Reconhecimento Inteligente de Sinais em LIBRAS: Um Modelo Computacional Assistivo para Educação Inclusiva em Ambientes Educacionais

Anderson Silva, Universidade Federal do Maranhão,
anderson.silva@nca.ufma.br, https://orcid.org/0009-0003-7721-0700
Sabryna Araújo, Universidade Federal do Maranhão
sabryna.ra@discente.ufma.br, https://orcid.org/0009-0001-5460-1261
Marcos Farias, Universidade Federal do Maranhão,
marvinfar852@nca.ufma.br, https://orcid.org/0009-0006-6836-6342
Isaque Santos, Universidade Federal do Maranhão,
isaque@nca.ufma.br, https://orcid.org/0009-0004-3170-5529
Alexandre Pessoa, Universidade Federal do Maranhão,
alexandre.pessoa@nca.ufma.br, https://orcid.org/0000-0003-4995-8909
Geraldo Braz Junior, Universidade Federal do Maranhão,
geraldo.braz@ufma.br, https://orcid.org/0000-0003-3731-6431

Resumo: Este trabalho apresenta uma proposta de tecnologia assistiva baseada em aprendizado profundo para o reconhecimento automático de sinais da Língua Brasileira de Sinais (LIBRAS), com foco no apoio à inclusão educacional de usuários da LIBRAS. O modelo desenvolvido combina as arquiteturas *EfficientNet* e LSTM com um bloco de atenção temporal, permitindo capturar tanto as características espaciais quanto as dinâmicas dos gestos ao longo do tempo. Os resultados demonstraram desempenho expressivo, alcançando até 96,48% de *F1-Score*, evidenciando a robustez e estabilidade do modelo mesmo em sinais com alta similaridade visual. Além da contribuição técnica, o estudo discute a utilização da solução em sistemas inteligentes para fins educacionais, indicando o potencial de integração da tecnologia em plataformas de ensino, tradutores automáticos e ambientes de aprendizagem inclusivos.

Palavras-chave: Reconhecimento de Sinais, LIBRAS, Aprendizado Profundo, Tecnologia Assistiva, Educação Inclusiva.

Intelligent Recognition of LIBRAS Signs: A Computational Assistive Model for Inclusive Education in Learning Environments

Abstract: This work presents an assistive technology proposal based on deep learning for the automatic recognition of Brazilian Sign Language (LIBRAS) signs, focusing on supporting the educational inclusion of LIBRAS users. The developed model combines the EfficientNet and LSTM architectures with a temporal attention block, enabling the capture of both spatial and dynamic features of gestures over time. The results showed expressive performance, reaching up to 94.41% F1-Score, demonstrating the robustness and stability of the model even for visually similar signs. Beyond its technical contribution, this study represents a significant step forward in the use of intelligent systems for educational purposes, highlighting the potential of integrating this technology into learning platforms, automatic translators, and inclusive educational environments.

Keywords: Sign Recognition, LIBRAS, Deep Learning, Assistive Technology, Inclusive Education.

1. Introdução

A efetivação da educação inclusiva representa um dos principais desafios do sistema de ensino brasileiro. Neste cenário, a Língua Brasileira de Sinais (LIBRAS) atua como um pilar para a comunicação e inclusão de estudantes surdos. Embora

sua presença em instituições de ensino seja legalmente garantida pela Lei Brasileira de Inclusão (Lei nº 13.146/2015), a implementação prática enfrenta obstáculos, criando uma lacuna entre as políticas públicas e a realidade escolar, que demanda a superação de uma cultura de segregação que historicamente ignora especificidades linguísticas e culturais da comunidade surda (Azevedo e Alencar, 2021). Uma das barreiras mais persistentes é a comunicação entre surdos e ouvintes, de forma ainda mais acentuada, em ambientes virtuais de aprendizagem. Neste contexto, a tecnologia assistiva, impulsionada pela Inteligência Artificial (IA), surge como uma via promissora para mitigar essa barreiras de comunicação e promover uma interação mais equitativa (Santos e Oliveira, 2024).

A justificativa para o desenvolvimento de soluções tecnológicas se ancora na carência de ferramentas automatizadas e acessíveis que ofereçam suporte à interação em LIBRAS (Damasceno *et al.*, 2025). Embora já existam iniciativas no Brasil, como aplicativos de tradução da Língua Portuguesa para LIBRAS, estudos apontam que essas soluções ainda enfrentam desafios significativos em termos de fidelidade, usabilidade e adaptação ao contexto educacional (Sakis e Bernardi, 2020). Nos últimos anos, os avanços em visão computacional têm possibilitado o desenvolvimento de sistemas capazes de reconhecimento e tradução de sinais visuais automaticamente, configurando-se como um recurso promissor de apoio à inclusão de estudantes surdos em ambientes educacionais (Stefano *et al.*, 2021).

Nesse contexto, a proposta central deste trabalho é o desenvolvimento de um modelo de aprendizado profundo para o reconhecimento automático de sinais em LIBRAS, visando validar por meio de experimentos uma arquitetura computacional que sirva como base para futuras ferramentas de apoio à educação inclusiva. Para tal, o modelo emprega uma arquitetura híbrida que combina Rede Neural Convolucional (CNN), Rede Neural Recorrente e blocos de atenção temporal, abordagens já consolidadas em outras tarefas de reconhecimento visual em sequências de vídeo (Yang *et al.*, 2023; Stefano *et al.*, 2021). A combinação dessas técnicas é projetada para capturar com eficácia tanto as características espaciais de cada sinal quanto a dinâmica temporal da gesticulação para o reconhecimento preciso dos sinais.

Para alcançar os objetivos propostos, este artigo está estruturado da seguinte maneira. A Seção 2 apresenta a fundamentação teórica, enquanto a Seção 3 apresenta os trabalhos relacionados. A Seção 4 detalha a proposta metodológica, incluindo o protocolo experimental. A Seção 5 expõe e discute os resultados obtidos. Por fim, a Seção 6 apresenta as considerações finais e sugestões para trabalhos futuros.

2. Fundamentação Teórica

2.1. Redes Neurais Convolucionais

A visão computacional é um campo interdisciplinar que busca capacitar computadores a realizar tarefas relacionadas à percepção visual humana (Caprì *et al.*, 2024). Nesse contexto, o Reconhecimento de Comportamento Humano (RCH) por vídeo constitui uma tarefa central, com aplicações em áreas como segurança, saúde e educação (Wang *et al.*, 2022; Vanneste *et al.*, 2021). Para esse tipo de problema, as redes neurais convolucionais são amplamente empregadas devido à sua capacidade de extrair automaticamente características espaciais relevantes a partir de imagens, por meio de uma hierarquia de camadas convolucionais (Krichen, 2023).

Uma operação de convolução é o processo de deslizar um *kernel* sobre uma imagem, multiplicando o filtro pelos valores de pixel nas posições correspondentes e então somando-os para obter um mapa de características. Neste trabalho, utiliza-se a arquitetura *EfficientNet* (Tan e Le, 2019), apresentada na Figura 1, por seu equilíbrio entre

desempenho e custo computacional. Sua estrutura é composta por blocos convolucionais MBConv (Sandler *et al.*, 2018) que reduzem o custo computacional das camadas convolucionais com o mecanismo *Squeeze and Excitation* (SE) (Hu; Shen e Sun, 2018). Esse mecanismo permite à rede enfatizar características mais informativas da imagem, contribuindo para uma representação espacial mais discriminativa dos sinais de LIBRAS.

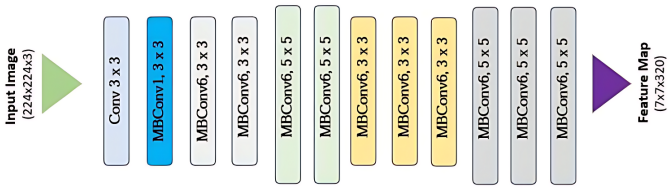

Figura 1. Arquitetura *EfficientNet* (Hisaria *et al.*, 2024)

2.2. *Long-Short Therm Memory* (LSTM)

De modo a eficientemente capturar as dependências temporais presentes nos dados, modelos recorrentes são comumente utilizados para lidar com tarefas envolvendo temporalidade. A arquitetura LSTM (Hochreiter e Schmidhuber, 1997) busca resolver o problema da explosão/desaparecimento de gradientes que tipicamente aparece ao se lidar com longas dependências (Houdt; Mosquera e Nápoles, 2020). Nesse sentido, o problema de RCH está profundamente relacionado a este tópico, visto que um algoritmo de predição recebe as características do comportamento humano e determina qual comportamento ocorrerá a seguir (Dunne *et al.*, 2023). Em consonância, a arquitetura LSTM apresentada na Figura 2, é composta por um conjunto de blocos de memória conectados recorrentemente. Essas estruturas são capazes de controlar o fluxo das informações passadas adiante pela rede, persistindo as características de maior relevância ao passo que permitem o esquecimento daquelas consideradas pouco relevantes.

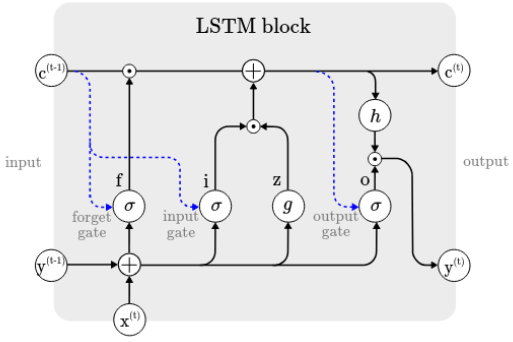

Figura 2. Bloco LSTM (Houdt; Mosquera e Nápoles, 2020)

2.3. Mecanismo de Atenção

Redes *Transformer*, propostas primeiramente por (Vaswani *et al.*, 2017), geraram uma enorme adoção pela comunidade de aprendizado de máquina recentemente. Essa arquitetura, baseada em blocos de atenção, é capaz superar o estado da arte em diversas frentes de pesquisa, especialmente o processamento de linguagem natural. Consequentemente, outras propostas baseadas em mecanismos de atenção começaram a ser adaptadas para os mais diferentes domínios, incluindo tarefas envolvendo imagens.

Mecanismos baseados em atenção são motivados pela noção de que podemos

induzir o modelo a focar em áreas específicas de toda a sequência de entrada (Fernando *et al.*, 2018). Esse comportamento permite que os modelos foquem em atributos importantes não só para uma única instância de entrada, mas para todo o contexto dos dados de treinamento. A função de atenção é matematicamente definida como um produto vetorial entre três matrizes nomeadas *Query*, *Key* e *Value*, as quais codificam representações intermediárias dos dados de entrada. A saída é computada como uma soma ponderada dos *Values*, onde o peso atribuído a cada valor é computado por uma função de compatibilidade das *Queries* com suas *Keys* correspondentes. O bloco de atenção presente na Figura 3 demonstra a visão geral desta arquitetura.

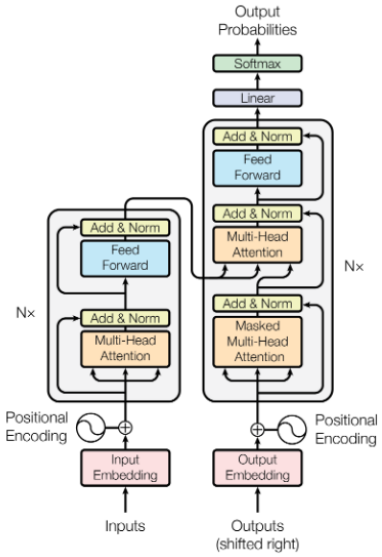

Figura 3. Bloco Transformer (Vaswani *et al.*, 2017)

3. Trabalhos Relacionados

Devido ao crescente interesse no uso de Inteligência Artificial como apoio à inclusão educacional, diversos estudos têm investigado como soluções computacionais podem auxiliar na redução de barreiras linguísticas e sensoriais em ambientes de ensino. Essa produção tem destacado que tecnologias baseadas em IA podem mediar comunicação, apoiar processos de aprendizagem e ampliar a participação de estudantes em contextos digitais e híbridos, especialmente quando há restrições de acesso à linguagem oral ou escrita (Gibson, 2024; Melo-López *et al.*, 2025).

O estudo de Corrêa et al. (2014) examina o uso de aplicativos de tradução entre Língua Portuguesa e LIBRAS (ProDeaf e HandTalk) como recurso de mediação comunicativa entre surdos e ouvintes. Os autores destacam que tais ferramentas podem atenuar barreiras em contextos formativos ao facilitar o diálogo, sobretudo em situações onde não há intérprete disponível, contribuindo para maior autonomia comunicacional de estudantes surdos. Também ressalta-se a importância dessas tecnologias na consolidação de práticas inclusivas mediadas por recursos digitais.

A pesquisa de Li et al. (2020) apresenta um conjunto de vídeos em larga escala para reconhecimento lexical da língua de sinais americana, reunindo sinais realizados por múltiplos sinalizadores em diferentes contextos. O estudo descreve detalhadamente o processo de coleta e anotação dos vídeos, destacando a importância da diversidade de sinalizadores e vocabulário para o desenvolvimento de modelos mais generalizáveis. Além disso, descreve o processo de comparação entre diversas arquiteturas, incluindo

redes convolucionais bidimensionais, redes recorrentes e modelos tridimensionais de convolução temporal, demonstrando que a ampliação da diversidade de sinalizador e vocabulários favorece o desempenho em tarefas de reconhecimento em vídeo.

O trabalho de Cerna et al. (2021) apresenta a base de dados LIBRAS-UFOP e realiza experimentos com redes neurais para o reconhecimento de pares mínimos. Em linguística, pares mínimos são sinais que diferem apenas em um dos parâmetros da língua - como movimento, ponto de articulação, configuração de mão ou expressão facial —, sendo fundamentais para distinguir significados próximos na LIBRAS. Os autores demonstram que modelos de visão computacional conseguem distinguir sinais visualmente semelhantes com consistência, além de disponibilizar uma base anotada que se tornou referência para estudos que investigam estratégias de reconhecimento automático em LIBRAS.

Já o trabalho de Benevides et al. (2024) propõe uma tecnologia assistiva capaz de reconhecer, em tempo real, o alfabeto datilológico de LIBRAS por meio de visão computacional. A solução, desenvolvida com base em um conjunto de 13 mil imagens estáticas, demonstra resultados expressivos na tradução de configurações manuais correspondentes às letras do alfabeto. Entretanto, o sistema concentra-se apenas no reconhecimento de sinais estáticos, voltados à soletração, o que limita sua aplicação em contextos mais complexos de comunicação.

Observa-se que os trabalhos concentram-se majoritariamente no desempenho de modelos e na estruturação de bases de dados, com menor atenção às implicações educacionais de seu uso. Ademais, o presente trabalho aborda o reconhecimento de sinais que envolvem movimento e articulação ao longo do tempo, aspecto importante para a construção de sistemas educacionais mais completos e inclusivos. Diante desse cenário, o presente estudo desenvolve e avalia um modelo de reconhecimento de sinais em LIBRAS, buscando contribuir ao aproximar resultados de visão computacional do campo da inclusão educacional, em direção a futuras aplicações assistivas em contextos reais de ensino.

4. Método Proposto

Nesta Seção são apresentadas as etapas do método empregadas no desenvolvimento da solução proposta. São descritos a base de dados utilizada, os procedimentos de pré-processamento das imagens, a estrutura do modelo desenvolvido e o protocolo de avaliação do método. O objetivo é detalhar de forma clara e reprodutível o processo experimental que fundamenta os resultados obtidos. A Figura 4 apresenta o fluxo do método proposto.

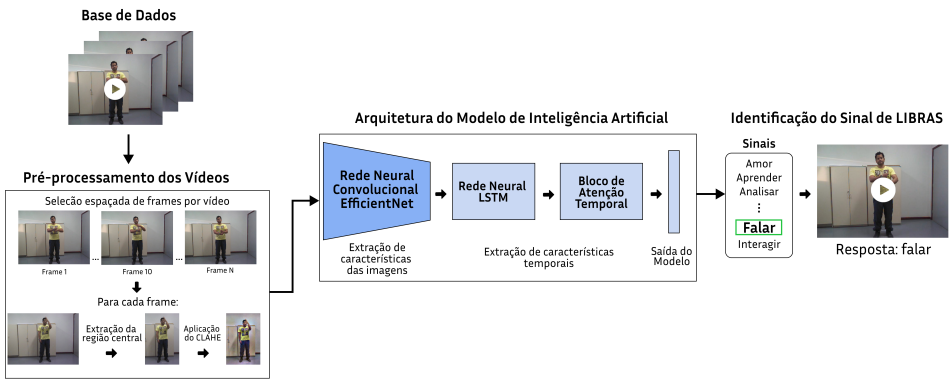

Figura 4. Fluxo do método proposto para identificar sinais de LIBRAS

4.1. Base de Dados

A base de dados utilizada neste trabalho é o *LIBRAS-UFOP Dataset*, um conjunto público desenvolvido pela Universidade Federal de Ouro Preto (Cerna *et al.*, 2021). Essa base contém informações multimodais - de cor (RGB), profundidade e esqueleto para cada vídeo. Ao todo, reúne 56 sinais distintos da LIBRAS, organizados em quatro categorias baseadas no conceito de pares mínimos — sinais que diferem apenas em um dos parâmetros fundamentais da língua, sendo eles configuração de mão, ponto de articulação, movimento, orientação da palma ou expressão facial. Essa estrutura torna o conjunto particularmente desafiador, pois inclui gestos com alta similaridade visual e semântica.

As gravações foram realizadas por cinco participantes (três mulheres e dois homens), cada um executando os sinais em média dez vezes, totalizando aproximadamente 3.040 vídeos. As filmagens foram feitas sob diferentes condições de iluminação e fundo. Além disso, todas as amostras foram revisadas e validadas por um especialista em LIBRAS, assegurando a precisão dos rótulos e a representatividade linguística dos gestos para viabilizar estudos de desenvolvimento de reconhecimento automático dos sinais.

4.2. Pré-processamento dos Vídeos

O pré-processamento consiste na preparação adequada dos dados a serem utilizados para treinamento e avaliação da solução proposta. Como cada sinal em LIBRAS possui durações diferentes, seguindo a proposta de Alves et al. (2024), definiu-se um número fixo de *frames* a serem utilizados por vídeo, coletados de forma espaçada ao longo da gravação. Essa estratégia garante que diferentes momentos do gesto sejam representados, capturando tanto o início quanto o fim do movimento. Quando o vídeo possui menos *frames* do que o número estabelecido, aplica-se a técnica de *padding*, que consiste em duplicar o último quadro até atingir a quantidade desejada. Essa padronização é feita visando analisar os sinais de forma equilibrada e coerente, aprendendo os padrões temporais reduzindo a influência da variação de duração entre os vídeos.

Em seguida, foi realizada a extração da região de interesse, que corresponde à área central onde ocorrem os movimentos das mãos. Essa etapa reduz partes do fundo da imagem em que o gesticulador não aparece, concentrando o aprendizado na ação gestual. Por fim, como proposto por Musa et al. (2018), utilizou-se a técnica CLAHE (*Contrast Limited Adaptive Histogram Equalization*), responsável por ajustar o contraste local das imagens para destacar detalhes sutis, como o formato dos dedos e a orientação das mãos. Esse realce de contraste é especialmente importante em vídeos com variação de iluminação, pois ajuda o modelo a identificar melhor os gestos mesmo em condições visuais desfavoráveis.

4.3. Modelo de Reconhecimento de Sinais

4.3.1. Arquitetura de Redes Neurais Convolucionais *EfficientNet*

O modelo utilizado para extração das características das imagens é baseado na arquitetura da *EfficientNet* (Tan e Le, 2019), uma rede neural convolucional amplamente empregada em tarefas de visão computacional. Esse tipo de rede é capaz de identificar automaticamente padrões visuais, como formas, texturas e contornos, auxiliando na diferenciação dos sinais de LIBRAS entre si. A *EfficientNet* foi escolhida por combinar alto desempenho com baixo custo computacional, o que permite sua aplicação em ambientes educacionais e sistemas assistivos com recursos limitados.

Além disso, a *EfficientNet* foi previamente treinada em um grande conjunto de

imagens, chamado *ImageNet* (Fei-Fei; Deng e Li, 2009), o que permite ao modelo aproveitar conhecimentos visuais já adquiridos e adaptá-los ao reconhecimento de sinais em LIBRAS. Essa técnica, chamada de *transfer learning* (Weiss; Khoshgoftaar e Wang, 2016), reduz o tempo de convergência treinamento e contribui para melhorias na tarefa proposta. Assim, a rede atua como um extrator de características espaciais, fornecendo informações essenciais sobre o formato e a posição de articulação dos gestos, que serão posteriormente combinadas às etapas de análise temporal para compreender a sequência completa do movimento.

4.3.2. Arquitetura de Redes Neurais Recorrentes LSTM

A rede LSTM (Houdt; Mosquera e Nápoles, 2020) foi empregada para analisar a sequência temporal dos quadros que compõem cada vídeo. Diferente das redes tradicionais, que tratam cada dado de entrada de forma independente, a LSTM é capaz de compreender a relação entre os diferentes momentos do gesto, reconhecendo padrões de movimento ao longo do tempo. Essa característica é importante no reconhecimento de sinais em LIBRAS, pois muitos gestos são definidos não apenas pela forma das mãos, mas também pela direção e pela continuidade do movimento. Assim, a LSTM, conectada na saída da rede *EfficientNet*, permite que o modelo interprete o gesto como uma sequência coerente, aproximando-se da forma como os humanos percebem e compreendem a linguagem de sinais.

4.3.3. Bloco de Atenção Temporal

Em tarefas de reconhecimento de padrões em vídeos, alguns *frames* podem ter importância maior para diferenciar as classes entre si. O bloco de atenção temporal, como proposto por Yang et al. (2023), foi incorporado ao modelo com o objetivo de aprimorar a análise da sequência de *frames*, destacando os momentos mais relevantes do gesto. O mecanismo de atenção permite que a rede aprenda a enfatizar nos trechos mais significativos da sequência, atribuindo pesos maiores aos quadros que mais contribuem para a decisão final do modelo.

No contexto deste trabalho, o uso da atenção temporal é particularmente vantajoso, pois a LIBRAS é uma língua visual-espacial cuja interpretação depende de sutis variações de movimento, orientação e posição das mãos ao longo do tempo. Ao integrar esse bloco ao modelo LSTM, o sistema consegue priorizar as partes do vídeo que melhor representam o gesto. Dessa forma, o mecanismo de atenção contribui para uma representação temporal mais coerente, aumentando a confiabilidade do reconhecimento e do sistema em contextos educacionais.

4.4. Métricas de Avaliação

Durante o treinamento e o teste da solução proposta, o desempenho do modelo foi avaliado utilizando métricas amplamente reconhecidas em aprendizado de máquina para problemas de classificação multiclasse (Alves; Boldt e Paixão, 2024). Essas métricas permitem medir o quanto o sistema foi capaz de identificar corretamente os sinais e distinguir gestos semelhantes. Nas expressões a seguir, VP (verdadeiro positivo) representa os sinais corretamente classificados em sua categoria real, VN (verdadeiro negativo) indica os demais sinais corretamente reconhecidos como não pertencentes àquela classe, FP (falso positivo) refere-se aos casos em que um gesto foi classificado incorretamente como outro, e FN (falso negativo) corresponde às situações em que o modelo não conseguiu identificar um gesto existente.

$$\text{Acurácia } (ACC) : (VP + VN)/(VP + VN + FP + FN) \quad (1)$$

$$\text{Precisão } (PRE) : VP/(VP + FP) \quad (2)$$

$$\text{Sensibilidade } (SEN) : VP/(VP + FN) \quad (3)$$

$$\textit{F1-Score } (F1) : 2 * (PRE * SEN)/(PRE + SEN) \quad (4)$$

Dessa forma, a acurácia representa a proporção total de classificações corretas. A precisão mede o quanto das predições feitas como sendo de um determinado sinal estavam realmente corretas. Essa métrica é importante quando diferentes sinais podem ser visivelmente parecidos, indicando se o sistema confunde gestos semelhantes. A sensibilidade mostra se o modelo consegue identificar corretamente todas as ocorrências dado um sinal, ou seja, o quanto o sistema consegue encontrar todos os exemplos reais de um gesto. Por fim, o *F1-Score* combina precisão e sensibilidade em uma única métrica harmônica, equilibrando casos em que o modelo pode acertar muitas vezes, mas ainda deixar de reconhecer alguns gestos. Dessa forma, as métricas combinadas oferecem uma visão completa sobre a capacidade do sistema em reconhecer sinais de LIBRAS.

5. Resultados

Esta seção apresenta e discute os resultados obtidos a partir dos experimentos realizados com o modelo proposto. São descritos o desempenho quantitativo de diferentes variações da arquitetura, as análises comparativas entre suas configurações e as implicações educacionais em contextos de ensino e acessibilidade dos resultados.

5.1. Configurações de Experimentos e Preparação da Base de Dados

Os experimentos foram realizados utilizando a categoria 3 da base de dados *LIBRAS-UFOP*, composta por oito sinais: amar, aprender, analisar, conversar, galinha, galo, interagir e trocar. Esses sinais compartilham o mesmo movimento e configuração de mão, diferenciando-se apenas pelo ponto de articulação, o que resulta em alta similaridade visual e gestual. Embora nem todos pertençam a um vocabulário diretamente educacional, essa característica os torna adequados para avaliar a capacidade do modelo em distinguir variações sutis entre gestos, um dos principais desafios no ensino e aprendizado da LIBRAS (Cerna *et al.*, 2021). A escolha dessa categoria permitiu reduzir a complexidade inicial do problema e realizar uma análise mais controlada do modelo. Cada vídeo foi padronizado para 30 *frames* com resolução de 224×224 *pixels*, utilizando *padding*, recorte central da região de interesse e a técnica CLAHE para realce de contraste, conforme descrito na Seção 4.2.

O ambiente de desenvolvimento utilizou a biblioteca PyTorch 2.0 (Imambi; Prakash e Kanagachidambaresan, 2021) e a divisão de dados foi feita por sinalizador, evitando o vazamento de informações de um mesmo sinalizador em conjuntos diferentes, garantindo que a avaliação reflita sua capacidade real de generalização para novos usuários. Três sinalizadores foram destinados ao treinamento, um à validação e um ao teste, mantendo essa proporção em todas as execuções. O treinamento seguiu 100 épocas, *batch size* de 4 e otimizador Adam (Kingma e Ba, 2014) com taxa de aprendizado inicial de 1×10^{-3}, aliado à política *ReduceLROnPlateau* e parada antecipada de 40 épocas (Alves; Boldt e Paixão, 2024). O protocolo *hold-out* (Dwork *et al.*, 2015) foi repetido cinco vezes de forma independente, reportando-se média e desvio padrão dos resultados, equilibrando custo computacional, desempenho e reprodutibilidade da análise.

5.2. Análise dos Resultados

A etapa de experimentação teve como objetivo avaliar o desempenho do modelo proposto e compreender o impacto de cada componente da arquitetura na tarefa de reconhecimento de sinais em LIBRAS. Para isso, o protocolo de avaliação adotado teve como objetivo comparar versões simplificadas e incrementais da arquitetura, permitindo validar o método proposto de forma gradual e auxiliando na compreensão do impacto de cada componente na qualidade do reconhecimento dos sinais. As métricas de acurácia, precisão, sensibilidade e *F1-Score* foram calculadas ao final de cada execução, e seus valores médios e desvios padrão são apresentados na Tabela 1.

Tabela 1. Média e desvio padrão (%) dos modelos avaliados no estudo de ablação.

Modelo	Acurácia	Precisão	Recall	F1-Score
EfficientNet	79,78 ± 9,44%	81,01 ± 12,84%	79,78 ± 9,44%	80,39 ± 12,31%
EfficientNet + LSTM	83,11 ± 7,04%	87,33 ± 7,09%	83,11 ± 7,04%	85,13 ± 6,84%
EfficientNet + Atenção	84,89 ± 4,37%	86,22 ± 5,95%	84,89 ± 4,37%	85,46 ± 4,42%
EfficientNet + LSTM + Atenção	**95,78 ± 3,07%**	**96,82 ± 1,98%**	**95,78 ± 3,07%**	**96,48 ± 2,08%**

Os resultados obtidos evidenciam a evolução gradual do desempenho conforme novos componentes foram incorporados à arquitetura. O modelo baseado apenas na *EfficientNet* apresentou acurácia de 79,78%, precisão de 81,01%, sensibilidade de 79,78% e *F1-Score* de 80,39%, mostrando que a rede foi capaz de extrair padrões espaciais relevantes das imagens, mas ainda limitada na compreensão do movimento completo. A inclusão da rede LSTM ampliou o desempenho geral, com ganhos perceptíveis em todas as métricas, atingindo 87,33% de precisão e 85,13% de *F1-Score*, e uma redução considerável no desvio padrão, evidenciando que o modelo se tornou mais estável e sensível às variações temporais dos gestos. Já o modelo com apenas o bloco de atenção apresentou resultados similares, mantendo o equilíbrio entre precisão e estabilidade.

O maior avanço foi obtido com a combinação completa entre a *EfficientNet*, a LSTM e o bloco de atenção temporal, que alcançou métricas superiores a 95%. Essa configuração demonstrou que o mecanismo de atenção é mais efetivo quando aliado à modelagem sequencial, permitindo ao modelo direcionar o foco para os *frames* mais informativos e reduzir erros em sinais com movimentos semelhantes. Além disso, a queda no desvio padrão para valores entre 2% e 3% indica um comportamento mais consistente e menos dependente das particularidades individuais dos sinalizadores. Essa redução nas oscilações demonstra que o modelo se tornou menos dependente das características particulares de quem executa o gesto, o que é fundamental para aplicações práticas em ambientes educacionais, onde a diversidade de usuários é alta.

Do ponto de vista técnico, o trabalho também se destaca pelo uso exclusivo das informações visuais contidas nos vídeos RGB da base LIBRAS-UFOP. Enquanto outros estudos da literatura exploram dados complementares, como mapas de profundidade e coordenadas esqueléticas (informações também disponíveis na base de dados) (Cerna *et al.*, 2021; Alves; Boldt e Paixão, 2024), o modelo proposto obteve resultados expressivos utilizando apenas os vídeos. Essa escolha simplifica a aplicação prática da solução. Dessa forma, o método proposto demonstra bons resultados quantitativos, demonstrando robustez e equilíbrio em seu desempenho.

5.3. Sugestões de Aplicações no Contexto Educacional

Sob a perspectiva educacional, a solução de visão computacional apresentada e avaliada nesse trabalho se configura como um iniciativa para reduzir a distância para a inclusão de pessoas deficientes auditivas. A realidade aponta que uma pequena parcela da população brasileira tem conhecimento de LIBRAS. Dessa forma, a instância tecnológica

pode ajudar na inserção e promoção de atividades educacionais. Modelos, como os desenvolvidos, podem ser embarcados em *apps* e/ou ambientes com monitoramento por câmera (corporativos ou não), com a finalidade de promover entendimento e inclusão. Esses modelos podem traduzir certos sinais, ampliando as possibilidades de comunicação linguística da LIBRAS em ambientes de aprendizagem remotos ao reduzir as barreiras linguísticas entre surdos e ouvintes, especialmente em contextos onde o domínio da LIBRAS ainda é limitado.

Professores também podem utilizá-lo para acompanhar o progresso dos estudantes, identificar dificuldades específicas, adaptar estratégias de ensino e até mesmo verificar os gestos durante atividades práticas. Assim, também é possível pensar que as pessoas envolvidas, tendo contato com a tecnologia, podem se valer do aprendizado dos sinais durante o processo. O mesmo ainda poderia ser aplicado em ambientes educacionais presenciais, com a finalidade direta de interação sobre atividades e/ou brincadeiras que de forma ativa criam conhecimento sobre os sinais de LIBRAS.

Em um cenário de crescente digitalização da educação, o modelo apresentado neste trabalho representa uma contribuição concreta para o desenvolvimento de ferramentas assistivas baseadas em IA. Embora o presente trabalho ainda se concentre na validação experimental do modelo, os resultados obtidos demonstram sua viabilidade técnica e reforçam a relevância de soluções desse tipo para a promoção da acessibilidade linguística e inclusão digital. A partir desse avanço, abre-se espaço para o desenvolvimento de interfaces educacionais acessíveis, sistemas de tradução automática de sinais e ambientes de aprendizagem mais inclusivos. Assim, este estudo reforça a importância do uso da IA como meio de promover não apenas inovação tecnológica, mas também igualdade de oportunidades no acesso à comunicação e à educação.

6. Conclusão

Este trabalho apresentou uma abordagem baseada em aprendizado profundo para o reconhecimento automático de sinais de LIBRAS, integrando extração espacial, análise temporal e mecanismo de atenção. A arquitetura proposta, composta pelas redes *EfficientNet* e LSTM, e pelo bloco de atenção temporal, demonstrou resultados promissores, atingindo altos índices de nas métricas apresentadas. As etapas de pré-processamento e o protocolo de avaliação adotado contribuíram para garantir a consistência dos experimentos e a capacidade de generalização do modelo frente a diferentes sinalizadores. Além do desempenho técnico, o estudo reforça o potencial de tecnologias de visão computacional como ferramentas de apoio à inclusão e ao ensino de LIBRAS, favorecendo uma maior acessibilidade comunicativa em ambientes educacionais.

Como trabalhos futuros, pretende-se expandir os experimentos para abranger os demais sinais disponíveis na base de dados, de modo a avaliar a escalabilidade e robustez da solução em contextos mais amplos. Também se planeja o desenvolvimento de uma interface interativa que permita a comunicação direta com o modelo, possibilitando sua utilização em plataformas de ensino, tradutores automáticos ou aplicativos voltados à mediação pedagógica. Dessa forma, o sistema poderá evoluir de uma prova de conceito para uma ferramenta prática e acessível, contribuindo de forma concreta para o fortalecimento da educação inclusiva e o ensino de LIBRAS.

Referências

Alves, C. E. G.; Boldt, F. D. A.; Paixão, T. M. Enhancing brazilian sign language recognition through skeleton image representation. In: IEEE. **2024 37th SIBGRAPI**

Conference on Graphics, Patterns and Images (SIBGRAPI). [S.l.], 2024. p. 1–6.
Azevedo, L. F.; Alencar, R. M. G. A importância do ensino da língua brasileira de sinais-(libras) para educação infantil e formação dos professores das séries iniciais. **Brazilian Journal of Development**, v. 7, n. 1, p. 5648–5671, 2021.
Benevides, M. P. *et al.* Sign talk assistive technology: Real-time recognition of the libras typical alphabet using artificial intelligence. **Revista de Gestão Social e Ambiental**, Centro Universitário da FEI, Revista RGSA, v. 18, n. 12, p. 1–13, 2024.
Caprì, T.; Cicceri, G.; Distefano, S.; Tartarisco, G. **Computer vision and human behaviour to recognize emotions**. [S.l.]: Frontiers Media SA, 2024. 1441678 p.
Cerna, L. R.; Cardenas, E. E.; Miranda, D. G.; Menotti, D.; Camara-Chavez, G. A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. **Expert Systems with Applications**, v. 167, p. 114179, 2021. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0957417420309143⟩.
Corrêa, Y.; Vieira, M. C.; Santarosa, L. M. C.; Biasuz, M. C. V. Tecnologia assistiva: a inserção de aplicativos de tradução na promoção de uma melhor comunicação entre surdos e ouvintes. **Revista Novas Tecnologias na Educação**, v. 12, n. 1, 2014.
Damasceno, A. R. P.; Holanda, S. F.; Lima, J. V. V.; Caldas, E. A. L.; Oliveira, F. C. d. M. B. Projetando avatares animados para promover a interação colaborativa entre pessoas surdas e ouvintes em ambientes educacionais de metaverso: um estudo com implicações para o dalverse. **RENOTE**, v. 23, n. 1, p. 483–595, 2025.
Dunne, R.; Matthews, O.; Vega, J.; Harper, S.; Morris, T. Computational methods for predicting human behaviour in smart environments. **Journal of Ambient Intelligence and Smart Environments**, SAGE Publications Sage UK: London, England, v. 15, n. 2, p. 179–205, 2023.
Dwork, C. *et al.* The reusable holdout: Preserving validity in adaptive data analysis. **Science**, American Association for the Advancement of Science, v. 349, n. 6248, p. 636–638, 2015.
Fei-Fei, L.; Deng, J.; Li, K. Imagenet: Constructing a large-scale image database. **Journal of vision**, The Association for Research in Vision and Ophthalmology, v. 9, n. 8, p. 1037–1037, 2009.
Fernando, T.; Denman, S.; Sridharan, S.; Fookes, C. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. **Neural networks**, Elsevier, v. 108, p. 466–478, 2018.
Gibson, R. **The Impact of AI in Advancing Accessibility for Learners with Disabilities**. 2024. EDUCAUSE Review. Acesso em: 18 out. 2025. Disponível em: ⟨https://er.educause.edu/articles/2024/9/the-impact-of-ai-in-advancing-accessibility-for-learners-with-disabilities⟩.
Hisaria, S.; Sharma, P.; Gupta, R.; Sumalatha, K. An analysis of multi-criteria performance in deep learning-based medical image classification: A comprehensive review. 2024.
Hochreiter, S.; Schmidhuber, J. Long short-term memory. **Neural Computation**, v. 9, p. 1735–1780, 11 1997.
Houdt, G. V.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. **Artificial intelligence review**, Springer, v. 53, n. 8, p. 5929–5955, 2020.
Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 7132–7141.
Imambi, S.; Prakash, K. B.; Kanagachidambaresan, G. Pytorch. In: **Programming with TensorFlow: solution for edge computing applications**. [S.l.]: Springer, 2021. p. 87–104.

Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

Krichen, M. Convolutional neural networks: A survey. **Computers**, MDPI, v. 12, n. 8, p. 151, 2023.

Li, D.; Opazo, C. R.; Yu, X.; Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**. [s.n.], 2020. p. 1459–1469. Disponível em: ⟨https://openaccess.thecvf.com/content_WACV_2020/papers/Li_Word-level_Deep_Sign_Language_Recognition_from_Video_A_New_Large-scale_WACV_2020_paper.pdf⟩.

Melo-López, V.-A.; Basantes-Andrade, A.; Gudiño-Mejía, C.-B.; Hernández-Martínez, E. The impact of artificial intelligence on inclusive education: A systematic review. **Education Sciences**, v. 15, n. 5, p. 539, 2025. Disponível em: ⟨https://www.mdpi.com/2227-7102/15/5/539⟩.

Musa, P.; Rafi, F. A.; Lamsani, M. A review: Contrast-limited adaptive histogram equalization (clahe) methods to help the application of face recognition. In: IEEE. **2018 third international conference on informatics and computing (ICIC)**. [S.l.], 2018. p. 1–6.

Sakis, I.; Bernardi, G. Rede social de aprendizagem colaborativa para surdos e deficientes auditivos. **RENOTE**, v. 18, n. 1, 2020.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 4510–4520.

Santos, M. C. R. dos; Oliveira, F. K. de. O uso de inteligência artificial (ia) no ensino de espanhol: superando barreiras no ambiente escolar. **RENOTE**, v. 22, n. 3, p. 164–173, 2024.

Stefano, G. dos S.; Passos, W. L.; Gois, J. N.; Araujo, G. M.; Lima, A. A. de. Um sistema de reconhecimento de sinais em libras usando cnn e lstm. 2021.

Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. **International conference on machine learning**. [S.l.], 2019. p. 6105–6114.

Vanneste, P. *et al.* Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement. **Mathematics**, MDPI, v. 9, n. 3, p. 287, 2021.

Vaswani, A. *et al.* Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

Wang, Z.; Zheng, Y.; Liu, Z.; Li, Y. A survey of video human behaviour recognition methodologies in the perspective of spatial-temporal. In: IEEE. **2022 2nd International Conference on Intelligent Technology and Embedded Systems (ICITES)**. [S.l.], 2022. p. 138–147.

Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. **Journal of Big data**, Springer, v. 3, n. 1, p. 9, 2016.

Yang, K.; Ding, Y.; Geng, F.; Jiang, H.; Zou, Z. A multi-sensor mapping bi-lstm model of bridge monitoring data based on spatial-temporal attention mechanism. **Measurement**, Elsevier, v. 217, p. 113053, 2023.