INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E
TECNOLOGIA DE SÃO PAULO
CAMPUS VOTUPORANGA

HEITOR CÂMARA COSTA FERNANDES

APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA TRADUÇÃO AUTOMÁTICA DE LIBRAS: DESENVOLVIMENTO E ANÁLISE PRELIMINAR DE RESULTADOS

VOTUPORANGA
2025

Heitor Câmara Costa Fernandes

APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA TRADUÇÃO AUTOMÁTICA DE LIBRAS: DESENVOLVIMENTO E ANÁLISE PRELIMINAR DE RESULTADOS

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso em Bacharelado em Sistemas de Informação do Instituto Federal de Educação, Ciência e Tecnologia, Câmpus Votuporanga.

Professor Orientador: Cecílio Merlotti Rodas.

Votuporanga

2025

Heitor Câmara Costa Fernandes

APLICAÇÃO DE REDES NEURAIS CONVOLUCIONAIS PARA TRADUÇÃO AUTOMÁTICA DE LIBRAS: DESENVOLVIMENTO E ANÁLISE PRELIMINAR DE RESULTADOS

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Bacharelado em Sistemas de Informação do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Campus Votuporanga.

Professor Orientador: Cecílio Merlotti Rodas.

Aprovado pela banca examinadora em xx de mês de 20XX.

BANCA EXAMINADORA:

_____

Prof. Dr. Cicrano da Silva (para feminino use Dra.)

_____

Prof. Me. Beltrano dos Santos (para feminino use Mª.)

_____

Prof. Esp. José Luis Brasil

FICHA CATALOGRÁFICA

...

...

...

...

...

...

...

...

AGRADECIMENTOS

Gostaria de agradecer primeiramente a todos os servidores do Instituto Federal do campus Votuporanga, especialmente àqueles que estiveram diretamente ligados à minha formação acadêmica. Aos professores em sua totalidade, que sempre deixam uma parte de si no ensino e nos ensinamentos – parte essa que cada aluno carregará para seu futuro, seguindo ou não na área. Agradeço nominalmente ao Professor Cecílio Merlotti, ao Professor Dr. Evandro Jardini, ao Professor Eduardo e ao Professor Ivan, que estiveram diretamente envolvidos no desenvolvimento deste trabalho.

Gostaria de agradecer imensamente à minha família, que sempre forneceu apoio em todas as minhas decisões e em todas as suas etapas. Sem eles, nada disso seria remotamente possível – a eles devo tudo. Agradeço também às amizades que me acompanharam até aqui, oferecendo suporte e companheirismo ao longo desta jornada.

Dedico ainda meu reconhecimento à comunidade mundial de software livre. Defendo que a cultura, o conhecimento e uma melhor qualidade de vida não deveriam ser privilégios apenas daqueles que podem comprá-los – todos deveriam ter esses direitos assegurados. A comunidade de software livre se propõe a garantir isso a uma grande parcela da população, abrindo mão muitas vezes do ganho monetário e dedicando seu trabalho ao acesso irrestrito à informação, à cultura e à qualidade de vida.

Foi essa filosofia que me inspirou a licenciar este trabalho sob uma licença de software livre. O sistema resultante não está perfeito e está longe de estar, mas ainda assim tenho confiança de que, conforme for me aprimorando como profissional e como pessoa, darei continuidade ao seu melhoramento – e talvez inspire outras pessoas a fazerem o mesmo. Se o sistema resultante deste Trabalho de Conclusão de Curso melhorar, nem que seja minimamente, a vida de uma pessoa, ou inspirar outros a seguirem este caminho, meu papel aqui estará cumprido e estarei feliz.

"A alma é tingida pela cor de seus pensamentos"

Marco Aurélio

RESUMO

O desenvolvimento do projeto envolve seis etapas fundamentais: (1) a coleta de dados, com o registro dos gestos do alfabeto da Libras; (2) a extração de parâmetros relevantes com a biblioteca *MediaPipe*, que realiza o rastreamento dos movimentos das mãos e articulações; (3) a normalização e o pré-processamento dos dados, para garantir consistência e padronização; (4) o treinamento de modelos de aprendizado de máquina baseados em *convolutional neural networks* (*CNNs*) para a classificação dos gestos; (5) a implementação de um sistema funcional capaz de operar em tempo real; e (6) a realização de testes e avaliação de desempenho.

A proposta se fundamenta em experiências prévias com interfaces baseadas em *hand-tracking*, evidenciando o potencial dessa tecnologia para soluções práticas e inclusivas. Os resultados esperados incluem a entrega de um sistema operacional capaz de realizar a tradução em tempo real das letras do alfabeto da Libras (A–Z), com o código-fonte disponibilizado sob licença de código aberto, permitindo o uso, a modificação e a contribuição da comunidade.

Pode-se afirmar que a integração entre visão computacional e inteligência artificial apresenta o potencial de uma solução inovadora, com capacidade de contribuir significativamente para a redução de barreiras comunicacionais e para a ampliação da inclusão de pessoas surdas em ambientes educacionais, profissionais e cotidianos.

Palavras-chave: reconhecimento de língua de sinais; *hand-tracking*; processamento de linguagem natural; inteligência artificial; interface homem-máquina.

ABSTRACT

The development of the project involves six fundamental stages: (1) data collection, through the recording of Brazilian Sign Language (Libras) alphabet gestures; (2) extraction of relevant parameters using the MediaPipe library, which performs real-time tracking of hand and joint movements; (3) data normalization and preprocessing, to ensure consistency and standardization; (4) training of machine learning models based on convolutional neural networks (CNNs) for gesture classification; (5) implementation of a functional system capable of operating in real time; and (6) testing and performance evaluation.

The proposal is based on previous experiences with interfaces that use hand-tracking, highlighting the potential of this technology for practical and inclusive solutions. The expected outcomes include the delivery of an operational system capable of performing real-time translation of the Libras alphabet (A–Z), with its source code made available under an open-source license, allowing the community to use, modify, and contribute to the tool.

It can be stated that the integration of computer vision and artificial intelligence presents the potential for an innovative solution, with the ability to significantly contribute to reducing communication barriers and expanding the inclusion of deaf individuals in educational, professional, and everyday environments.

Keywords: sign language recognition; hand-tracking; natural language processing; artificial intelligence; human-computer interaction.

Índice de figuras

Índice de tabelas

Sumário

1 INTRODUÇÃO

A comunicação exerce um papel central no desenvolvimento pessoal, social e profissional dos indivíduos, sendo um fator determinante para assegurar a inclusão em diferentes esferas da sociedade. No entanto, pessoas surdas ainda enfrentam obstáculos significativos ao interagirem em ambientes majoritariamente compostos por ouvintes, o que compromete sua plena participação e limita suas oportunidades. Nesse contexto, a Língua Brasileira de Sinais (Libras) representa um importante recurso de expressão visual-gestual amplamente utilizado pela comunidade surda, permitindo o estabelecimento de relações sociais e o acesso à informação de forma eficaz.

Apesar de seu valor cultural e comunicacional, a Libras ainda encontra barreiras, principalmente devido à ausência de soluções tecnológicas acessíveis que possibilitem sua tradução em tempo real para texto. Essa lacuna impacta diretamente a autonomia das pessoas surdas, especialmente em contextos educacionais, profissionais e de convívio cotidiano, acentuando processos de exclusão social.

O presente trabalho propõe o desenvolvimento de um software de código aberto que utiliza técnicas de visão computacional e *machine learning* para traduzir, em tempo real, os gestos da Libras em texto. É importante destacar que o sistema terá como escopo exclusivo a tradução das letras do alfabeto manual da Libras (de A a Z), não abrangendo, nesta etapa, a totalidade da língua de sinais. Além disso, a saída gerada será exibida apenas em formato textual, sem síntese de voz.

A metodologia adotada envolve o uso da biblioteca *MediaPipe*, que permite a detecção precisa dos movimentos das mãos por meio de uma câmera, e a aplicação de modelos baseados em *convolutional neural networks* (*CNNs*) para a classificação dos gestos capturados. O sistema desenvolvido terá seu código-fonte disponibilizado de forma aberta, com o objetivo de fomentar adaptações e melhorias contínuas por parte da comunidade.

Com essa proposta, busca-se demonstrar que é possível empregar ferramentas como visão computacional e aprendizado de máquina no desenvolvimento de tecnologias acessíveis, voltadas à inclusão comunicacional. O sistema tem o potencial de ampliar a autonomia das pessoas surdas, melhorar sua qualidade de vida e promover maior integração social. Ao facilitar a interação entre surdos e ouvintes, a iniciativa contribui diretamente para a redução das barreiras linguísticas e para o fortalecimento de práticas mais equitativas em diferentes contextos da vida cotidiana.

2 OBJETIVOS

2.1 OBJETIVOS GERAIS

Desenvolver um software de código aberto que utilize técnicas de visão computacional e aprendizado de máquina para traduzir, em tempo real, os gestos da Língua Brasileira de Sinais (Libras) em texto, com o propósito de promover a acessibilidade e a inclusão social. O sistema tem como escopo a tradução exclusiva do alfabeto manual da Libras (letras de A a Z), não abrangendo, nesta etapa, a totalidade da língua de sinais. Essa delimitação visa garantir viabilidade técnica, controle experimental e clareza quanto aos objetivos do projeto.

2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste trabalho envolvem, inicialmente, a coleta e estruturação de um conjunto de dados contendo gestos correspondentes às letras do alfabeto da Língua Brasileira de Sinais (Libras), que servirá de base para o treinamento dos modelos de aprendizado de máquina. Em seguida, pretende-se implementar algoritmos de detecção e rastreamento dos movimentos das mãos utilizando a biblioteca *MediaPipe*, permitindo a extração dos dados visuais necessários para o reconhecimento dos sinais. A etapa seguinte consiste no desenvolvimento de modelos baseados em *convolutional neural networks* (*CNNs*), com o intuito de classificar, de forma precisa e eficiente, os gestos representativos das letras de A a Z. Será desenvolvida uma interface gráfica direta e acessível, destinada à visualização em tempo real das traduções realizadas pelo sistema. O desempenho da ferramenta será avaliado com o objetivo de verificar sua robustez e aplicabilidade. Por fim, o código-fonte do sistema será disponibilizado sob uma licença de código aberto, permitindo que a comunidade acadêmica e de desenvolvedores possa adaptar, expandir e contribuir com melhorias contínuas à solução proposta.

3 REVISÃO DE LITERATURA/REFERENCIAL TEÓRICO

3.1 O PAPEL DA TECNOLOGIA DA INFORMAÇÃO NA INCLUSÃO DA COMUNIDADE SURDA

A busca por uma sociedade mais equitativa e inclusiva impulsiona o desenvolvimento de soluções que visam mitigar barreiras e promover a plena participação de todos os cidadãos. Nesse contexto, a Tecnologia da Informação (TI) emerge como um catalisador fundamental, especialmente no que tange à acessibilidade para a comunidade surda. Ao oferecer ferramentas que mediam a comunicação entre surdos e ouvintes, a TI não apenas facilita interações cotidianas, mas também fortalece a autonomia, a identidade cultural e o exercício da cidadania. Este capítulo estabelece o alicerce social e tecnológico que justifica a presente pesquisa, contextualizando-a no campo da Tecnologia Assistiva e analisando o ecossistema de inovações digitais voltadas para a Língua Brasileira de Sinais (Libras).

3.1.1 TECNOLOGIA ASSISTIVA (TA) COMO FERRAMENTA DE AUTONOMIA E CIDADANIA

A Tecnologia Assistiva (TA) é formalmente definida como uma área de conhecimento interdisciplinar que abrange produtos, recursos, metodologias, estratégias, práticas e serviços. O seu objetivo principal é promover a funcionalidade, relacionada à atividade e participação, de pessoas com deficiência, incapacidades ou mobilidade reduzida, visando sua autonomia, independência, qualidade de vida e inclusão social. É crucial desmistificar a noção de que a TA se restringe a artefatos digitais; o termo engloba desde soluções simples, como rampas de acesso, até sistemas computacionais complexos.

No Brasil, o acesso à Tecnologia Assistiva é um direito assegurado pela Lei Brasileira de Inclusão da Pessoa com Deficiência (LBI), Lei nº 13.146/2015, que a reconhece como um elemento essencial para a superação de barreiras e para a promoção da igualdade de oportunidades. A relevância dessa legislação se torna evidente ao considerarmos a demografia da comunidade surda no país. Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) indicam que a população com algum grau de surdez ultrapassa 2,3 milhões de pessoas. Para uma parcela significativa dessa comunidade, a Libras é a língua materna, sendo que, globalmente, a Federação Mundial dos Surdos estima que 80% das pessoas surdas não são fluentes em línguas escritas, dependendo primariamente das línguas de sinais para a comunicação.

Esse cenário demográfico e linguístico sublinha a necessidade crítica de tecnologias que operem na modalidade visual-gestual. As TAs para surdos são, portanto, ferramentas de acessibilidades projetadas para facilitar o cotidiano, promovendo autonomia em atividades que, de outra forma, dependeriam da audição ou da interação com ouvintes não fluentes em Libras. Exemplos variam desde alertas visuais ou vibratórios para campainhas e alarmes de incêndio até complexos softwares de tradução, todos com o objetivo comum de quebrar as barreiras de acessibilidade que permeiam a sociedade.

3.1.2 A TECNOLOGIA DA INFORMAÇÃO COMO PONTE COMUNICACIONAL NO CENÁRIO BRASILEIRO

As Tecnologias da Informação e Comunicação (TICs) tornaram-se o principal vetor para o desenvolvimento de TAs modernas, ampliando exponencialmente as possibilidades de inclusão educacional e social para a comunidade surda. O avanço tecnológico reflete não apenas um progresso técnico, mas um reconhecimento crescente da diversidade linguística e cultural, promovendo maior acessibilidade e autonomia. No Brasil, diversas iniciativas de software destacam-se por criar pontes comunicacionais entre o português e a Libras, posicionando o país como um polo relevante no desenvolvimento de soluções de acessibilidade digital.

Uma das iniciativas mais proeminentes é a Suíte VLibras, um conjunto de ferramentas de código aberto e gratuito desenvolvido em uma parceria estratégica entre o governo federal — através do Ministério da Gestão e Inovação em Serviços Públicos (MGISP) e do Ministério dos Direitos Humanos e da Cidadania (MDHC) — e a academia, representada pela Universidade Federal da Paraíba (UFPB). O objetivo do VLibras é traduzir conteúdos digitais (texto, áudio e vídeo) do português para a Libras, utilizando um avatar 3D para realizar os sinais. Com ampla adoção em websites governamentais e corporativos, além de aplicativos para desktops e dispositivos móveis, o VLibras representa um marco na política pública de acessibilidade digital no Brasil.

Em paralelo, no setor privado, a plataforma Hand Talk consolidou-se como uma solução líder de mercado. Atuando há mais de uma década, a ferramenta já traduziu bilhões de palavras e se destaca por sua interface amigável, protagonizada pelos avatares Hugo e Maya. O aplicativo móvel da Hand Talk não só oferece tradução, mas também funciona como uma

ferramenta educacional, com dicionários e módulos de aprendizado, fomentando a disseminação da Libras entre ouvintes.

Além dessas duas grandes plataformas, outras ferramentas como o ProDeaf e o Rybená também contribuem para este ecossistema, oferecendo tradutores automáticos que podem ser integrados a páginas web e Ambientes Virtuais de Aprendizagem (AVAs), facilitando o acesso à informação e favorecendo o aprendizado do aluno surdo.

A análise desse cenário revela uma tendência clara e um avanço significativo na tradução na direção *português para Libras*. Ferramentas baseadas em avatares, como VLibras e Hand Talk, cumprem o papel crucial de tornar o vasto conteúdo digital em português acessível à comunidade surda. Contudo, essa é apenas uma via da ponte comunicacional. Permanece uma lacuna tecnológica significativa no sentido inverso: a tradução em tempo real de *Libras para texto ou voz*. Este desafio, consideravelmente mais complexo do ponto de vista técnico, é fundamental para garantir a autonomia expressiva da pessoa surda em interações com ouvintes não sinalizadores. O presente trabalho de conclusão de curso se insere exatamente nesta fronteira, buscando desenvolver uma solução que contribua para a construção da outra metade desta ponte, focando no Reconhecimento de Língua de Sinais (*Sign Language Recognition* – SLR).

3.2 A COMPLEXIDADE DA LIBRAS COMO DESAFIO COMPUTACIONAL

Para desenvolver um sistema de reconhecimento automático eficaz, é imperativo primeiro compreender a estrutura do objeto a ser reconhecido. A Libras, como qualquer língua natural, não é um mero conjunto de gestos, mas um sistema linguístico complexo e altamente estruturado. Sua modalidade visual-espacial, em contraste com a oral-auditiva das línguas faladas, impõe um conjunto único e formidável de desafios para a modelagem computacional. A tradução automática não é, portanto, um simples problema de classificação de imagens, mas uma tarefa que reside na interseção da visão computacional, do aprendizado de máquina e da linguística.

3.2.1 ESTRUTURA LINGUÍSTICA MULTIDIMENSIONAL: OS CINCO PARÂMETROS

A linguística, como ciência dedicada à investigação dos princípios que regem as línguas naturais, estabelece que a Libras possui a mesma complexidade estrutural de qualquer língua oral, com léxico e gramática próprios. Conforme estabelecido por Quadros e Karnopp (2004), a unidade fundamental da Libras, o sinal, é formada pela combinação de cinco parâmetros básicos que funcionam de maneira análoga aos fonemas nas línguas orais. São eles:

1. **Configuração de Mão (CM):** A forma que a mão assume durante a produção do sinal.

2. **Movimento (M):** A trajetória, a velocidade e a repetição do movimento das mãos.

3. **Locação (L):** O local no corpo ou no espaço à frente do sinalizador onde o sinal é executado.

4. **Orientação da Palma (O):** A direção para a qual a palma da mão aponta (para cima, para baixo, para o corpo, etc.).

5. **Expressões Não Manuais (ENM):** Movimentos da face, dos olhos, da cabeça e do tronco que acompanham o sinal manual.

O principal desafio computacional decorre da simultaneidade com que esses parâmetros são articulados. Enquanto os fonemas de uma língua oral se organizam de forma linear e sequencial (e.g., /k/-/a/-/z/-/a/), os parâmetros da Libras são produzidos, em grande parte, ao mesmo tempo para formar um único sinal. Nenhum parâmetro isolado possui significado completo; o sentido emerge da composição integrada. Para um sistema de visão computacional, isso significa que não basta reconhecer uma forma de mão estática; é preciso capturar e interpretar simultaneamente a forma, sua localização, sua orientação, seu movimento e as expressões faciais que a acompanham, um fluxo de dados multidimensional e de alta complexidade.

3.2.2 A RELEVÂNCIA GRAMATICAL DOS SINAIS NÃO MANUAIS (NMFS)

As Expressões Não Manuais, frequentemente referidas na literatura internacional como *Non-Manual Features* (NMFs), merecem um destaque particular, pois representam uma das maiores barreiras para o reconhecimento automático completo da língua. É um erro comum supor que essas expressões sirvam apenas para adicionar um tom emocional à comunicação. Na realidade, os NMFs são um componente gramaticalizado e essencial da Libras e de outras línguas de sinais. Sua função é comparável à entonação, ao ritmo e à prosódia nas línguas faladas, sendo capazes de alterar fundamentalmente o significado de uma sentença.

A importância gramatical dos NMFs se manifesta de várias formas:

- **Funções Lexicais:** Podem ser o único elemento distintivo entre dois sinais manuais idênticos. Um exemplo clássico da *American Sign Language* (ASL) é o sinal para "NOT YET" (ainda não), que exige um movimento específico da língua e da cabeça. Sem esses NMFs, o mesmo sinal manual seria interpretado como "LATE" (atrasado).
- **Funções Frasais:** Muitos marcadores gramaticais são puramente não manuais. Em ASL e outras línguas, a transformação de uma frase afirmativa em uma pergunta do tipo “sim/não” é realizada pelo levantar das sobrancelhas e uma leve inclinação da cabeça para a frente. A negação, a topicalização e a construção de orações condicionais também dependem fortemente de marcadores faciais e posturais.
- **Funções Discursivas:** O contato visual e os movimentos da cabeça são cruciais para a regulação da troca de turnos em um diálogo, funcionando como marcadores de coesão e coerência textual.

A captura e interpretação desses sinais sutis e simultâneos representam uma fronteira de pesquisa em visão computacional. A necessidade de rastrear não apenas as mãos, mas também múltiplos pontos de referência no rosto e no tronco, e de modelar a sincronia entre eles, aumenta exponencialmente a complexidade do problema. Diante disso, a delimitação do escopo deste trabalho ao alfabeto manual da Libras (datilologia) constitui uma abordagem metodológica estratégica. Os sinais do alfabeto são primariamente manuais, com uma dependência muito menor de NMFs gramaticais complexos, permitindo isolar e focar no desafio do reconhecimento da configuração e do movimento das mãos, que é, por si só, um problema fundamental e não trivial.

3.2.3 O DESAFIO DO FLUXO CONTÍNUO: COARTICULAÇÃO E SEGMENTAÇÃO EM CSLR

O objetivo final da pesquisa em SLR é o Reconhecimento Contínuo de Língua de Sinais (*Continuous Sign Language Recognition* - CSLR), ou seja, a capacidade de traduzir frases e discursos completos, assim como eles ocorrem na comunicação natural. No entanto, a transição do reconhecimento de sinais isolados (

Isolated Sign Language Recognition - ISLR) para o CSLR introduz dois fenômenos interligados que representam alguns dos maiores desafios da área: a coarticulação e a segmentação.

A coarticulação é o processo linguístico no qual a morfologia de um sinal é afetada pelos sinais vizinhos (o anterior e o posterior). Na prática, isso significa que os sinais não são produzidos de forma discreta e separada, mas se fundem em um fluxo contínuo de movimento. A forma da mão, a posição e a trajetória no final de um sinal e no início do seguinte podem ser significativamente alteradas para criar uma transição mais fluida. Como resultado, a aparência visual de um mesmo sinal pode variar drasticamente dependendo do seu contexto na frase, tornando seu reconhecimento muito mais difícil.

Diretamente ligado à coarticulação está o desafio da segmentação, que é a tarefa de identificar os pontos de início e fim de cada sinal individual dentro do fluxo contínuo. Como os sinais se misturam devido à coarticulação e aos movimentos de transição (epêntese), não existem pausas claras e consistentes entre eles. A segmentação manual de vídeos para a criação de datasets de treinamento é um processo laborioso, e a segmentação automática em tempo real é um problema de pesquisa ainda em aberto.

Esses fenômenos demonstram que o reconhecimento de uma frase em Libras é muito mais complexo do que simplesmente reconhecer uma sequência de sinais isolados. É necessário modelar dependências temporais e contextuais de longo alcance. Portanto, o foco em ISLR, como o adotado neste projeto para as letras do alfabeto, é um passo metodológico necessário. Dominar o reconhecimento de unidades isoladas e estáticas é a base sobre a qual sistemas mais complexos, capazes de lidar com a dinâmica da coarticulação e da segmentação, podem ser construídos.

3.3 EVOLUÇÃO E ANÁLISE CRÍTICA DAS ABORDAGENS DE RECONHECIMENTO AUTOMÁTICO

A jornada para a criação de sistemas de reconhecimento de língua de sinais é marcada por uma evolução tecnológica contínua, impulsionada pela busca de maior precisão, naturalidade e acessibilidade. Cada paradigma tecnológico, desde os primeiros dispositivos com sensores até as mais recentes técnicas de visão computacional, representou um avanço significativo, mas também revelou novas camadas de desafios. A análise crítica dessa trajetória não apenas contextualiza o estado da arte, mas também fornece uma justificativa robusta para as escolhas metodológicas adotadas neste trabalho, em particular a transição de uma abordagem baseada em imagens brutas para uma mais abstrata e resiliente, baseada em marcos esqueléticos.

3.3.1 ABORDAGENS INVASIVAS: SISTEMAS BASEADOS EM SENSORES (LUVAS DE DADOS)

Os primeiros esforços sistemáticos para o reconhecimento automático de língua de sinais, iniciados na década de 1980, recorreram a abordagens baseadas em hardware especializado, notadamente as luvas de dados (*data gloves*). Esses dispositivos são equipados com sensores, como sensores de flexão (

flex sensors) e unidades de medição inercial (IMUs), que capturam diretamente dados sobre a curvatura dos dedos, a orientação da mão e seu movimento no espaço 3D.

A principal vantagem dessa abordagem reside na sua alta precisão e na robustez dos dados coletados. Ao medir diretamente a geometria e a cinemática da mão, os sistemas baseados em luvas são imunes a problemas que afligem as abordagens visuais, como variações de iluminação, fundos complexos ou oclusões parciais. Os dados gerados são numéricos e de baixa dimensão, simplificando o processo de classificação subsequente.

No entanto, as desvantagens são significativas e limitam fundamentalmente a sua aplicabilidade prática e escalabilidade. Primeiramente, são sistemas **intrusivos**: a necessidade de vestir um equipamento afeta a naturalidade da sinalização e cria uma barreira para o uso cotidiano. Em segundo lugar, o custo do hardware é geralmente elevado e o equipamento pode ser volumoso, tornando a tecnologia inacessível para a maioria dos usuários e inadequada para implantação em larga escala. A meta de criar uma ferramenta de comunicação transparente e onipresente é, portanto, incompatível com a dependência de dispositivos especializados.

3.3.2 A TRANSIÇÃO PARA A NATURALIDADE: SISTEMAS VISUAIS (BASEADOS EM CÂMERA)

Em resposta às limitações dos sistemas baseados em sensores, a comunidade de pesquisa promoveu uma mudança de paradigma em direção a abordagens baseadas em visão computacional. Utilizando câmeras convencionais e de baixo custo — como as webcams presentes em notebooks e smartphones — como único sensor, esses sistemas buscam uma interação humano-computador muito mais natural e não intrusiva. Essa transição se alinha perfeitamente com o objetivo de desenvolver tecnologias assistivas que se integrem de forma transparente ao dia a dia dos usuários.

Essa nova abordagem, no entanto, transferiu a complexidade do hardware para o software. Em vez de receber dados limpos e estruturados de sensores, o sistema agora precisa extrair informações significativas a partir de um fluxo de dados brutos e ruidosos: os pixels de um vídeo. Isso introduziu um conjunto de desafios clássicos da visão computacional, como a necessidade de segmentar a mão do fundo, a sensibilidade a variações de iluminação, a oclusão da mão por objetos ou pela outra mão, e a variabilidade na aparência do sinalizador (e.g., tom de pele, roupas).

3.3.3 LIMITAÇÕES DAS ABORDAGENS VISUAIS DIRETAS E O PROBLEMA DO "RUÍDO SEMÂNTICO"

Uma estratégia comum dentro do paradigma visual é aplicar modelos de aprendizado profundo, especialmente Redes Neurais Convolucionais (CNNs), diretamente sobre as imagens dos sinais. As CNNs são extremamente poderosas para aprender características hierárquicas a partir de dados brutos, tendo revolucionado o campo do reconhecimento de imagens. A premissa é que, ao ser treinada com um grande volume de imagens de sinais, a rede aprenderia autonomamente a identificar os padrões visuais distintivos de cada um.

Contudo, a aplicação prática dessa abordagem direta revela uma falha crítica, que se manifesta como um severo sobreajuste (*overfitting*) a características contextuais irrelevantes. Conforme detalhado na seção de metodologia deste trabalho, a experiência inicial com um modelo de CNN treinado em imagens resultou em um sistema que, embora apresentasse alta acurácia nos dados de validação, falhava drasticamente em generalizar para novas amostras em tempo real. A análise indicou que o modelo não estava aprendendo o conceito geométrico abstrato do gesto, mas sim "memorizando" características espúrias do conjunto de

treinamento, como a iluminação específica do ambiente, o padrão do fundo, e até mesmo particularidades idiossincráticas da mão do usuário.

Este fenômeno, que pode ser denominado “ruído semântico”, é um desafio bem documentado na literatura. Sistemas de reconhecimento baseados em características de cor, por exemplo, são inerentemente frágeis a variações de iluminação e a fundos complexos. Modelos treinados em datasets com pouca diversidade podem falhar ao serem expostos a tons de pele não vistos durante o treinamento, ou a diferentes condições de fundo em cenários do mundo real. A alta capacidade das CNNs, se não for devidamente regularizada ou guiada, pode levá-las a aprender correlações estatísticas que são fortes nos dados de treino, mas que não têm qualquer relevância para a definição linguística do sinal.

3.3.4 A SOLUÇÃO ROBUSTA: RECONHECIMENTO BASEADO EM ESQUELETO (SKELETON-BASED)

Para superar as limitações tanto das abordagens invasivas quanto das visuais diretas, emergiu uma terceira via que combina o melhor de ambos os mundos: o reconhecimento baseado em esqueleto (*skeleton-based*). Esta abordagem mantém a naturalidade não intrusiva dos sistemas baseados em câmera, mas introduz uma camada de abstração que filtra o ruído semântico e foca na informação essencial do gesto.

O processo consiste em duas etapas: primeiro, um modelo de visão computacional pré-treinado é utilizado para detectar e extrair as coordenadas espaciais (2D ou 3D) de um conjunto de pontos de referência anatômicos chave (*landmarks*) — por exemplo, as articulações dos dedos, os pulsos, e pontos no rosto. Em seguida, em vez de alimentar a imagem inteira ao modelo de classificação, apenas este vetor de coordenadas é utilizado como entrada.

A principal vantagem desta metodologia é a **invariância contextual**. Ao operar exclusivamente sobre a geometria esquelética, o modelo de classificação nunca é exposto à iluminação, ao fundo, ao tom de pele ou a outras variáveis visuais. Ele é forçado a aprender a distinguir os sinais com base unicamente naquilo que os define linguisticamente: a forma, a pose e as relações espaciais entre as articulações. Isso resulta em sistemas com uma adaptabilidade muito maior a fundos complexos e circunstâncias dinâmicas, resolvendo diretamente o problema de sobreajuste que afligiu a abordagem anterior.

Além da robustez, esta abordagem oferece uma drástica redução de dimensionalidade. A entrada do modelo de classificação passa de uma matriz com dezenas de milhares de pixels para um vetor com algumas dezenas de coordenadas. Isso não apenas reduz massivamente o custo computacional do treinamento e da inferência, mas também mitiga o risco de sobreajuste, conforme o princípio da Navalha de Ockham.

A viabilidade prática desta abordagem foi enormemente impulsionada por bibliotecas de código aberto como o MediaPipe do Google, que oferecem modelos de detecção de *landmarks* de mão, face e corpo de alta performance, capazes de operar em tempo real em hardware de consumo. A disponibilidade dessas ferramentas democratizou o acesso à tecnologia de extração de esqueleto, tornando-a a escolha metodológica preferencial para muitos sistemas de SLR modernos, incluindo o desenvolvido neste trabalho.

A tabela a seguir sintetiza a análise comparativa das três principais abordagens de SLR, evidenciando a superioridade da metodologia baseada em esqueleto para o desenvolvimento de sistemas robustos e acessíveis.

Característica	**Abordagem Baseada em Sensores (Luvas)**	**Abordagem Baseada em Imagem (Pixels)**	**Abordagem Baseada em Esqueleto (Landmarks)**
Precisão de Captura	Alta	Média-Alta	Alta (depende do detector de landmarks)
Robustez a Variações Ambientais	Muito Alta	Baixa	Muito Alta
Custo e Acessibilidade	Alto (hardware específico)	Baixo (webcam padrão)	Baixo (webcam padrão)
Nível de Intrusão/Naturalidade	Alto (invasivo)	Nulo (não invasivo)	Nulo (não invasivo)
Complexidade Computacional	Baixa	Muito Alta (CNNs profundas)	Baixa-Média (MLPs, LSTMs)
Risco de Overfitting Contextual	Baixo	Muito Alto	Muito Baixo

3.4 MODELAGEM DE APRENDIZADO DE MÁQUINA PARA DADOS ESQUELÉTICOS

Uma vez estabelecida a abordagem baseada em esqueleto como a mais robusta e eficiente para a extração de características em SLR, a etapa seguinte consiste em selecionar e treinar um modelo de aprendizado de máquina apropriado para classificar esses dados. A natureza dos dados esqueléticos — vetores numéricos estruturados que representam a geometria e a dinâmica dos gestos — abre um leque de possibilidades de modelagem, desde classificadores simples para gestos estáticos até arquiteturas sequenciais complexas para o reconhecimento contínuo.

3.4.1 CLASSIFICAÇÃO DE GESTOS ESTÁTICOS COM MODELOS PARA DADOS ESTRUTURADOS

Para a tarefa de reconhecer sinais estáticos e isolados, como as letras do alfabeto manual da Libras, a abordagem baseada em esqueleto transforma fundamentalmente a natureza do problema. O que antes era uma tarefa de classificação de imagens de alta dimensão torna-se um problema de classificação de dados tabulares ou estruturados. Cada instância de um gesto é representada por um vetor de tamanho fixo contendo as coordenadas x,y,z de todos os *landmarks* detectados.

Nesse cenário, modelos computacionalmente intensivos como as CNNs profundas, projetadas para extrair características de dados brutos como pixels, não são mais necessários e podem até ser subótimos. Em vez disso, modelos mais leves e adequados para dados vetoriais podem ser empregados com grande eficácia. O **Perceptron de Múltiplas Camadas (MLP)**, uma arquitetura de rede neural totalmente conectada, é uma escolha natural e eficiente para essa tarefa. Conforme implementado neste trabalho, um MLP com poucas camadas ocultas é capaz de aprender as relações não-lineares complexas entre as coordenadas das articulações, mapeando a geometria da mão à sua respectiva classe (letra). A validade dessa escolha é corroborada pela drástica redução no tempo de treinamento — de horas para segundos — e pela eliminação do sobreajuste visual, demonstrando que o MLP é um modelo com a capacidade adequada para a complexidade dos dados de entrada, sem o excesso de capacidade que levava ao sobreajuste no modelo baseado em imagem.

3.4.2 O HORIZONTE DA PESQUISA: MODELOS SEQUENCIAIS PARA SINAIS DINÂMICOS

Embora o MLP seja suficiente para gestos estáticos, a modelagem de sinais dinâmicos ou o reconhecimento contínuo (CSLR) exige a captura de informações temporais. Nesses casos, não é apenas a pose da mão em um instante que importa, mas a evolução dessa pose ao longo do tempo. A entrada do modelo deixa de ser um único vetor e passa a ser uma sequência de vetores esqueléticos, um para cada quadro (*frame*) do vídeo. Para lidar com essa dimensão temporal, a pesquisa em SLR tem se voltado para modelos sequenciais, muitos dos quais foram originalmente desenvolvidos para o Processamento de Linguagem Natural (PLN).

As **Redes Neurais Recorrentes (RNNs)** e, mais especificamente, as redes de **Memória de Longo e Curto Prazo (LSTMs)**, foram as primeiras arquiteturas de aprendizado profundo a serem amplamente aplicadas para modelar sequências temporais em SLR. As LSTMs são projetadas para manter um estado de "memória" que lhes permite capturar dependências ao longo do tempo, tornando-as adequadas para entender a trajetória e a dinâmica de um sinal. Arquiteturas híbridas, que combinam modelos espaciais como as Redes Neurais de Grafos (GCNs) para processar a estrutura do esqueleto em cada

frame com LSTMs para modelar a evolução temporal, têm demonstrado resultados promissores. No entanto, as LSTMs podem ter dificuldades em capturar dependências de muito longo prazo devido ao problema do desvanecimento do gradiente (

vanishing gradient) e seu processamento inerentemente sequencial limita a paralelização e a eficiência computacional.

Mais recentemente, a arquitetura **Transformer**, que revolucionou o PLN, tem sido adaptada para o SLR com sucesso notável. Os Transformers dispensam a recorrência e, em vez disso, utilizam um mecanismo de

autoatenção (*self-attention*) para ponderar a importância de todos os *frames* em uma sequência ao processar cada *frame* individual. Isso permite que o modelo capture eficientemente tanto dependências locais (movimentos rápidos) quanto globais (a estrutura geral de um sinal longo). Modelos baseados em Transformers estão alcançando o estado da arte em diversos

benchmarks de SLR, muitas vezes com arquiteturas mais leves (menos parâmetros) do que abordagens anteriores, demonstrando alta eficácia e eficiência.

A evolução da modelagem de dados esqueléticos — de MLPs para LSTMs e agora para Transformers — reflete uma maturação do campo. A transição para uma representação esquelética permitiu que os pesquisadores de SLR se afastassem de problemas puramente de visão computacional e se aproximassem de problemas de modelagem de sequências. Isso viabilizou a importação e adaptação das arquiteturas mais poderosas do PLN, tratando a língua de sinais não como uma série de imagens, mas como aquilo que ela de fato é: uma linguagem com estrutura espaço-temporal complexa. O presente trabalho, ao dominar a classificação de gestos estáticos com um modelo eficiente, estabelece uma base sólida e alinhada com as melhores práticas, ao mesmo tempo em que reconhece e aponta para o horizonte da pesquisa, onde modelos sequenciais avançados prometem finalmente decifrar a complexidade do discurso contínuo em Libras.

4 METODOLOGIA

A construção de um sistema de reconhecimento de padrões eficaz é um processo iterativo, que exige reavaliações à medida que os desafios práticos emergem. A metodologia empregada neste trabalho reflete essa dinâmica, tendo evoluído de uma abordagem inicial baseada em imagem para uma solução mais eficiente, centrada em dados geométricos. Esta seção introdutória detalha a justificativa para essa transição, delineando os obstáculos encontrados e a fundamentação teórica que suporta a nova abordagem.

Inicialmente, o projeto foi concebido sob o paradigma do reconhecimento visual direto, utilizando Redes Neurais Convolucionais (CNN's). Essa escolha se alinhava com a abordagem consolidada por trabalhos seminais como o de LeCun *et al.* (1998), que demonstraram a capacidade excepcional das CNN's de aprender características hierárquicas diretamente de dados brutos, como *pixels* de uma imagem, para tarefas complexas de classificação. A premissa era que, ao alimentar o modelo com um vasto conjunto de imagens dos gestos da Libras, a rede aprenderia autonomamente a extrair os padrões visuais distintivos de cada letra.

Contudo, a implementação prática desta abordagem revelou dois desafios críticos e interligados, que comprometeram a viabilidade do sistema. O primeiro foi o custo computacional proibitivo. O treinamento de uma CNN com dezenas de milhares de imagens de alta resolução é uma tarefa intrinsecamente lenta e que demanda recursos significativos, com ciclos de treinamento que se estendiam por mais de 15 horas. Essa morosidade tornava o processo de experimentação e ajuste de hiperparâmetros impraticável. O segundo e mais fundamental problema foi a manifestação de um severo sobreajuste (*overfitting*). O modelo exibia uma *performance* quase perfeita nos dados de treinamento e validação, mas falhava drasticamente em generalizar para novas amostras em tempo real. A análise desse comportamento indicou que a alta capacidade da CNN a levava a "memorizar" características espúrias e contextuais do conjunto de dados — como a iluminação específica do ambiente, o fundo da imagem e particularidades idiossincráticas da mão do usuário — em vez de aprender o conceito geométrico abstrato de cada gesto. Este é um desafio conhecido em sistemas de reconhecimento de língua de sinais, onde a variabilidade dos dados pode facilmente levar modelos a aprender correlações irrelevantes.

Diante desses obstáculos, foi realizado um pivô metodológico fundamental: a transição de uma abordagem baseada em imagem para uma baseada em marcos geométricos

(*landmarks*). Em vez de processar *pixels*, o sistema foi reestruturado para extrair e operar exclusivamente sobre as coordenadas 3D dos 21 pontos de referência da mão, detectados pela biblioteca *MediaPipe*. Essa mudança estratégica ataca a raiz dos problemas identificados. Primeiramente, ao trabalhar com vetores numéricos de baixa dimensão (63 coordenadas por amostra) armazenados em arquivos *CSV*, o tempo de treinamento foi reduzido de horas para meros segundos. Mais importante, essa abordagem elimina por completo a fonte do sobreajuste visual. O modelo nunca é exposto ao fundo, à iluminação ou à textura da pele; ele é forçado a aprender unicamente a partir da geometria esquelética da mão — a forma, a pose e as relações espaciais entre as articulações. Pesquisas na área corroboram essa estratégia, demonstrando que sistemas baseados em *landmarks* podem alcançar desempenho robusto, especialmente em cenários com fundos complexos, pois enviam uma representação de dados mais enxuta e focada para o módulo de reconhecimento.

Portanto, a metodologia detalhada nas seções subsequentes descreve um fluxo de trabalho otimizado e resiliente, que abrange desde a coleta e o pré-processamento de dados geométricos até o treinamento de um modelo de aprendizado de máquina mais simples e apropriado *(*um Perceptron de Múltiplas Camadas - MLP), culminando na implementação de um sistema de tradução em tempo real mais preciso e com uma melhor capacidade de generalização.

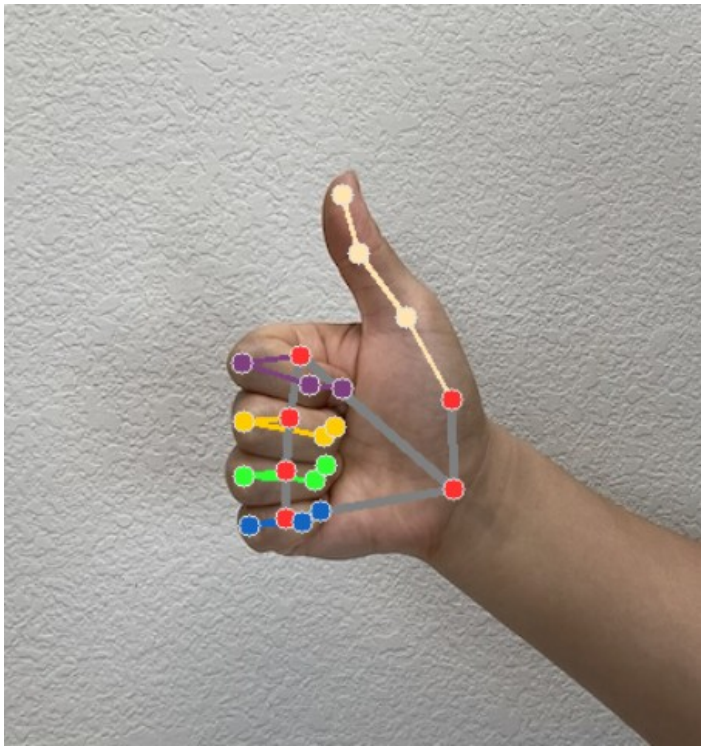
4.1 COLETA DE DADOS GEOMÉTRICOS

A etapa de coleta de dados constitui o alicerce sobre o qual a capacidade de generalização e a precisão do modelo são construídas. Conforme ressaltado por Bishop (2006), a forma como os dados são representados e pré-processados é um dos fatores mais determinantes para o sucesso de um sistema de reconhecimento de padrões. Em contraste direto com a abordagem anterior, que se baseava no armazenamento massivo de imagens, a metodologia revisada foca na criação de um conjunto de dados (*dataset*) que representa a essência geométrica de cada gesto, abstraindo-o de seu contexto visual. Este processo visa gerar um *dataset* leve, informacionalmente denso e otimizado para o treinamento de modelos de *machine learning* eficientes.

A ferramenta central para esta nova abordagem é a biblioteca MediaPipe, uma solução de código aberto do Google que oferece modelos pré-treinados de alta performance para tarefas de visão computacional. Especificamente, foi utilizado o módulo *Hand Landmarker*, que é capaz de detectar e rastrear 21 pontos de referência tridimensionais (*landmarks*) que mapeiam a topologia da mão em *real-time*. A eficácia do MediaPipe para extrair características relevantes para o reconhecimento de gestos já foi validada em diversos estudos, como o de Sundar e Bagyammal (2022), que empregaram a mesma estrutura para um sistema de reconhecimento de gestos manuais. A decisão de utilizar os marcos 3D (coordenadas x, y, z) é deliberada. Enquanto os marcos 2D (x, y) podem ser ambíguos e suscetíveis a variações de perspectiva, a inclusão da coordenada de profundidade (z) permite ao modelo compreender as relações espaciais entre as articulações de forma muito mais robusta. Conforme demonstrado por Cerlinca e Pentiuc (2010), o uso de informação de profundidade facilita um reconhecimento de mãos mais robusto em comparação com abordagens 2D.

O processo de captura opera da seguinte forma: o sistema inicializa a *webcam* e processa o fluxo de vídeo quadro a quadro. Para cada quadro, a imagem é convertida para o formato *RGB* e enviada ao modelo do MediaPipe. Se uma mão é detectada com um nível de confiança superior a um limiar pré-definido, o modelo retorna as coordenadas 3D para cada um dos 21 marcos. O sistema então extrai essas coordenadas e as organiza em um vetor numérico. Especificamente, para cada marco, as três coordenadas são coletadas, resultando em um vetor de 63 características (*features*) para cada captura. Este vetor é então “achatado” (*flattened*) e salvo como uma nova linha em um arquivo de valores separados por vírgula (*CSV*), cujo nome corresponde à letra que está sendo capturada. Esta abordagem de exportar coordenadas para um arquivo *CSV* é uma prática para criar *datasets* estruturados a partir de

dados visuais, como demonstrado em metodologias de reconhecimento de emoções faciais baseadas em *landmarks*.



A necessidade de um conjunto de dados volumoso e diversificado é um pilar do *machine learning*, especialmente em domínios como o reconhecimento de língua de sinais, onde a variabilidade entre usuários é alta e o risco de *overfitting* é uma preocupação constante, como apontado por Huamani *et al.* (2023). Um número elevado de amostras por classe permite que o modelo seja exposto a uma vasta gama de pequenas variações na execução de um mesmo gesto — como ligeiras diferenças de ângulo, rotação e posicionamento da mão. Essa variabilidade no treinamento é essencial para o desenvolvimento da capacidade de generalização do modelo, ou seja, sua habilidade de reconhecer corretamente um gesto mesmo que ele não seja idêntico aos exemplos vistos durante o treinamento. Ao final desta etapa, o resultado é um conjunto de dados composto por arquivos *CSV*, onde cada arquivo representa uma classe e cada linha representa uma instância única da geometria daquele gesto. Esta estrutura de dados é não apenas computacionalmente eficiente, mas também intrinsecamente resiliente ao *overfitting* visual que prejudicou a abordagem metodológica anterior.

4.2 EXTRAÇÃO DE PARÂMETROS

Na metodologia revisada, a etapa de extração de parâmetros é fundida ao processo de coleta de dados, representando uma mudança fundamental na forma como a informação é representada para o modelo de *machine learning*. Os parâmetros, ou características (*features*), que alimentarão o modelo não são mais os *pixels* de uma imagem, mas sim o vetor de 63 coordenadas que descreve a geometria da mão. Esta abordagem transita de um paradigma de aprendizado de representação, onde a própria rede neural é responsável por extrair *features* de dados brutos, para uma abordagem baseada em engenharia de características, onde o conhecimento do domínio é utilizado para fornecer ao modelo uma representação de dados já purificada e de relevância informacional.

Esta mudança oferece algumas vantagens. A primeira é uma drástica redução de dimensionalidade. Conforme Bishop (2006) discute, a "maldição da dimensionalidade" é um desafio central em reconhecimento de padrões, onde o volume do espaço de entrada cresce exponencialmente com o número de *features*. Ao substituir uma matriz de imagem (por exemplo, 224x224 *pixels*, totalizando 50.176 valores) por um vetor de apenas 63 coordenadas, a complexidade do problema é massivamente reduzida. Isso não apenas diminui a carga computacional, mas também mitiga o risco de *overfitting*, pois o modelo opera em um espaço de características muito mais conciso.

A segunda e mais crucial vantagem é a invariância a fatores contextuais. Conforme argumentado por Gil-Martín *et al.* (2023), uma abordagem baseada em *landmarks* tem a vantagem de enviar menos informação ao módulo de reconhecimento em comparação com abordagens tradicionais de visão computacional (coordenadas de *landmarks versus* uma imagem completa), alcançando um desempenho robusto mesmo em cenários com fundos complexos. Ao extrair apenas a geometria esquelética, o modelo torna-se inerentemente imune a variações de iluminação, cor de fundo, tom de pele e outros ruídos visuais que não possuem relevância semântica para o gesto em si. Essa robustez é um dos principais benefícios de se trabalhar com dados esqueléticos para o reconhecimento de ações, como destacado em diversos estudos na área. O modelo é forçado a aprender a classificar os gestos com base unicamente naquilo que os define: a forma e a configuração espacial das articulações da mão. Portanto, a saída desta etapa não é apenas um conjunto de dados, mas um conjunto de representações numéricas otimizadas, que encapsulam a informação essencial para a tarefa de classificação, estabelecendo uma base sólida para a fase subsequente de normalização e treinamento.

4.3 NORMALIZAÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

Com o conjunto de dados geométricos coletado, a etapa de pré-processamento assume um papel fundamental para garantir a integridade e a adequação dos dados ao ambiente de treinamento. Esta fase, que antecede a exposição dos dados ao modelo de *machine learning*, é executada por um processo automatizado que aplica um conjunto de transformações padronizadas com o objetivo de otimizar a performance e a estabilidade do aprendizado. O pré-processamento é composto por três passos sequenciais: a codificação dos rótulos, a divisão do conjunto de dados e o escalonamento das características.

O primeiro passo é a codificação de rótulos. Os rótulos das classes, inicialmente representados como *strings* (por exemplo, ‘A’, ‘B’, ‘C’), são incompatíveis com a entrada numérica exigida pela maioria dos algoritmos de aprendizado de máquina. Para resolver isso, utiliza-se um codificador que converte cada rótulo textual em um valor inteiro único, criando um mapeamento consistente entre o nome da classe e sua representação numérica (ex: ‘A’ → 0, ‘B’ → 1). Este mapeamento é salvo para ser utilizado posteriormente no processo de predição, permitindo a conversão do resultado numérico do modelo de volta para a letra correspondente.

O segundo passo é a divisão do conjunto de dados. Para realizar uma avaliação imparcial da capacidade de generalização do modelo, é imperativo que ele seja testado com dados que não foram utilizados durante o seu treinamento. Para este fim, o *dataset* completo é particionado em um conjunto de treinamento (80% dos dados) e um conjunto de teste (20% dos dados). Uma consideração crítica nesta etapa é o uso da divisão estratificada. Conforme discutido na documentação do *scikit-learn*, para tarefas de classificação, a estratificação garante que a divisão dos dados preserve a proporção de amostras para cada classe que existia no conjunto de dados original. Em um problema com 26 classes como este, uma divisão aleatória simples poderia, por acaso, resultar em um conjunto de teste com poucas ou nenhuma amostra de uma determinada letra, o que tornaria a avaliação do modelo enviesada e pouco confiável. A estratificação mitiga esse risco, assegurando que ambos os subconjuntos, de treino e de teste, sejam microcosmos fiéis da distribuição de dados original.

O terceiro e último passo é o escalonamento de características. Modelos de redes neurais, como o Perceptron de Múltiplas Camadas (MLP), são sensíveis à escala das características de entrada. Conforme a documentação do *scikit-learn*, algoritmos que utilizam otimização baseada em gradiente convergem muito mais rápido e de forma mais estável quando os dados estão em uma escala uniforme. No caso dos *landmarks* da mão, embora as coordenadas já sejam normalizadas pelo MediaPipe em relação à imagem, suas variâncias podem diferir. Para padronizá-las, aplica-se a técnica de estandardização, que transforma cada uma das 63 características para que tenham uma média de 0 e um desvio padrão de 1. É crucial que o objeto escalonador seja ajustado (*fit*) apenas com os dados de treinamento e, em seguida, a mesma transformação seja aplicada tanto ao conjunto de treinamento quanto ao de teste. Esta prática evita o "vazamento de dados" (*data leakage*) do conjunto de teste para o processo de treinamento, garantindo que a avaliação final do modelo permaneça completamente imparcial. Ao final desta etapa, o *dataset* encontra-se preparado: rotulado, dimensionalmente consistente, particionado de forma representativa e numericamente normalizado, pronto para ser utilizado na fase de treinamento do modelo.

4.4 TREINAMENTO DO MODELO

A etapa de treinamento do modelo representa o núcleo do processo de aprendizado, onde os dados pré-processados são utilizados para ajustar os parâmetros internos de um classificador a fim de que ele possa mapear os vetores de características geométricas às suas respectivas classes. A transição de uma abordagem baseada em imagem para uma baseada em *landmarks* permitiu a substituição da computacionalmente custosa Rede Neural Convolucional (*CNN*) por um modelo mais leve e apropriado para dados tabulares: o Perceptron de Múltiplas Camadas (MLP). Embora redes neurais profundas dominem domínios de dados não estruturados, as arquiteturas MLP, quando devidamente configuradas, podem ser extremamente eficazes para dados tabulares, rivalizando e por vezes superando modelos mais complexos como os baseados em árvores de decisão.

A arquitetura do MLP foi definida com duas camadas ocultas, contendo 128 e 64 neurônios, respectivamente. Esta estrutura oferece capacidade suficiente para aprender as relações não-lineares complexas entre as 63 coordenadas dos *landmarks*, representando um equilíbrio entre poder de representação e a necessidade de evitar o *overfitting* que assolou a abordagem anterior. O treinamento foi conduzido utilizando a implementação do *scikit-learn*, que emprega a função de ativação *ReLU* (*Rectified Linear Unit*) por padrão nas camadas ocultas. Conforme demonstrado por Glorot *et al.* (2011), a *ReLU* oferece vantagens significativas, como a mitigação do problema do "desvanecimento do gradiente" (*vanishing gradient*) e uma maior eficiência computacional em comparação com funções de ativação tradicionais como a sigmoide, o que acelera a convergência durante o treinamento.

Para otimizar os pesos do modelo, o MLP do *scikit-learn* utiliza o otimizador Adam (*Adaptive Moment Estimation*). Proposto por Kingma e Ba (2014), o Adam é um algoritmo de otimização estocástica baseado em gradiente que se destaca por sua eficiência e robustez. Ele calcula taxas de aprendizado adaptativas para cada parâmetro individualmente, mantendo estimativas de primeiro e segundo momentos dos gradientes. Essa abordagem combina as vantagens de outros otimizadores, como a capacidade de lidar com gradientes esparsos e de se adaptar a objetivos não-estacionários, tornando-o particularmente bem-sucedido em uma vasta gama de problemas de *deep learning*. O processo de treinamento foi configurado para ocorrer de forma iterativa ao longo de 100 épocas. Uma época representa uma passagem completa do algoritmo por todo o conjunto de dados de treinamento. A cada época, a função de perda do modelo era registrada, permitindo o monitoramento da curva de aprendizado e a verificação da convergência.

Após a conclusão do ciclo de treinamento, o modelo final é avaliado no conjunto de teste, que foi mantido isolado durante todo o processo para garantir uma avaliação imparcial. A métrica utilizada para a avaliação final é a acurácia, que mede a proporção de previsões corretas. Finalmente, três artefatos essenciais são serializados e salvos em disco para serem utilizados pelo sistema de predição em *real-time*: o modelo MLP treinado, o objeto escalonador ajustado nos dados de treinamento e o mapeamento das classes codificadas. Salvar o escalonador é um passo de importância crítica, pois garante que a mesma transformação de normalização aplicada aos dados de treinamento seja consistentemente aplicada aos novos dados durante a inferência, uma premissa fundamental para a correta operação do modelo.

4.5 IMPLEMENTAÇÃO E PREDIÇÃO EM TEMPO REAL

A etapa final da metodologia consiste na implementação do sistema de tradução em *real-time*, que integra os artefatos gerados durante o treinamento em uma aplicação. Este módulo é projetado para operar de forma contínua, capturando dados visuais da *webcam*, processando-os através do *pipeline* de reconhecimento e fornecendo um *feedback* visual imediato ao usuário. O objetivo é criar uma experiência de interação homem-computador (*HCI*) fluida e com baixa latência, cumprindo o propósito central do projeto.

O fluxo de operação em *real-time* inicia-se com o carregamento dos artefatos do modelo previamente salvos: o classificador *MLP* treinado, o objeto escalonador e o mapeamento das classes. Em seguida, o sistema entra em um laço de execução contínuo, onde cada iteração corresponde ao processamento de um único quadro (*frame*) de vídeo. Para cada quadro capturado, a biblioteca *MediaPipe* é novamente empregada para detectar a presença de uma mão e extrair as coordenadas 3D dos 21 *landmarks*. Este processo espelha exatamente a etapa de coleta de dados, garantindo a consistência na extração das características.

Uma vez que o vetor de 63 coordenadas é extraído, ele é submetido a um pré-processamento em *real-time* que é idêntico ao aplicado durante o treinamento. O vetor é transformado utilizando o objeto escalonador carregado, que aplica a mesma estandardização de média e variância. Este passo é de importância crítica. Conforme a documentação do *scikit-learn*, é indispensável que os dados apresentados ao modelo durante a inferência (predição) passem pela mesma transformação de pré-processamento que os dados de treinamento. A falha em aplicar a mesma escala resultaria em uma entrada com uma distribuição estatística diferente daquela que o modelo aprendeu a interpretar, levando a predições inconsistentes e imprecisas.

Com o vetor de *landmarks* devidamente normalizado, ele é então passado para o método de predição do modelo *MLP* carregado. O modelo processa a entrada e retorna uma predição na forma de um rótulo numérico. Este rótulo é, por fim, decodificado de volta para a letra correspondente do alfabeto, utilizando o mapeamento de classes salvo. O resultado final é exibido na tela sobre a imagem da *webcam*, completando o ciclo de tradução. Este processo se repete para cada quadro, permitindo que o sistema responda dinamicamente aos gestos do usuário, criando uma ferramenta de comunicação interativa.

5 CONSIDERAÇÕES FINAIS

6 REFERÊNCIAS

FONSECA, Fabiana Ferreira. VISÃO COMPUTACIONAL APLICADA AO RECONHECIMENTO DE IMAGENS RELACIONADAS À LÍNGUA BRASILEIRA DE SINAIS. Orientador: Eduardo A. B. da Silva/Gabriel Matos Araújo. 2020. 105 f. TCC (Graduação) - Curso de Engenharia Eletrônica e de Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2020. Disponível em: https://monografias.poli.ufrj.br/monografias/monopoli10031301.pdf. acesso em: 1 jun. 2025.

PASSOS, Rosana. PARÂMETROS FÍSICOS DO MOVIMENTO EM LIBRAS: UM ESTUDO SOBRE INTENSIFICADORES. Orientador: Profa. Dra. Thaïs Cristófaro Alves da Silva. 2014. 244 f. Tese (Doutorado) - Curso de Letras, UFMG, Belo Horizonte, 2014. Disponível em: https://repositorio.ufmg.br/bitstream/1843/RMSA-ALAGMG/1/1282d.pdf. Acesso em: 1 jun. 2025.

QUADROS, Ronice Müller de; KARNOPP, Lodenir Becker. Língua de sinais brasileira: Estudos lingüísticos. Porto Alegre: Artmed, 2004.

GOOGLE DEVELOPERS. **ai.google.dev.** Rastreamento de mãos com Hand Landmarker. *[S.l.]*. Google, 2025. Disponível em: https://ai.google.dev/edge/mediapipe/solutions/vision/gesture_recognizer/python?hl=pt-br. Acesso em: 3 jun. 2025.

LECUN, Yann *et al.* Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278-2324, 1998. Disponível em: http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf.

SRIVASTAVA, Nitish *et al.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, v. 15, p. 1929-1958, 2014. Disponível em: http://jmlr.org/papers/v15/srivastava14a.html.

7 GLOSSÁRIO

...

...

...

...

...

...

...

...