais com unidades de memória de longo e curto prazo (do inglês, Long-Short Term Memory) na tradução automática de LIBRAS, buscando avaliar

sua viabilidade e eficácia em termos de precisão dos resultados. Espera-se que este estudo forneça informações relevantes sobre o uso dessas abordagens, especialmente os resultados obtidos por meio de treinamentos específicos destacados no capítulo de resultados.

2 REVISÃO BIBLIOGRÁFICA E ESTADO DA ARTE

Este capítulo apresenta brevemente os conceitos necessários na elaboração deste trabalho, com o objetivo de embasar o entendimento do sistema proposto. A seção 2.1 apresenta alguns princípios e desenvolvimentos da LIBRAS. A seção 2.2 aborda a visão computacional. Na seção 2.3 é explicado os conceitos básicos do aprendizado de máquinas, com foco em redes neurais (subseção 2.2 e 2.3.1). Após, a seção 2.3.1.1 apresenta a estrutura e o conceito de uma rede neural LSTM (memória de longo e curto prazo, do inglês Long Short-Term Memory). Por fim o estado da arte é abordado na seção 2.5, ressaltando o potencial da combinação de visão computacional, aprendizado de máquina e redes neurais na tradução automática de sinais.

2.1 LIBRAS

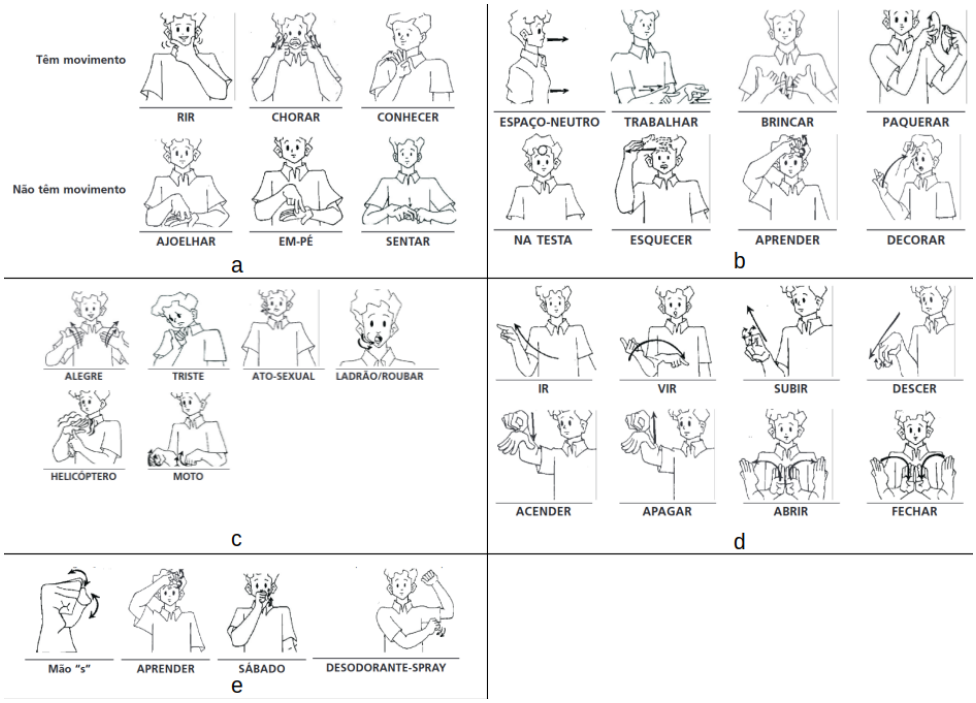
Considerando que a percepção das informações linguísticas ocorre por meio da visão e sua transmissão é realizada através de gestos manuais, as línguas de sinais são denominadas línguas de modalidade gesto-visual (KARNOPP, 2015). De acordo com Cooper, Holt e Bowden (2011), a língua de sinais vai além de uma simples coleção de gestos bem especificados. No estudo da fonologia das línguas de sinais, são ressaltados cinco parâmetros principais: movimento da mão (a), locação da mão (b), aspectos não-manuais (c),orientação da mão (d) e configuração de mão (e). Esses parâmetros são apresentados na Figura 1. Conforme Quadros e Karnopp (2007), a LIBRAS, utilizada pela comunidade surda no território nacional, apresenta uma organização espacial de complexidade comparável às línguas orais-auditivas. Os autores também enfatizam a relevância da investigação fonológica, morfológica e sintática no contexto das línguas de sinais.

O desenvolvimento de novas tecnologias no campo da tradução da língua de sinais também é destacado por sua importância na inclusão da comunidade surda (GALA, 2023).

2.2 Visão Computacional

A visão computacional é uma área da ciência da computação que se dedica ao desenvolvimento de algoritmos e técnicas capazes de extrair informações e compreender o conteúdo visual de imagens e vídeos. Segundo Szeliski (2011), a visão computacional tem como objetivo descrever o mundo em termos de imagens, reconstruindo suas propriedades, como forma, iluminação e distribuição de cores. Suas aplicações são amplas e abrangem diversos setores, como medicina (FUTURES, 2023), automação industrial (EDWARDS, 2023), segurança, robótica, realidade aumentada, monitoramento de tráfego (ANDRADE, 2020), reconhecimento facial, entre outros.

Figura 1: *Exemplo dos cinco parâmetros.*



Fonte: Felipe e Monteiro (2006)

A visão computacional exige o processamento de imagens. Uma imagem digital é representada por uma função discreta $f(x, y)$, onde x e y são as coordenadas da imagem. Essa função pode ser representada por uma matriz bidimensional MxN (GONZALEZ; WOODS, 2008):

$$\begin{bmatrix} f(0,0) & f(0,1) & f(0,2) \\ f(1,0) & f(1,1) & f(1,2) \\ f(2,0) & f(2,1) & f(2,2) \end{bmatrix}$$

Cada coordenada da matriz é chamada de pixel, e apresenta um valor (SATAPATHY et al., 2015). Esse valor pode ser composto por três canais (RGB, representando respectivamente as cores vermelho, verde e azul), com cada pixel da imagem sendo representado por uma combinação de intensidades dessas cores, ou possuir apenas um canal, onde cada pixel é representado por um valor único de intensidade luminosa. Enquanto o primeiro tipo permite obter uma vasta gama de cores, no segundo as cores são substituídas por diferentes tons de cinza, variando entre preto e branco conforme a intensidade do pixel. Imagens binárias são muito usadas em visão computacional. Nessas imagens os valores dos pixels variam entre zero (branco) e um (preto).

2.3 Aprendizado de máquina

O Aprendizado de Máquina (ML, em inglês, *Machine Learning*) é um campo da inteligência artificial que busca desenvolver algoritmos e técnicas capazes de permitir que sistemas computacionais aprendam e melhorem seu desempenho a partir da experiência adquirida com dados. De acordo com Burkov (2019), o aprendizado de máquinas também pode ser definido como um processo de reunir dados e, com esses, construir um modelo estatístico. Os algoritmos de aprendizado de máquina têm sido aplicados em uma ampla

variedade de áreas, como reconhecimento de voz, visão computacional, processamento de linguagem natural, análise de dados e tomada de decisão.

Diversas técnicas são utilizadas nesse campo, incluindo redes neurais artificiais, árvores de decisão, algoritmos genéticos e regressão linear (GOODFELLOW; BENGIO; COURVILLE, 2016). O estudo contínuo do aprendizado de máquina e o desenvolvimento de novas abordagens têm impulsionado avanços significativos na área da inteligência artificial. Conforme Hastie, Tibshirani e Friedman (2009), as técnicas de aprendizado de máquina são essenciais para lidar com grandes volumes de dados e extrair informações relevantes para auxiliar na tomada de decisão. Esses avanços têm permitido a criação de sistemas inteligentes e autônomos que podem aprender a partir da experiência, melhorando sua precisão e eficiência ao longo do tempo.

2.3.1 Rede Neural

Uma rede neural (RN) é um modelo computacional inspirado no funcionamento do cérebro humano, projetado para processar informações de maneira semelhante aos neurônios biológicos. É composta por unidades de processamento chamadas de neurônios artificiais, conectados por sinapses artificiais, que transmitem sinais entre eles (COSTA; BIANCHI; RIBEIRO, 2018). Matematicamente, uma rede neural é uma composição de transformações lineares e não lineares. Cada neurônio na rede recebe um conjunto de entradas ponderadas, realiza uma operação de soma ponderada dessas entradas e aplica uma função de ativação não linear para produzir uma saída, transmitindo esse resultado para outros neurônios conectados (FLECK et al., 2016). Uma RN é capaz de aprender a reconhecer padrões complexos nos dados de entrada e realizar tarefas como classificação, regressão e reconhecimento de padrões através de um treinamento adequado. Quando as redes neurais são aplicadas no processamento e análise de dados visuais, permitindo o desenvolvimento de sistemas capazes de identificar objetos, reconhecer rostos, interpretar cenas, entre outros aspectos relacionados à visão, tem-se a conexão com a área de visão computacional (RAWAT; WANG, 2017).

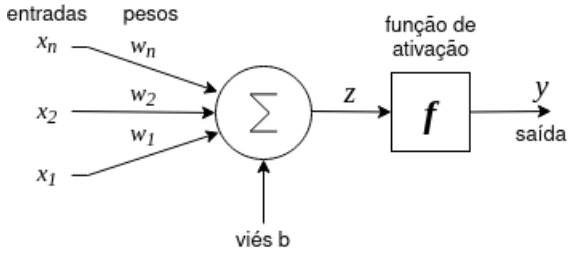
Considerando um treinamento supervisionado da rede, com o objetivo de minimizar uma função de custo (também chamada de perda, do inglês *loss*), que mede a diferença entre as saídas previstas da rede e os valores reais, durante o treinamento da rede neural, os pesos e os termos de viés são ajustados iterativamente com base em um algoritmo de otimização, método conhecido como retro propagação do erro (do inglês *backpropagation*) (SOUZA, 2010).

As redes neurais são utilizadas para extrair características relevantes das imagens e aprender a realizar tarefas específicas, contribuindo para avanços significativos na área de Visão Computacional. O método de aprendizado automático que envolve o uso de redes neurais artificiais com múltiplas camadas para aprender representações complexas e hierárquicas de dados é conhecido como *deep learning* (aprendizado profundo), no qual cada camada da rede neural processa informações em níveis de abstração crescentes, com as camadas mais profundas capturando características mais complexas e abstratas.

Uma rede neural pode ter L camadas. Em cada uma das camadas, os neurônios recebem um vetor de entradas x e um vetor de pesos w. A soma ponderada dessas entradas é calculada como o produto escalar entre os vetores de entrada e pesos, seguido de um termo de viés b, obtendo-se a saída z do neurônio:$\boldsymbol{z} = \boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b}$. A saída do neurônio é obtida aplicando uma função de ativação não linear f ao valor $\boldsymbol{z}$: $\boldsymbol{y} = f(\boldsymbol{z})$. A Figura 2

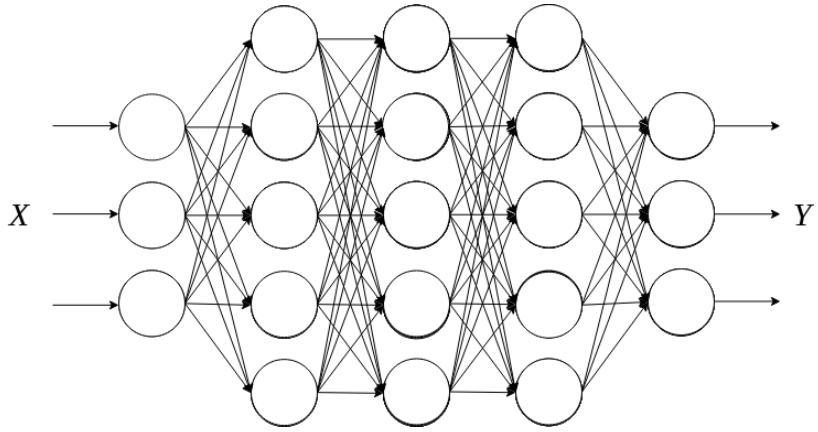
ilustra a estrutura de um neurônio artificial, enquanto a Figura 3 apresenta um exemplo de arquitetura de uma rede neural.

Figura 2: *Estrutura de um neurônio artificial.*



Fonte: Autor

Figura 3: *Exemplo de uma estrutura de uma rede neural com três camadas completamente conectadas.*



Fonte: Autor

2.3.1.1 LSTM

Dentre as possíveis estratégias de redes neurais, tem-se um tipo especial de célula de memória recorrente chamado LSTM (Long Short-Term Memory). Matematicamente, o funcionamento de uma LSTM envolve a manipulação de vetores de entrada, vetores de estado oculto (hidden state vectors) e vetores de célula de memória (memory cell vectors) (ALMEIDA, R. C. DE, 2019). Suponha que temos uma LSTM com uma entrada $x(t)$, um vetor de estado oculto $h(t)$ e um vetor de célula de memória $c(t)$ no tempo t. A LSTM realiza uma série de cálculos para atualizar esses vetores de acordo com as informações de entrada e as informações retidas na memória.

Os cálculos na LSTM envolvem portões (*gates*), que são camadas de unidades de processamento responsáveis por controlar o fluxo de informações, permitindo que a rede aprenda a esquecer informações irrelevantes, lembrar informações importantes e emitir saídas relevantes. Existem três tipos principais de portões em uma LSTM: o portão de esquecimento (f, *forget gate*), o portão de entrada (i, *input gate*) e o portão de saída (o, *output gate*).

O primeiro passo é calcular os valores dos portões através das equações (1), (2) e (3):

$$\mathbf{f_t} = \sigma(\mathbf{W_f} \cdot [\mathbf{h_{(t-1)}}, \mathbf{x_t}] + \mathbf{b}_f) \tag{1}$$

$$\mathbf{i_t} = \sigma(\mathbf{W_i} \cdot [\mathbf{h_{(t-1)}}, \mathbf{x_t}] + \mathbf{b}_i) \tag{2}$$

$$\mathbf{o_t} = \sigma(\mathbf{W_o} \cdot [\mathbf{h_{(t-1)}}, \mathbf{x_t}] + \mathbf{b}_o) \tag{3}$$

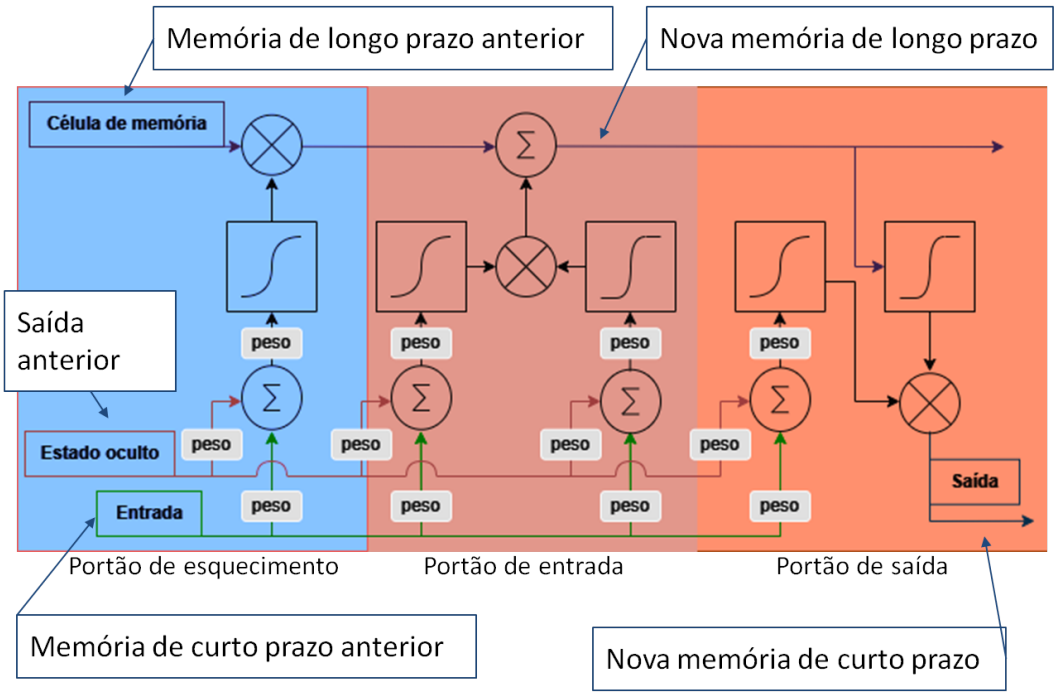
Em seguida, atualiza-se a célula de memória c_t na equação 4:

$$\mathbf{c_t} = \mathbf{f_t} \cdot \mathbf{c_{(t-1)}} + \mathbf{i_t} \cdot \tanh(\mathbf{W_c} \cdot [\mathbf{h_{(t-1)}}, \mathbf{x_t}] + \mathbf{b_c}) \tag{4}$$

Nesses cálculos, σ representa a função de ativação *sigmoide*, $*$ indica multiplicação elemento a elemento, *tanh* é a função hipérbole tangente e $(t-1)$ se refere ao estado anterior ao atual (valores da célula anterior).

A estrutura da LSTM (Figura 4) (LE et al., 2019) permite que ela aprenda a capturar dependências temporais de longo prazo, tornando-a especialmente útil em tarefas como processamento de linguagem natural, tradução automática, reconhecimento de voz e previsão de séries temporais, que envolvem sequências de dados.

Figura 4: *Estrutura de uma camada LSTM.*



Fonte: adaptada de Le et al. (2019)

O primeiro processo da rede LSTM, chamado de portão de esquecimento tem como entrada a memória de curto prazo da camada anterior ($\mathbf{h_{(t-1)}}$), a entrada atual ($\mathbf{x_t}$) e a memória de longo prazo anterior ($\mathbf{c_{(t-1)}}$), e as informações que não são mais consideradas

úteis são removidas. O segundo portão, onde uma porcentagem da nova memória é salva, é chamado de portão de entrada. Por fim, o portão de saída gera um nova memória de curto prazo (**h**), que será a entrada da próxima célula.

Utilizando técnicas avançadas de processamento de imagem e aprendizado de máquina em conjunto com redes neurais é possível realizar a construção de modelos de classificação e reconhecimento de sinais da LIBRAS, treinados a partir de um conjuntos de dados para assim desenvolver sistemas capazes de identificar e compreender os movimentos das mãos, expressões faciais e posturas corporais presentes na comunicação em LIBRAS.

2.4 Bibliotecas e Pacotes Utilizados

Nesta seção, serão apresentadas as principais bibliotecas e pacotes utilizados no desenvolvimento e implementação da pesquisa. Cada uma dessas ferramentas desempenhou um papel fundamental na análise de dados, processamento de imagens e aprendizado de máquina, contribuindo para o sucesso do projeto.

A manipulação de dados foi realizada com auxílio de uma biblioteca chamada Numpy. Essa biblioteca é essencial para computação numérica em Python, permitindo realizar operações matemáticas complexas com objetos do tipo array de forma eficiente e otimizada (almeida, m., 2023).

O pré-processamento das imagens utilizadas no estudo usou o OpenCV (Open Source Computer Vision Library). Essa biblioteca é amplamente utilizada para processamento de imagens e visão computacional (howse, 2013), e oferece uma vasta gama de funcionalidades para aquisição, processamento e análise de imagens e vídeos.

Para o treinamento das redes neurais foram usados o TensorFlow e o Keras. O primeiro é uma plataforma de aprendizado de máquina de código aberto desenvolvida pela Google. O Keras é uma API de alto nível para redes neurais que roda sobre o TensorFlow, simplificando a criação e treinamento de modelos de redes neurais. Neste trabalho o Keras foi usado para implementar e otimizar arquiteturas de redes neurais.

Todas as bibliotecas e pacotes mencionados foram programados utilizando a linguagem de programação Python. Segundo Carvalho (2023), a versalidade dessa linguagem possibilita seu uso em diversos projetos, se tornando uma das mais populares do mundo nos últimos anos.

2.5 Estado da arte

O estado da arte em Visão Computacional, impulsionado pelo advento das redes neurais artificiais, revolucionou a capacidade de computadores entenderem e interpretarem o conteúdo visual do mundo ao seu redor. As redes neurais convolucionais (do inglês *Convolutional Neural Network* - CNN) e as redes neurais recorrentes (do inglês *Recurrent Neural Network* - RNN) são fundamentais no o desenvolvimento de ferramentas essenciais para o contexto do dia a dia da sociedade (marques, 2016). A aplicação de CNNs estendeu-se para áreas como segmentação semântica e transferência de estilo, permitindo a identificação precisa de objetos em imagens complexas e a geração de conteúdo visualmente estilizado. Além disso, no contexto de tradução de línguas naturais, as redes neurais recorrentes (RNNs) e, em particular, a Long Short-Term Memory networks (LSTM), que será utilizada neste trabalho, têm desempenhado um papel crucial. Essas arquiteturas de

aprendizado profundo têm a capacidade de capturar padrões temporais em sequências de palavras, viabilizando a tradução automática eficiente e precisa de textos entre idiomas, o que tem sido uma área ativa de pesquisa e inovação (MARQUES, 2016).

Diversos estudos vêm sendo desenvolvidos com potencial para contribuir com o desenvolvimento de sistemas informatizados de tradução de línguas de sinais. Para o cenário da língua de sinais indiana Athira, Sruthi e Lijiya (2022) propõem uma metodologia em três passos: pré-processamento, extração de características e classificação. Na fase de pré-processamento os sinais são extraídos de vídeos usando segmentação pela cor da pele. Esse sinal passa por uma eliminação de co-articulações e resulta na geração de um vetor de características, que, por sua vez, é enviado para classificação através de um algoritmo de aprendizado de máquina supervisionado SVM (*Support Vector Machine*). Segundo os autores, o sistema permitiu uma acurácia de 91% na identificação do alfabeto manual e de 89% na identificação de sinais com apenas uma mão.

Xiao, Qin e Yin (2020) propuseram a comunicação bidirecional (entre ouvintes e surdos) através da geração e do reconhecimento de CSL (Chinese Sign Language - Língua de Sinais Chinesa) baseado em uma detecção dos principais pontos do corpo e em uma rede neural recorrente (RNN), chegando em uma acurácia de 82.55% para 500 sinais de CSL.

Börstell (2023) propôs o uso da ferramenta *Mediapipe* para extrair informações sobre articulações de sinais da língua sueca de sinais. O autor concluiu mostrando a eficácia da ferramenta para detectar se o sinal utiliza uma, duas ou ambas as mãos, qual a mão dominante e a principal localização do sinal. Em seu trabalho de conclusão de curso, Amaral (2021) utilizou a ferramenta *Mediapipe*. O procedimento adotado iniciou com a aquisição da imagem e extração dos pontos, após a qual uma sequência de cálculos foi executada para identificar se o sinal presente na imagem corresponde a uma das letras do alfabeto. Segundo o autor foram corretamente identificadas 20 das 26 letras do alfabeto manual de Libras, 11 delas com detecção razoável.

3 METOLOGIA

Este capítulo descreve a metodologia aplicada para a detecção de marcos das mãos e do corpo para a identificação dos sinais, bem como o tratamento dos dados. A seção de Detecção de Marcos (3.1) explora a identificação dos pontos de interesse, enquanto a etapa de Pré-processamento e Aumento de Dados (presentes na seção 3.2) aborda a manipulação e enriquecimento do conjunto de dados. Os sinais de LIBRAS selecionados para o estudo são apresentados na seção 3.2.3. A seção 3.4 apresenta os treinamentos aplicados à rede neural, bem como as métricas de desempenho utilizadas. Na seção 3.5 os modelos obtidos durante o treinamento são utilizados em um novo teste. Por fim, a seção 3.3 apresenta a sequência de operações aplicada aos dados utilizados.

3.1 Detecção de Marcos

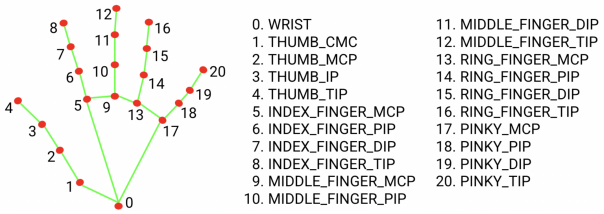
Para a identificação dos sinais é necessária a identificação dos movimentos realizados pelo indivíduo, o que exige a correta detecção dos marcos da mão e corpo. Neste trabalho esse procedimento foi realizado com auxílio do MediaPipe, uma estrutura de código aberto desenvolvida pelo Google que permite a construção de cadeias de processamento de mídia em tempo real. Essa ferramenta é uma biblioteca eficiente e flexível para o desenvolvimento de aplicações de visão computacional, reconhecimento de gestos, detecção e rastreamento de objetos, entre outras tarefas relacionadas ao processamento de mídia. O *MediaPipe* oferece uma variedade de módulos pré-construídos que podem ser facilmente integrados e configurados para criar fluxos de processamento personalizados. Sua arquitetura modular e sua capacidade de processamento em tempo real o tornam uma opção popular para aplicações que envolvem interações naturais com as mãos em ambientes virtuais ou aumentados, rastreamento de gestos ou outras aplicações de interface homem-máquina.

3.1.1 Pontos tratados

O Mediapipe possui uma ferramenta chamada *Hand Landmarker*, que utiliza uma rede neural convolucional (CNN) treinada especificamente para detectar a presença de mãos em uma imagem ou quadro de vídeo. Essa rede é capaz de identificar regiões da imagem que contêm mãos. Uma vez que a mão é detectada, o Hand Landmarker utiliza outra rede neural treinada para estimar a posição dos pontos-chave da mão, como as articulações dos dedos e a base da palma. Esses pontos-chave (Figura 5) são fundamentais para representar a pose da mão.

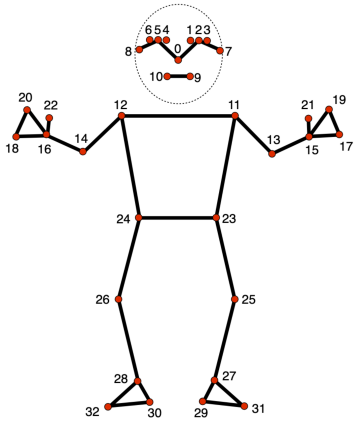
De forma semelhante, há também o *Pose landmark* (Figura 6) que identifica os principais pontos de uma pose de corpo inteiro.

Figura 5: *Pontos da mão.*



Fonte: Mediapipe (2023a)

Figura 6: *Pontos do corpo.*



Fonte: Mediapipe (2023b)

3.2 Dados

O conjunto de dados (do inglês, *dataset*) desempenha um papel fundamental em qualquer projeto envolvendo visão computacional e aprendizado de máquina, sendo assim de extrema importância no presente estudo. De acordo com Goetz (2022), é necessário a construção de uma base de dados para treinar algoritmos para aplicações que realizem algum tipo de análise sobre imagens. Um conjunto de dados de alta qualidade e representativo é essencial para treinar um modelo de tradução preciso e robusto: quanto mais diversificado e abrangente for o conjunto de dados, melhor será a capacidade do modelo de generalizar e traduzir corretamente os gestos da LIBRAS em diferentes contextos e variações.

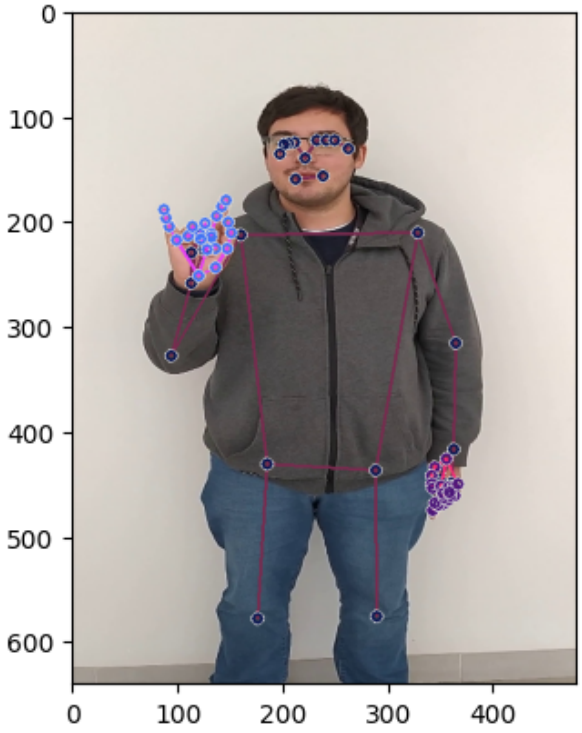
Foram adotadas duas estratégias iniciais no que se refere ao conjunto de dados: enquanto a primeira consistiu na busca na internet por fontes de dados existentes, a segunda estratégia envolveu a criação de um conjunto de dados próprio. Posteriormente, procedeu-se a um treinamento de teste empregando ambos os conjuntos de dados. Os resultados obtidos demonstraram maior precisão nos conjuntos de dados criados para o presente estudo, permitindo um desempenho mais satisfatório do modelo em questão.

3.2.1 Pré-processamento dos dados

Para a etapa inicial do pré-processamento houve a extração de imagens, que resulta na em um conjunto de 30 imagens por vídeo. O tamanho do conjunto de imagens foi definido visando reduzir a necessidade de memória e processamento no treinamento, permitindo representar um vídeo completo de forma mais leve. As imagens obtidas foram então

submetidas a uma fase de redimensionamento, com o propósito de padronizar as dimensões e otimizar os procedimentos manipulativos subsequentes. Posteriormente foram aplicadas técnicas para o aumento dos dados (detalhadas na Seção 3.2.2), com o intuito de identificar sua influência sobre o processo de treinamento. Espera-se que essas técnicas confiram maior robustez ao treinamento, como observado por Hanel (2021). Após a aplicação dessas transformações, todas as imagens foram convertidas para o formato de arquivo *.npy*, o que as tornou compatíveis para a utilização subsequente em atividades de análise e treinamento de modelos. A Figura 7 apresenta um exemplo de imagem com os pontos principais destacados após redimensionamento.

Figura 7: *Imagem retirada de vídeo, redimensionada e com pontos de destaque.*



Fonte: Autor

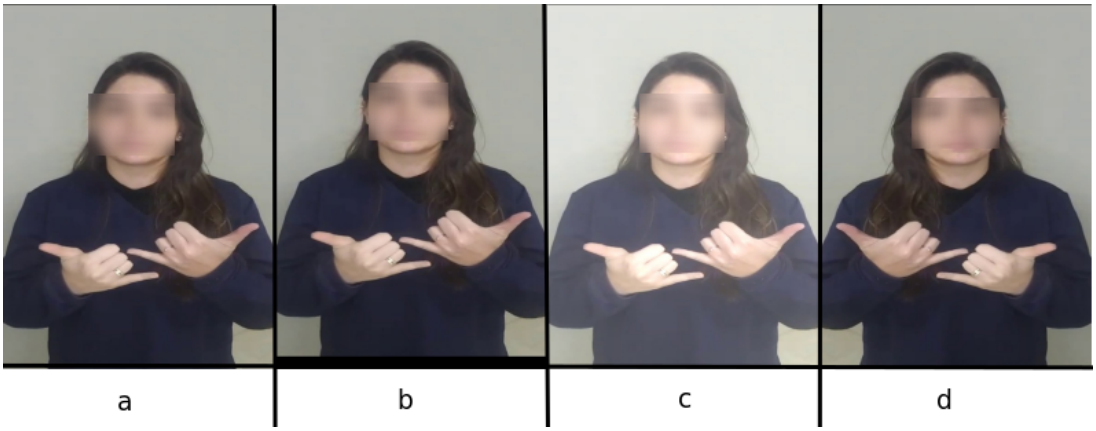
3.2.2 Data augmentation

Data augmentation, ou aumento de dados, é um conjunto de técnicas amplamente empregado no domínio de aprendizado de máquina e processamento de dados. Esse conceito consiste em aplicar transformações diversas nos dados existentes para gerar novas instâncias artificialmente. Essas transformações podem incluir rotações, espelhamento, ampliações, reduções, entre outras operações geométricas e de manipulação de cores. A relevância do *data augmentation* reside no fato de que ele permite aumentar significativamente a quantidade e a diversidade dos dados disponíveis para o treinamento de modelos de aprendizado, tornando-os mais robustos e generalizados. Com a geração de exemplos adicionais, o modelo é exposto a uma variedade maior de cenários, possibilitando uma melhor adaptação a diferentes situações reais e reduzindo o risco de *overfitting*, no qual o

modelo se torna excessivamente especializado nos dados de treinamento e não consegue generalizar para novos dados.

No decorrer deste estudo foram conduzidos treinamentos em dez conjuntos de dados distintos utilizando a rede neural LSTM, com o objetivo de comparar diversas técnicas de aumento de dados. Essas técnicas incluem o espelhamento horizontal das imagens, translação das imagens e manipulação do brilho das imagens, como ilustra a Figura 8. O intuito dessa abordagem é investigar o impacto dessas estratégias no desempenho das redes neurais LSTM. Ao realizar esses treinamentos comparativos, busca-se compreender de forma objetiva e imparcial como cada técnica afeta a capacidade do modelo em generalizar e lidar com cenários diversos, contribuindo, assim, para uma análise mais abrangente e informada sobre o melhor tratamento dos dados para obter resultados otimizados na tradução de LIBRAS para a língua portuguesa.

Figura 8: *Exemplos de imagens manipuladas.*



Fonte: Autor

Nota: A imagem apresenta, da esquerda para a direita, exemplos da aplicação a uma imagem de redimensionamento (a), translação(b), aumento do brilho (c), e espelhamento horizontal (d).

3.2.3 Sinais escolhidos

Para a realização deste estudo foram selecionados 3 sinais manuais distintos, visando contemplar diversidade no que se refere ao formato das mãos utilizadas, à localização espacial dos sinais e aos movimentos correspondentes. Os sinais escolhidos (Figura 9) para compor o conjunto de dados são os seguintes: 'Eu te amo', 'Brincar' e 'Casa'. A escolha desses sinais busca proporcionar uma representação de gestos manuais comumente empregados em diferentes contextos e situações comunicativas, possibilitando realizar uma análise do desempenho dos algoritmos de reconhecimento de sinais da LIBRAS a serem avaliados neste trabalho.

Figura 9: *Sinais utilizados no trabalho.*

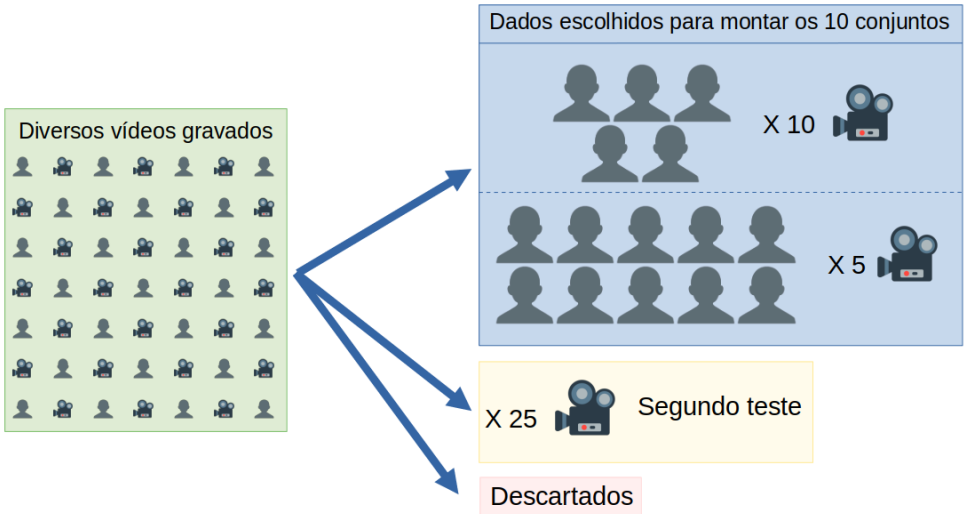
Fonte: Autor

Nota: Da esquerda para a direita são apresentados os sinais correspondente a "Casa", "Brincar"e "Eu te amo".

3.3 Gerenciamento de amostras e treinamentos da rede neural

Foram registradas múltiplas gravações audiovisuais, entretanto, determinadas dentre elas foram eliminadas devido a deficiências na execução do sinal, ou em virtude de inadequações no cenário visual foram separadas para um segundo teste, procedimento ilustrado na Figura 10.

Figura 10: *Gerenciamento dos vídeos gravados*

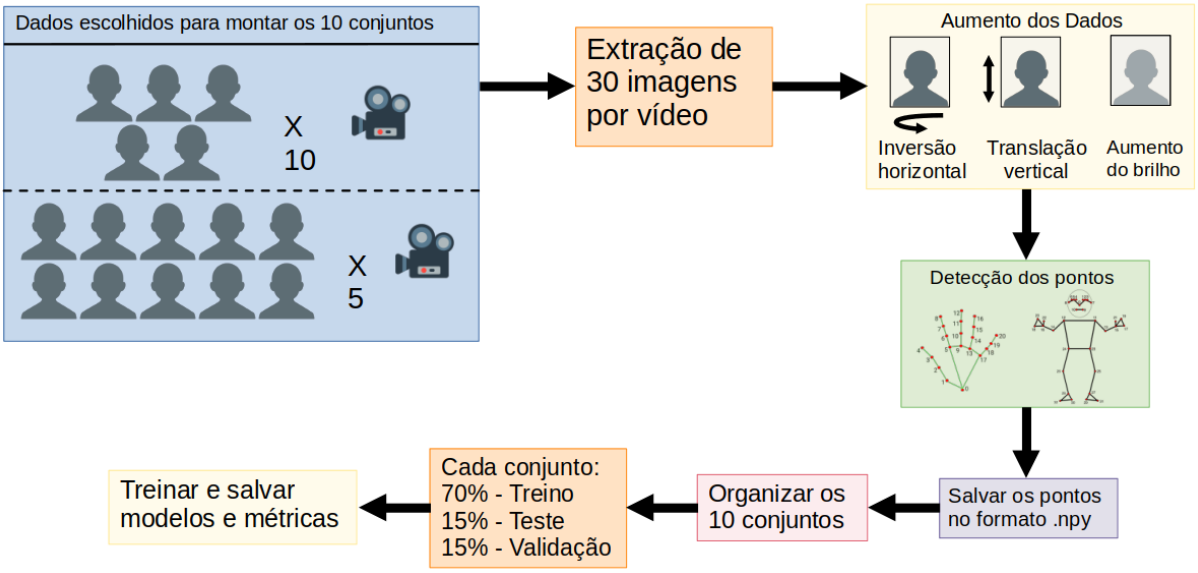


Fonte: Autor

Com o intuito de viabilizar a condução dos treinamentos da rede neural e a obtenção das métricas almejadas, cada um dos vídeos selecionados passaram pela extração de imagens, aumento de dados, detecção de pontos, salvar no formato .npy, organizado nos dez conjuntos. Cada um deles foi então dividido em três segmentos distintos, atribuindo-se

70% para a fase de treinamento, 15% para a etapa de teste e 15% para a fase de validação. O fluxo de procedimentos empreendidos, desde a seleção dos vídeos até a finalização dos treinamentos, é apresentado de maneira sequencial na Figura 11.

Figura 11: *Gerenciamento dos vídeos selecionados*



Fonte: Autor

3.4 Análise das redes neurais

Um dos objetivos deste estudo é analisar a influência de diferentes conjuntos de dados no treinamento da rede neural utilizada. Essa análise se configura como um elemento essencial para investigar a eficácia e eficiência dos diferentes conjuntos de dados em relação ao objetivo específico deste estudo. A avaliação abrangerá métricas pertinentes, tais como acurácia e perda, permitindo identificar os métodos mais promissores e fornecendo subsídios para a tomada de decisões na implementação de futuros sistemas de aprendizado de máquina voltados à tradução de LIBRAS.

Foram conduzidos treinamentos com os dez diferentes conjuntos de dados, cujas composições são apresentadas na Tabela 1).O estudo partiu de dois conjuntos de dados iniciais (Conjunto 0 e Conjunto 5), cada um com cinquenta amostras: o Conjunto 0 teve dez participantes, e utilizou, para cada sinal, dez vídeos gravados por cada participante; para o Conjunto 5 cada participante (5) gravou dez vídeos para cada sinal. No total, cada condição inicial consistiu em um *dataset* contendo cinquenta amostras. Os três conjuntos seguintes à condição inicial foram realizados com as técnicas de *data augmentation* avaliadas: aumento de brilho (Conjuntos 1 e 6), espelhamento (Conjuntos 2 e 7) e translação (Conjuntos 3 e 8). Por fim, um *dataset* contendo todas as transformações aplicadas foi utilizado para cada condição inicial (Conjuntos 4 e 9).

Os resultados foram avaliados através do uso de três métricas: Acurácia, Perda e F1_Score. A acurácia é uma métrica utilizada no treinamento de redes neurais que avalia a proporção de previsões corretas em relação ao total de exemplos de um conjunto de dados, fornecendo uma medida geral da exatidão do modelo (equação). A perda, por outro lado, é uma medida quantitativa da diferença entre as previsões do modelo e os valores reais

Tabela 1: *Composição de cada conjunto de dados.*

Conjunto	Normal	Mais Brilho	Espelhada	Transladada	Nº total de vídeos
0	X	-	-	-	50
1	X	X	-	-	100
2	X	-	X	-	100
3	X	-	-	X	100
4	X	X	X	X	200
5	X	-	-	-	50
6	X	X	-	-	100
7	X	-	X	-	100
8	X	-	-	X	100
9	X	X	X	X	200

Fonte: Autor

Nota: X representa a presença da técnica no conjunto de dados.

dos dados de treinamento. Durante o treinamento de uma rede neural busca-se minimizar essa métrica, através do ajuste dos pesos da rede. Por fim, o F1_Score é uma métrica que combina a precisão (*precision* - relação de verdadeiros positivos sobre todos os positivos previstos) e a revocação (*recall* - relação de verdadeiros positivos sobre todos os positivos reais) em um único valor. Essas métricas são calculadas através das equações 5, 6, 7 e (8), onde VP são os verdadeiros positivos, aqueles que foram classificados corretamente como pertencentes a uma classe, VN são os verdadeiros negativos, aqueles que foram classificados corretamente como não pertencentes a uma classe, FP são os falsos positivos, aqueles que foram classificados incorretamente como pertencentes a uma classe e o FN são os falsos negativos, aqueles que foram classificados incorretamente como não pertencentes a uma classe:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{5}$$

$$Precisão = \frac{VP}{VP + FP} \tag{6}$$

$$Revocação = \frac{VP}{VP + FN} \tag{7}$$

$$F1\_Score = 2 \cdot (precisão \cdot revocação)/(precisão + revocação) \tag{8}$$

3.5 Segundo teste

Considerando a expectativa de um desempenho superior para os treinos realizados com maior diversidade no conjunto de dados (Conjuntos 4 e 9), uma segunda avaliação foi conduzida especificamente nesses cenários. Nessa avaliação, os modelos resultantes dos treinos nos Conjuntos 4 e 9 foram utilizados para identificar sinais em vídeos que foram inicialmente descartados por apresentarem muitos elementos de fundo, resultando

em um fundo visualmente poluído (Figura 12). Esses vídeos foram considerados como potenciais candidatos para um segundo teste nas redes, visto que eles não estavam presentes em nenhum dos dez conjuntos. Antecipa-se que durante esse teste os modelos possam apresentar desempenhos inferiores em comparação com as etapas de treino, validação e teste iniciais, uma vez que os vídeos possuem uma complexidade maior devido à maior poluição visual.

Figura 12: *Exemplos de vídeos utilizados no teste*

Fonte: Autor

4 RESULTADOS E DISCUSSÃO

A evolução da acurácia e da perda no decorrer dos ciclos de treinamento e validação é apresentada na Figura 13. Esses conjuntos foram os que apresentaram resultados mais promissores (os gráficos dos demais resultados são apresentados no Apêndice A). É interessante observar que os modelos advindos dos Conjuntos 4 e 9 manifestaram resultados mais sólidos em relação aos modelos dos Conjuntos 2 e 7. Tal constatação decorre da observação de uma menor variabilidade, representada por oscilações menos pronunciadas nos gráficos de evolução, o que pode indicar uma maior habilidade desses modelos em lidar com novos dados. Isso pode ter relação com o maior tamanho dos *datasets* usados nesses treinamentos devido ao uso das três técnicas de *data augmentation*. Outro ponto importante é que, a fim de evitar overfitting, a rede foi modelada para que o treinamento fosse interrompido (*early stopping*) quando a perda na validação começasse a crescer (GENÇAY; QI, 2001). Com base nisso, uma convergência mais próxima entre os dados de validação e os de treinamento foi observada nos modelos dos Conjuntos 4 e 9.

Os resultados dos treinamentos realizados estão apresentados na Tabela 2, que inclui as métricas de perda, acurácia e $F1\_Score$.

Tabela 2: *Resultados dos treinamentos.*

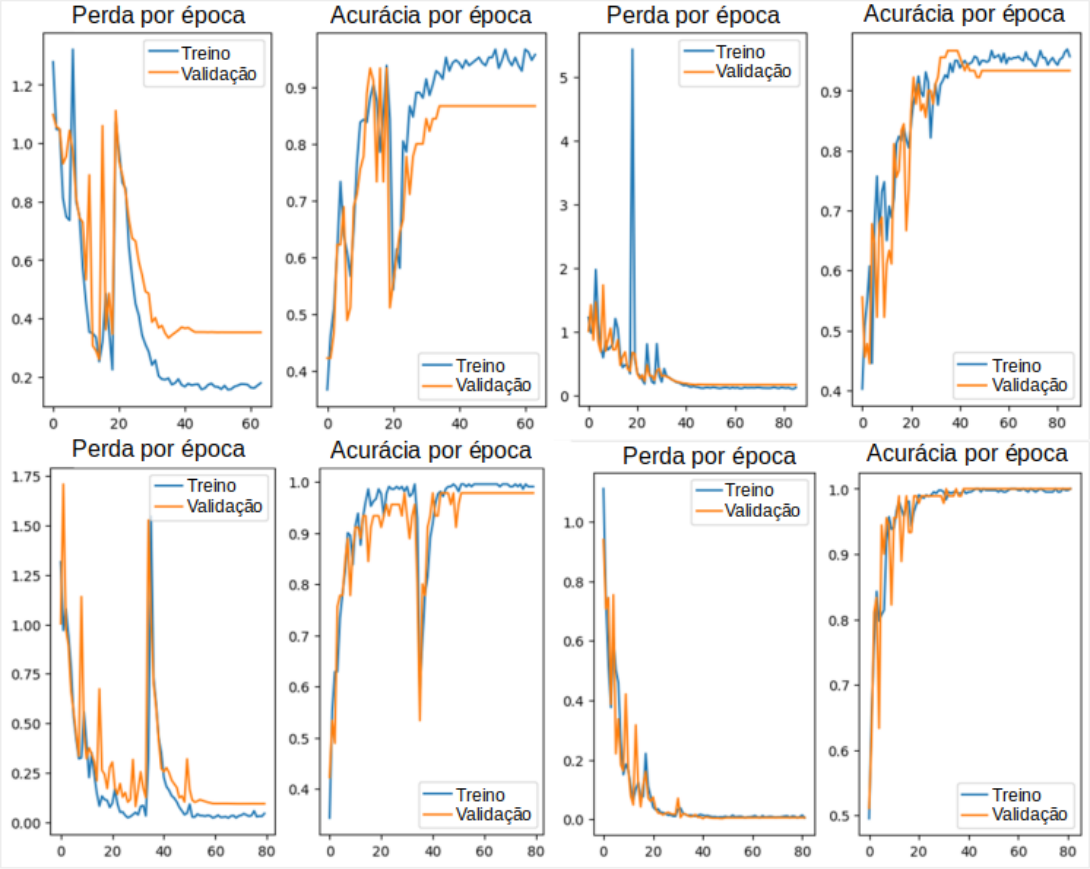
Conjunto	Nº de pessoas	Nº de vídeos/sinal	Perda no treinamento	Acurácia no treinamento	F1_score do teste
0	10	5	0.6896	0.8190	0,80/0,93/0,86
1	10	5	0.3612	0.8476	0.81/0.93/0.86
2	10	5	0.1791	0.9571	0.97/0.97/0.93
3	10	5	0.1126	0.9476	0.81/0.90/0.88
4	10	5	0.1273	0.9571	0.89/0.94/0.87
5	5	10	0.1604	0.9333	1.00/1.00/1.00
6	5	10	0.1313	0.9619	0,83/0,77/0,93
7	5	10	0.044	0.9905	0.97/1.00/0.97
8	5	10	0.0286	0.9905	0.90/0.90/0.93
9	5	10	0.0053	1	1.00/1.00/1.00

Fonte: Autor

Nota: O F1_Score é calculado para cada classe individualmente, e obedece a ordem Brincar/Casa/Eu te amo.

A análise dos dados presentes na Tabela 2, permite identificar diversas tendências. Para os cinco primeiros conjuntos de treinamento, em que dez participantes gravaram

Figura 13: *Gráficos de acurácia e perda por ciclo no treino e validação dos Conjuntos 2, 4, 7 e 9.*



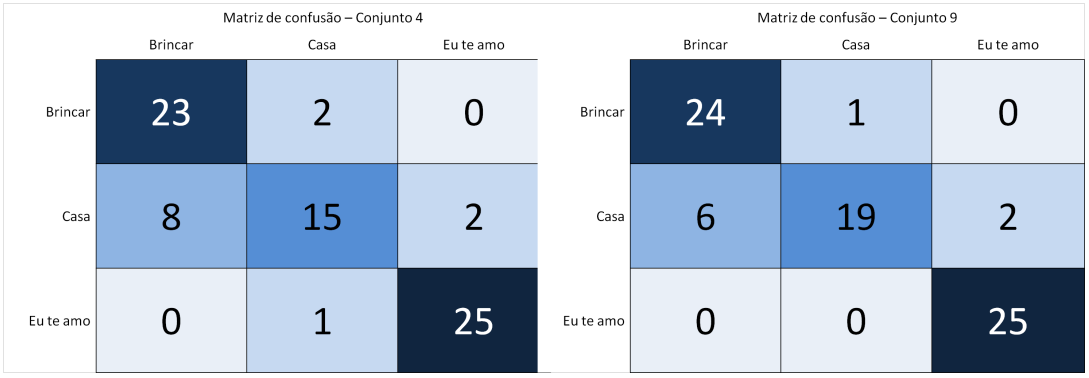
Fonte: Autor

cinco vídeos para cada sinal, os melhores desempenhos foram obtidos pelos treinamentos dos conjuntos 2 e 4. Para os cinco conjuntos de treinamento seguintes, os conjuntos de índices 7 e 9 foram os que exibiram métricas mais otimizadas. É importante ressaltar que tanto o segundo quanto o sétimo treinamento foram conduzidos utilizando conjuntos de dados compostos por imagens não manipuladas mais as imagens refletidas horizontalmente (espelhamento). Por outro lado, os treinamentos dos conjuntos 4 e 9 incorporam não somente as imagens em sua forma original, mas também aquelas que passaram pelos três processos de manipulação descritos na seção 3.2.2 referente ao aumento dos dados. Esses resultados indicam que, para as condições avaliadas neste estudo, o método de *data augmentation* mais eficiente foi o espelhamento horizontal.

Uma matriz de confusão é uma tabela usada em problemas de classificação para avaliar de forma rápida o desempenho de um modelo. Ela apresenta a contagem de previsões corretas (na diagonal principal) e incorretas (nos elementos fora da diagonal principal) para cada classe, permitindo a análise dos acertos (verdadeiros positivos e verdadeiros negativos) e dos erros (falsos positivos e falsos negativos), fornecendo uma melhor percepção sobre a eficácia do modelo na classificação das diferentes categorias (FONSECA, 2019).

A Figura 14 apresenta as matrizes de confusão obtidas após o teste com as amostras descartadas devido à maior poluição visual dos vídeos. Uma diferença significativa (16%) foi identificada entre os resultados obtidos pela rede após os treinos nos Conjuntos 4 e 9. Conforme esperado, o Conjunto 9 manteve a melhor colocação no teste.

Figura 14: *Matrizes de confusão da segunda avaliação*



Fonte: Autor

Nota: A Matriz da esquerda se refere aos resultados do treino no Conjunto 4, e na direita ao Conjunto 9.

5 CONCLUSÃO

Este trabalho investigou a viabilidade do uso de redes neurais, associadas com visão computacional, na identificação de sinais da LIBRAS. Através do uso de técnicas de *data augmentation* foi possível criar conjuntos de dados maiores a partir de um número limitado de amostras. Através das técnicas abordadas foi possível obter resultados promissores nos treinamentos da rede, em especial quando as três técnicas de *data augmentation* foram utilizadas em conjunto, indicando a viabilidade do uso dessas tecnologias na tradução automática de LIBRAS e outras línguas de sinais.

Dentre os resultados obtidos nesta pesquisa se destacaram os treinamentos dos Contjuntos 2, 4, 7 e, especialmente, o Conjunto 9 como o mais eficaz, evidenciando métricas superiores para uma notável variação no conjunto de dados. A utilização de técnicas de visão computacional e redes neurais LSTM na tradução automática de LIBRAS demonstrou ser uma abordagem promissora e eficaz para superar as barreiras de comunicação entre surdos e ouvintes. A capacidade das redes neurais LSTM em capturar dependências de longo prazo em sequências visuais, aliada a um cuidadoso tratamento e variedade dos dados, resultou em melhorias significativas na precisão e no desempenho em tempo real do sistema de tradução automática.

O trabalho aqui apresentado contribui de forma substantiva para o avanço da área de tradução automática de LIBRAS, não apenas explorando e aplicando técnicas de visão computacional e redes neurais LSTM, mas também fornecendo percepções valiosas sobre a importância da variação no conjunto de dados e a influência positiva dessa abordagem nos resultados. Espera-se que estudos e tecnologias nessa área avancem, para assim promover um ambiente mais inclusivo e igualitário.

Trabalhos futuros apresentam diversas áreas passíveis de investigação, abrangendo uma ampla gama de tópicos. Uma das possibilidades é a expansão da aplicação para um conjunto mais abrangente de sinais, tendo em vista que o Dicionário de Libras publicado em 2017 abriga mais de 13 mil sinais (CAPOVILLA et al., 2017). Além disso, explorar a complexidade associada à detecção de sinais altamente semelhantes representa um desafio relevante.

Outra perspectiva de pesquisa poderia se concentrar na análise de sequências de sinais, considerando a ordem em que são utilizados, bem como a avaliação da intensidade de cada sinal. Essa intensidade muitas vezes se manifesta através de expressões faciais e movimentos corporais, o que pode adicionar uma dimensão adicional de complexidade à detecção.

Por fim, a concepção e desenvolvimento de um aplicativo móvel constituiria um avanço substancial. Tal aplicativo teria o potencial de contribuir significativamente tanto para o ensino e aprendizado da Língua Brasileira de Sinais (LIBRAS) quanto para melhorar a acessibilidade do português para a comunidade surda.

REFERÊNCIAS

ALMEIDA, M. *Numpy: trabalhando computação científica com Python*. [S.l.: s.n.], mar. 2023. Disponível em: <https://www.alura.com.br/artigos/numpy-computacao-cientifica-com-python>.

ALMEIDA, R. C. DE. *Uso de Redes Neurais Long Short-Term Memory como Estratégia de Algorithmic Trading*. [S.l.: s.n.], 2019. Disponível em: <https://ele.ufes.br/sites/engenhariaeletrica.ufes.br/files/field/anexo/projeto_de_graduacao_-_rafael_costa_-_final_-_revisado.pdf>.

AMARAL, F. L. A. *Aplicação para o reconhecimento das letras do alfabeto manual em Libras*. [S.l.: s.n.], 2021. Disponível em: <https://github.com/fernandolucasaa/project_vision>. (accessed: 14.08.2023).

ANDRADE, M. M. DE. *Aplicação de Visão Computacional para Rastreamento e Contagem de Veículos em Rodovias*. [S.l.: s.n.], 2020. Disponível em: <http://repositorio.poli.ufrj.br/monografias/monopoli10032456.pdf>.

ATHIRA, P.; SRUTHI, C.; LIJIYA, A. A Signer Independent Sign Language Recognition with Co-articulation Elimination from Live Videos: An Indian Scenario. *Journal of King Saud University - Computer and Information Sciences*, v. 34, n. 3, p. 771–781, 2022. ISSN 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2019.05.002. Disponível em: <https://www.sciencedirect.com/science/article/pii/S131915781831228X>.

BINH, N. D.; EJIMA, T. Hand Gesture Recognition Using Fuzzy Neural Network. In. Disponível em: <https://api.semanticscholar.org/CorpusID:59924333>.

BÖRSTELL, C. *Extracting Sign Language Articulation from Videos with MediaPipe*. [S.l.: s.n.], 2023. Disponível em: <https://aclanthology.org/2023.nodalida-1.18.pdf>.

BURKOV, A. *The Hundred- Page Machine Learning Book*. [S.l.: s.n.], 2019. Disponível em: <http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf>.

CAPOVILLA, F. C. et al. *Dicionário da Língua de Sinais do Brasil: a libras em suas mãos*. [S.l.: s.n.], 2017. Disponível em: <https://repositorio.usp.br/item/002841837>.

CARVALHO, C. *O que é Python? História, Sintaxe e um Guia para iniciar na Linguagem*. [S.l.: s.n.], fev. 2023. Disponível em: <https://www.alura.com.br/artigos/python>.

COOPER, H.; HOLT, B.; BOWDEN, R. Sign Language Recognition. In: *Visual Analysis of Humans: Looking at People*. Edição: Thomas B. Moeslund. London: Springer London, 2011. P. 539–562. DOI: 10.1007/978-0-85729-997-0_27. Disponível em: <https://doi.org/10.1007/978-0-85729-997-0_27>.

COSTA, A. H. R.; BIANCHI, R. A.; RIBEIRO, C. *Redes Neurais Artificiais*. [S.l.: s.n.], 2018. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4461048/mod_resource/content/3/2018-NN.pdf>.

DIGITAL, G. *Tradução automática para tornar a Web mais acessível*. 2020. Disponível em: <https://www.gov.br/governodigital/pt-br/vlibras>. (accessed: 16.08.2023).

EDWARDS, C. *Generative AI Opens New Era of Efficiency Across Industries | NVIDIA Blog*. [S.l.: s.n.], jul. 2023. Disponível em: <https://blogs.nvidia.com/blog/2023/07/13/generative-ai-for-industries/>.

FEDERICI, S.; SCHERER, M. *Assistive Technology Assessment Handbook*. 2018. Disponível em: <https://www.researchgate.net/publication/321683103_Assistive_Technology_Assessment_Handbook>. (accessed: 16.08.2023).

FELIPE, T. A.; MONTEIRO, M. S. *Libras em Contexto: Curso Básico : Livro do Professor*. 2006. Disponível em: <https://jucienebertoldo.files.wordpress.com/2018/03/libras-em-contexto.pdf>. (accessed: 03.08.2023).

FLECK, L. et al. *Redes Neurais Artificiais: Princípios Básicos*. [S.l.: s.n.], 2016. Disponível em: <https://periodicos.utfpr.edu.br/recit/article/viewFile/4330/Leandro>.

FONSECA, M. G. B. D. Redes Neurais Artificiais Aplicadas à Classificação de Gestos da Mão através de Sinais Eletromiográficos, 2019. Disponível em: <https://repositorio.ufba.br/bitstream/ri/29351/1/REDES%20NEURAIS%20ARTIFICIAIS%20APLICADAS%20%C3%80%20CLASSIFICA%C3%87%C3%83O%20DE%20GESTOS.pdf>.

FUTURES, M. H. *Project InnerEye – Democratizing Medical Imaging AI*. [S.l.: s.n.], jul. 2023. Disponível em: <https://www.microsoft.com/en-us/research/project/medical-image-analysis/>.

GALA, A. S. *Profissionais de tecnologia e acessibilidade: quais são eles?* [S.l.: s.n.], mai. 2023. Disponível em: <https://www.handtalk.me/br/blog/profissionais-de-tecnologia-e-acessibilidade/>.

GENÇAY, R.; QI, M. Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. *Neural Networks, IEEE Transactions on*, v. 12, p. 726–734, ago. 2001. DOI: 10.1109/72.935086.

GOETZ, H. S. CVLabel : uma plataforma online para rotulação de imagens. *Ufrgs.br*, 2022. DOI: http://hdl.handle.net/10183/252539. Disponível em: <https://lume.ufrgs.br/handle/10183/252539>.

GONZALEZ, R. C.; WOODS, R. E. Digital Image Processing. *Pearson Education*, Pearson International Edition, 2008. Disponível em: <https://dl.ebooksworld.ir/motoman/Digital.Image.Processing.3rd.Edition.www.EBooksWorld.ir.pdf>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>.

HANDTALK. *Somos a maior plataforma de tradução automática para Línguas de Sinais do mundo*. 2023. Disponível em: <https://www.handtalk.me/br/blog/tecnologia-assistiva-surdos/#:~:text=O%20acesso%20%C3%A0%20tecnologia%20assistiva,Libras%20em%20shows%20e%20eventos>. (accessed: 16.08.2023).

HANEL, M. S. Análise de métodos de data augmentation para melhoria do desempenho de uma rede neural de detecção de defeitos em superfícies metálicas, 2021. Disponível em: <https://lume.ufrgs.br/handle/10183/235162>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. [S.l.]: Springer New York, 2009. DOI: https://doi.org/10.1007-978-0-387-84858-7. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-84858-7>.

HOWSE, J. *OpenCV Computer Vision with Python*. [S.l.: s.n.], abr. 2013. Disponível em: <https://www.academia.edu/36437176/OpenCV_Computer_Vision_with_Python>.

KARAMI, A.; ZANJ, B.; SARKALEH, A. K. Persian sign language (PSL) recognition using wavelet transform and neural networks. *Expert Systems with Applications*, v. 38, n. 3, p. 2661–2667, 2011. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2010.08.056. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417410008523>.

KARNOPP, L. *FONÉTICA E FONOLOGIA*. [S.l.: s.n.], 2015. Disponível em: <https://www.libras.ufsc.br/colecaoLetrasLibras/eixoFormacaoBasica/foneticaEFonologia/assets/359/FoneticaFonologia_TextoBase.pdf>.

LE, X. H. et al. *Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting*. [S.l.]: MDPI, jul. 2019. Disponível em: <https://www.researchgate.net/publication/334268507_Application_of_Long_Short-Term_Memory_LSTM_Neural_Network_for_Flood_Forecasting>.

MAGNO, R. *As Dificuldades da Pessoa Surda na Sociedade Brasileira | Jusbrasil*. [S.l.: s.n.], 2021. Disponível em: <https://www.jusbrasil.com.br/artigos/as-dificuldades-da-pessoa-surda-na-sociedade-brasileira/1176514129>.

MARQUES, E. A. L. *Estudo sobre Redes Neurais de Aprendizado Profundo com Aplicações em Classificação de Imagens*. 2016. Disponível em: <https://bdm.unb.br/bitstream/10483/15147/1/2016_EduardaAlmeidaLeaoMarques.pdf>. (accessed: 17.08.2023).

MEDIAPIPE. *Hand landmark detection guide*. 2023a. Disponível em: <https://developers.google.com/mediapipe/solutions/vision/hand_landmarker>. (accessed: 07.08.2023).

MEDIAPIPE. *Pose landmark detection guide*. 2023b. Disponível em: <https://developers.google.com/mediapipe/solutions/vision/pose_landmarker#get_started>. (accessed: 07.08.2023).

QU, X.; QIN, M.; YIN, Y. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, Elsevier BV, v. 125, p. 41–55, mai. 2020. DOI: https://doi.org/10.1016/j.neunet.2020.01.030. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S089360802030040X>.

QUADROS, R.; KARNOPP, L. *Língua de sinais brasileira: estudos lingüísticos*. [S.l.]: Artmed, 2007. ISBN 9788536303086. DOI: https://doi.org/9788536303086. Disponível em: <https://bds.unb.br/handle/123456789/948>.

RAWAT, W.; WANG, Z. *Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review*. [S.l.: s.n.], 2017. DOI: 10.1162/neco_a_00990.

SATAPATHY, S. C. et al. proceedings of the second international conference on computer and communication technologies. *SpringerLink*, Springer India, 2015. DOI: https://doi.org/10.1007-978-81-322-2526-3. Disponível em: <https://link.springer.com/book/10.1007/978-81-322-2526-3>.

SOUZA, L. R. DE. *Algoritmo para Reconhecimento e Acompanhamento de Trajetória de Padrões em Imagens Móveis*. [S.l.: s.n.], 2010. Disponível em: <http://www.univasf.edu.br/~brauliro.leal/tcc/LenivaldoRSouza/LenivaldoRSouza.pdf>.

SZELISKI, R. In: COMPUTER Vision: Algorithms and Applications. Cham: Springer International Publishing, 2011. ISBN 978-1-84882-934-3. DOI: 10.1007/978-1-84882-935-0. Disponível em: <https://doi.org/10.1007/978-1-84882-935-0>.

VANINI, E. *Intérpretes de Libras viralizam e ganham os palcos do Brasil em shows*. 2023. Disponível em: <https://oglobo.globo.com/ela/gente/noticia/2023/01/interpretes-de-libras-viralizam-e-ganham-os-palcos-do-brasil-em-shows.ghtml>. (accessed: 16.08.2023).

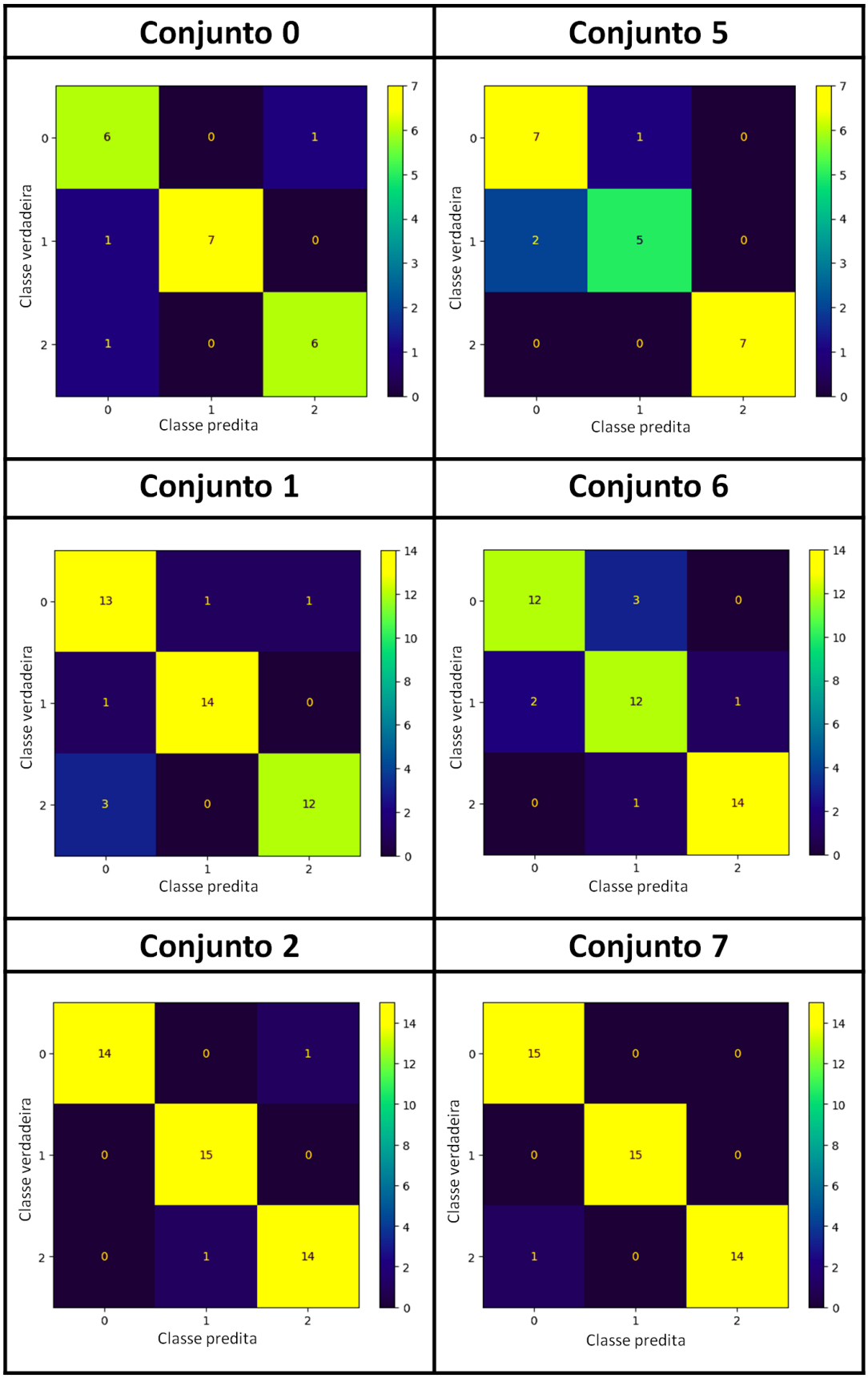
XIAO, Q.; QIN, M.; YIN, Y. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural Networks*, v. 125, p. 41–55, 2020. ISSN 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2020.01.030. Disponível em: <https://www.sciencedirect.com/science/article/pii/S089360802030040X>.

Apêndices

Apêndice A - RESULTADOS

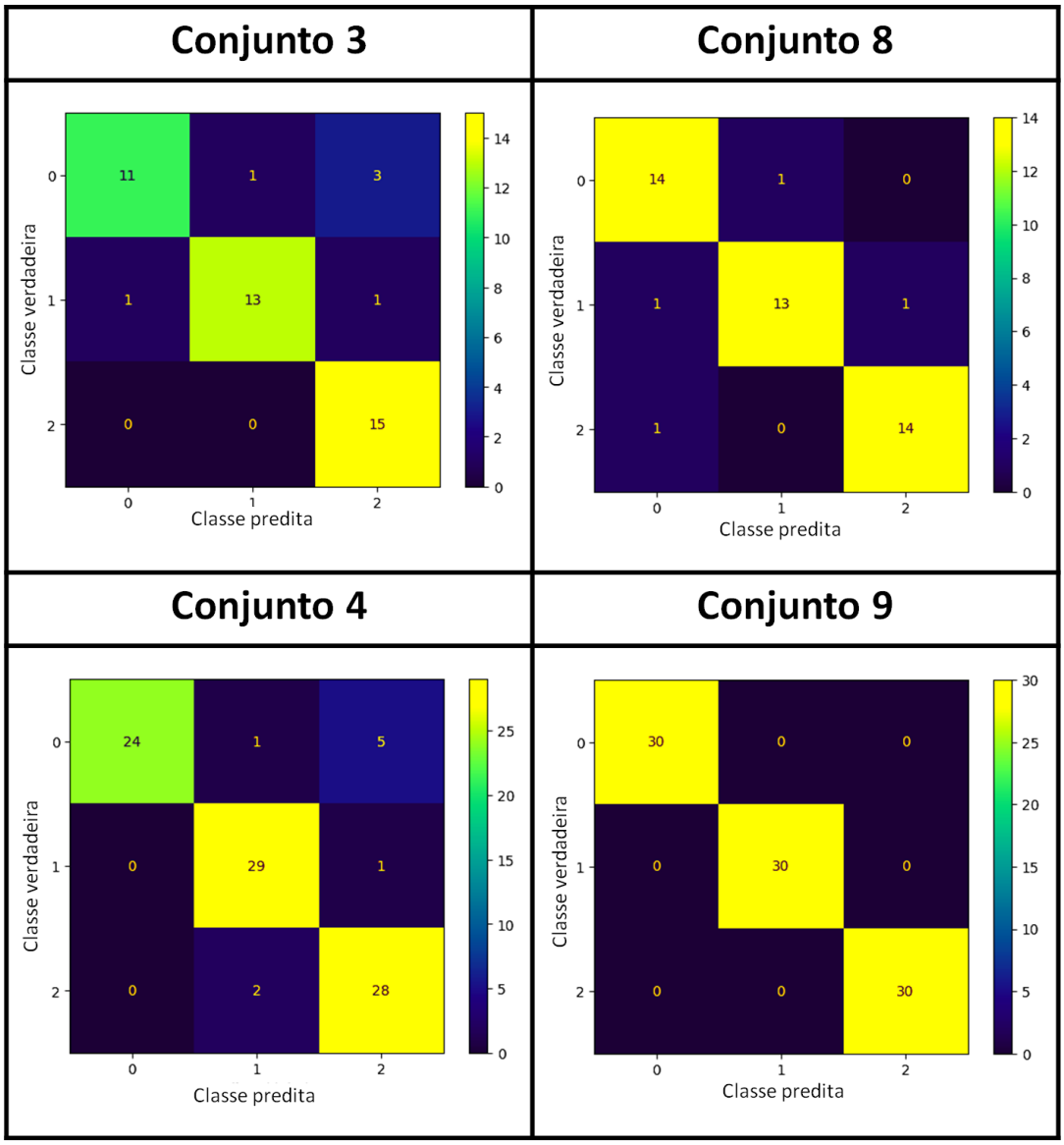
As Figuras 15 e 16 apresentam as matrizes de confusão (onde 0 se refere ao sinal de 'Brincar', 1 a 'Casa' e 2 a 'Eu te amo') dos treinamentos e estão dispostas de forma a poder-se comparar de forma mais simples a diferença os treinos com os conjunto de dados contendo dez ou cinco indivíduos. As Figuras 17 e 18 apresentam os gráficos de perda e acurácia ao longo dos treinamentos e validações.

Figura 15: *Matriz de confusão do conjunto de teste (15%) de cada um dos dez conjuntos.*



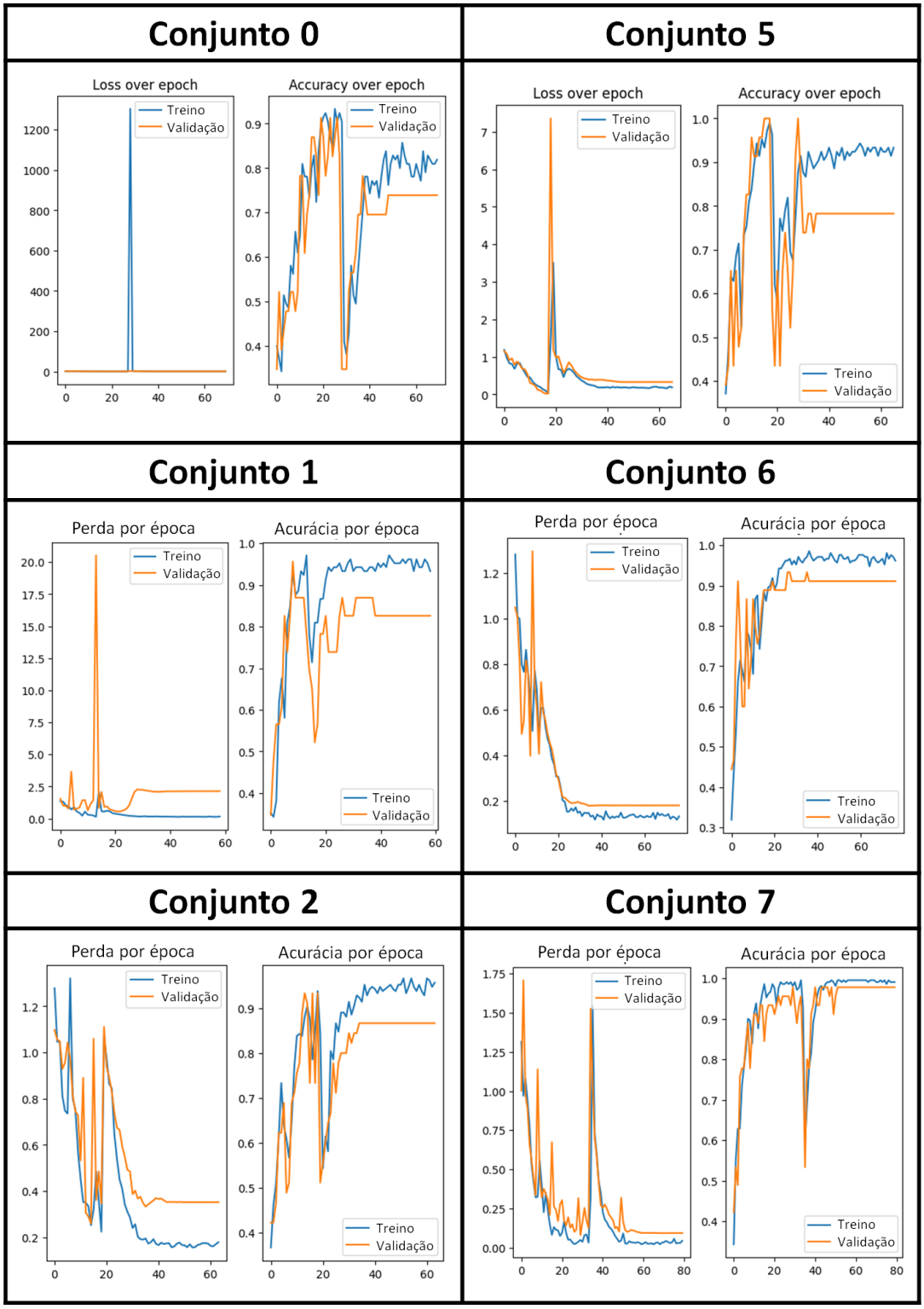
Fonte: Autor

Figura 16: *Matriz de confusão do conjunto de teste (15%) de cada um dos dez conjuntos.*



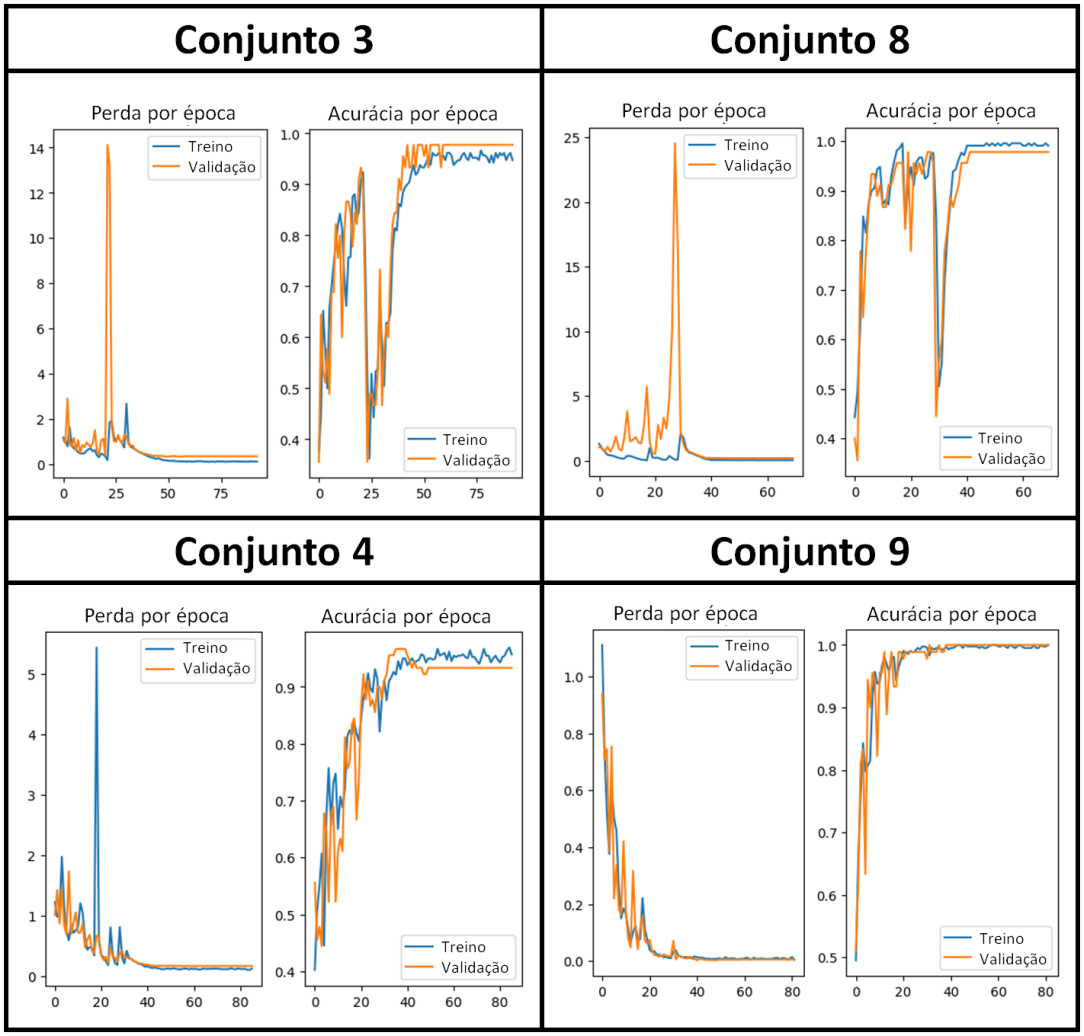
Fonte: Autor

Figura 17: *Gráficos de perda e acurácia, no treinamento e na validação.*



Fonte: Autor

Figura 18: *Gráficos de perda e acurácia, no treinamento e na validação.*



Fonte: Autor