Aprendizado Profundo para Reconhecimento de Sinais Isolados em Libras

Gabriela T. Barreto, Orientador: Pedro Olmo Stancioli Vaz de Melo
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais

***Abstract*—This work investigates deep learning approaches for the recognition of isolated signs in Brazilian Sign Language (Libras) using pose- and hand-based landmark sequences extracted with MediaPipe Holistic. Two neural architectures were evaluated: a bidirectional LSTM encoder and a compact Transformer encoder. Experiments were conducted on two Brazilian RGB video datasets, MINDS-Libras and V-Librasil, considering signer-dependent, signer-independent, cross-dataset transfer, and mixed-domain training scenarios. The results show that the Transformer model achieves higher performance when sufficient data is available, outperforming the LSTM in both signer-dependent and signer-independent evaluations on MINDS-Libras. However, both models exhibit limited generalization across domains, achieving near-zero transferability when trained on MINDS-Libras and tested directly on V-Librasil, highlighting substantial differences between datasets despite visual similarity. Mixed-dataset training partially mitigates these effects but still reveals strong dataset dependency. Overall, the findings reinforce the need for larger, more diverse and standardized Libras datasets, and suggest that domain robustness remains a key challenge for automatic sign recognition in Brazilian Sign Language.**

I. Introdução

De acordo com o censo mais recente sobre pessoas com deficiência auditiva no Brasil, realizado pelo Instituto Brasileiro de Geografia e Estatística(IBGE), e datado de 2019, cerca de 10 milhões de pessoas no país são portadoras de algum grau de deficiência auditiva [1], e dessas, aproximadamente 2,7 milhões possuem surdez profunda, ou seja, não escutam absolutamente nada. Especificamente entre as pessoas com deficiência auditiva de 5 a 40 anos, estima-se que apenas $22{,}4\%$ sabem usar a Língua Brasileira de Sinais (Libras). Já no grupo que não consegue ouvir de forma alguma, esse percentual é de $61{,}3\%$.

Em relação ao número de pessoas ouvintes que se comunicam em Libras, ainda não há uma estimativa oficial. No entanto, especialistas apontam que esse número é reduzido e mal documentado, refletindo uma lacuna nas políticas públicas e nas estatísticas nacionais [2].

A população surda sofre um alto nível de exclusão social. Esse grupo utiliza a Libras como principal meio de comunicação, mas em diversos espaços — como atendimentos, serviços, entretenimento e ambientes públicos — a pesso surda precisa recorrer ao português escrito para interagir com pessoas ouvintes. Nessas situações, recai sobre o surdo, enquanto minoria social no Brasil, o esforço contínuo de adaptação, comunicação e sobrevivência em um mundo que não foi pensado para ela.

Dessa forma, um tradutor de Libras para o português poderia funcionar como uma solução tecnológica de comunicação que contraria a lógica dominante, em que grupos excluídos precisam se adaptar ao mundo ouvinte. Em vez disso, propõe-se uma inversão: que os ouvintes também assumam a responsabilidade de se adaptar e se comunicar com a população surda. Além disso, tal ferramenta permitiria que pessoas surdas se comunicassem com um número maior de pessoas, não apenas em interações presenciais, mas também por meio da internet. Isso possibilitaria que pessoas surdas, que têm a Libras como forma primária de comunicação, pudessem também criar conteúdo, se expressar e alcançar públicos, da mesma forma que fazem os criadores ouvintes — diminuindo as barreiras entre comunidades surdas e ouvintes.

No Brasil, e principalmente no meio internacional, pesquisadores têm investigado abordagens diversas para tradução automática de outras línguas de sinais e reconhecimento de sinais, como a American Sign Language (ASL), Chinese Sign Language(CSL), entre outras linguagens, empregando técnicas de visão computacional, aprendizado de máquina e processamento de linguagem natural. Diante desse cenário, o projeto realizado na disciplina Projeto Orientado em Computação I consistiu em uma revisão da literatura nacional e internacional sobre reconhecedores de sinais de línguas de sinais para texto escrito, com ênfase nos reconhecedores de Libras para o português.

A. Trabalhos relacionados

A partir da leitura de cerca de 20 trabalhos, nacionais e internacionais, a pesquisa mapeou as principais abordagens, ferramentas, desafios e tendências tecnológicas presentes na área, contribuindo para a consolidação do conhecimento científico e para a identificação de lacunas relevantes. Esse levantamento foi realizado em duas análises: uma no contexto nacional e outra no internacional, nas quais foram identificadas as particularidades de cada cenário.

Na leitura dos textos, observou-se um padrão: o reconhecimento funciona em duas etapas — primeiro há uma de extração de features e, em seguida, a de reconhecimento. Assim, foi feito um levantamento das principais tecnologias utilizadas para cada uma dessas fases.

Na etapa de extração de características, destacam-se o MediaPipe, algoritmos tradicionais de visão computacional e CNNs, empregados para extrair informações dos sinais representados em imagens ou vídeos. Já na etapa de reconhecimento de sinais, os métodos mais recorrentes envolvem redes neurais,

com predomínio do uso de CNNs. Mais recentemente, têm sido explorados também os Vision Transformers (ViTs), além de arquiteturas híbridas.

Do projeto apresentado no último semestre, foi possível levantar os principais problemas em aberto na área, assim como os desafios. Alguns deles são:

- Falta de bases de dados amplas e diversificadas, tanto no contexto nacional quanto internacional;
- Dificuldade de desenvolver modelos agnósticos a domínio;
- Pesquisas que utilizam sensores de luvas, o que não é muito prático para soluções escaláveis;
- Pesquisas em fingerspelling que não têm grande aplicabilidade;
- Tradução contínua ainda é um problema longe estar resolvido;
- Soluções ainda não conseguem lidar com oclusão de parte do sinalizador;
- Desenvolvimento de sistemas em tempo real;
- Ausência de bases de dados para tradução contínua da Libras;
- Poucos trabalhos de tradução da Libras até o momento, e ainda menos utilizando transformers.

Por fim, essa etapa inicial foi crucial para criar o embasamento teórico para o desenvolvimento de uma solução de reconhecimento de sinais de Libras na disciplina de POC II. Este é um passo importante que, posteriormente, poderá viabilizar a criação de um sistema tradutor de Libras para o português escalável.

B. Presente trabalho

Dessa forma, o presente trabalho propôs realizar uma série de experimentos, com o objetivo de avaliar diferentes abordagens em redes neurais aplicadas ao reconhecimento de sinais da Libras para o português escrito. Um dos principais objetivos desta pesquisa foi investigar modelos agnósticos ao domínio, ou seja, que são capazes de performar bem em diferentes bases e com diferentes sinalizadores não vistos durante o treinamento.

Essa investigação foi importante, dado que há pouca pesquisa no Brasil sobre o tema, e ainda menos trabalhos com a proposta de desenvolver modelos agnósticos ao domínio. Além disso, no contexto atual da pesquisa, muitos artigos analisados apresentam o problema de utilizar o mesmo modelo em bases diferentes, com o modelo performando bem em uma base, mas não em outra. Assim, ao avaliar as arquiteturas LSTM e Transformer em várias configurações experimentais — cenários signer-dependent, signer-independent, transferência entre domínios e treino com bases combinadas — este trabalho busca identificar limitações, comportamentos e potencialidades de cada abordagem. Com isso, pretende-se avançar na direção do desenvolvimento de sistemas mais robustos e escaláveis para uso real, onde a variabilidade entre usuários e ambientes é inevitável.

II. Bases de dados

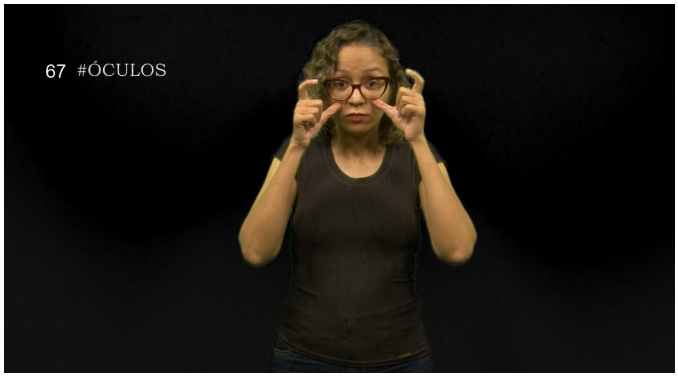
A escolha das bases de dados foi baseada nas leituras feitas na POC I. Neste trabalho, optou-se por usar bases de vídeo em RGB. Foram selecionadas duas bases brasileiras importantes desenvolvidas nos últimos anos: MINDS-Libras [3] e V-Librasil [4].

Apesar de ambas terem sido construídas de forma semelhante — com sinalizadores em ambientes controlados, filmados do tronco para cima e com fundo de chroma key — elas apresentam diferenças em número de classes, quantidade de amostras e variabilidade entre sinalizadores, o que impacta diretamente a capacidade de generalização dos modelos.

A base MINDS-Libras, desenvolvida na UFMG, é composta por 20 classes. Cada sinalizador realizou cinco repetições de cada sinal, o que permite capturar variações naturais de movimento, velocidade e posicionamento. A presença de 12 sinalizadores constitui um diferencial, pois fornece maior diversidade entre sujeitos que o modelo vê, o que é essencial para avaliar modelos em cenários signer independent.

Sinalização de Acontecer, amostra do MINDS-Libras

Já a base V-Librasil é uma das mais extensas bases disponíveis em número de classes, contendo 1327 sinais distintos. Porém, ela possui uma limitação significativa quanto ao número de sinalizadores, são apenas três e cada sinal é executado apenas uma vez por cada um deles, resultando em um total de apenas 3991 amostras.



Sinalização de Óculos, amostra do V-Librasil

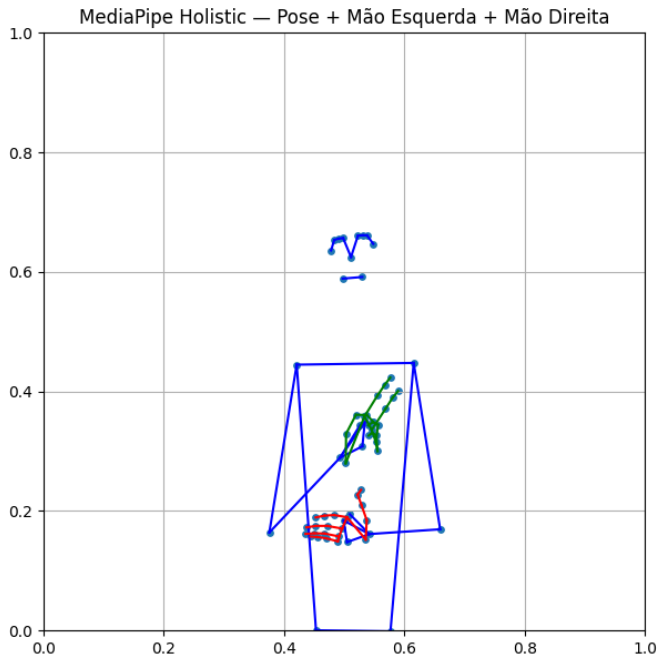
Em ambas as bases, os vídeos apresentam visão frontal estática, sem mudanças de ângulo, e foram capturados sobre um fundo simples - verde (chroma key) no MINDS, e preto no V-Librasil. Ainda assim, existem algumas diferenças sutis

— como distância da câmera, área ocupada pelo corpo, iluminação, fluência dos sinais e estilo individual do sinalizador — que podem afetar o desempenho dos modelos e a transferência entre bases.

A. Extração de características

A etapa de extração de características é crucial para o desempenho dos modelos, visto que ela determina a forma de entrada dos dados e precisa ser eficaz para captar os movimentos das mãos, tronco e rosto do sinalizador. Para este trabalho, optou-se por utilizar o MediaPipe Holistic, que fornece landmarks por frame do vídeo. Escolheu-se utilizar landmarks corporais, das mãos e do tronco, que são as regiões responsáveis pela maior parte da informação linguística em sinais isolados.

Assim, cada frame das amostras foi representado por um total de 75 pontos, sendo 33 pontos do corpo (pose), que incluem também alguns pontos da cabeça — como olhos, nariz, orelhas e canto da boca — e que formam um contorno facial simplificado, além de 21 pontos da mão esquerda e 21 pontos da mão direita, correspondentes às articulações dos dedos e do punho. Cada landmark foi armazenado no formato (x,y,z,visibility), resultando em uma representação compacta que captura os movimentos essenciais do sinalizador.



Landmarks extraídos de amostra do MINDS-Libras

A principal vantagem do MediaPipe, em comparação com outras abordagens, é que ele funciona de forma ágil e robusta em tempo real, sendo uma alternativa mais eficiente e leve em relação a métodos baseados em CNNs tridimensionais.

Além disso, o uso de landmarks facilita o treinamento de modelos em bases especialmente grandes, pois a quantidade de parâmetros a serem aprendidos é drasticamente reduzida.

III. Modelos

A. LSTM Bidirecional

As redes LSTM (Long Short-Term Memory) são amplamente utilizadas em tarefas de modelagem de sequências devido à sua capacidade de aprender dependências de longo prazo. As LSTMs possuem mecanismos de gates de memória que permitem aprender padrões e aprender dependências temporais entre os frames. No modelo usado, a inclusão de uma camada bidirecional permite que o modelo avalie simultaneamente a sequência na direção temporal normal e na reversa, capturando padrões tanto no início quanto no final do sinal. Neste trabalho, a LSTM foi utilizada como baseline devido à sua simplicidade e ampla adoção na literatura, sendo uma rede frequente nos trabalhos analisados na POC I. Ela apresenta vantagens em datasets menores devido à sua menor complexidade e, para a tarefa de reconhecimento de sinais isolados, costuma apresentar boa performance; já para o reconhecimento contínuo (em nível de frase), nem tanto.

A arquitetura LSTM implementada neste trabalho consiste em um codificador sequencial bidirecional construído sobre o módulo de LSTM do PyTorch. Ela foi configurado com duas camadas empilhadas, tamanho de estado oculto de 256 e dropout interno de 0,3. O uso de recorrência bidirecional permite ao modelo capturar dependências temporais tanto no sentido passado-futuro quanto futuro-passado. O encoder produz uma representação compacta por meio da média temporal dos estados ocultos gerados pela última camada da LSTM, juntando informações dinâmicas da sequência inteira em um único embedding. Assim, o modelo implementado representa uma abordagem recorrente clássica, eficiente para conjuntos de dados moderados e capaz de capturar padrões temporais de curta e média duração nos sinais.

B. Trasformer Encoder

O Transformer revolucionou tarefas sequenciais ao substituir mecanismos recorrentes por mecanismos de atenção. Sua capacidade de capturar relações entre todos os elementos da sequência simultaneamente o torna altamente eficiente e escalável. Nos últimos anos, avanços feitos na pesquisa de reconhecimento de sinais isolados e contínuos devem-se, em grande parte, ao avanço do Transformer. Porém, no Brasil não foram identificados muitos estudos que utilizem Transformers, observando-se uma predominância de redes convolucionais e recorrentes nos trabalhos encontrados. Portanto, a proposta de usar o Transformer está alinhada com a pesquisa mais recente na literatura internacional, além de trazer resultados novos para o problema no contexto brasileiro. Transformers geralmente superam LSTMs, mas, por outro lado, são mais sensíveis à quantidade de dados — uma característica que ficou evidente nos resultados obtidos.

O modelo Transformer utilizado tem um formato compacto, usando exclusivamente o bloco codificador, denominado aqui de TinyTransformer. Ele foi implementado com o módulo TransformerEncoder do PyTorch. Nele, cada vetor de entrada é projetado linearmente para um espaço latente de

dimensão $d_{model} = 256$, de onde segue para uma pilha de quatro camadas Transformer, cada uma contendo mecanismos de autoatenção multi-cabeças e camadas feed-forward expandidas. Um token de classificação é adicionado no início da sequência, permitindo que o modelo agregue informação global por meio da autoatenção. Essa arquitetura elimina a dependência recorrente da LSTM e permite capturar relações de longo alcance entre frames, tornando-a mais expressiva em contextos com grande variabilidade espacial e temporal.

IV. Treinamentos

Os experimentos foram estruturados em quatro configurações principais, com o objetivo de estudar não somente o desempenho dos modelos, mas também sua capacidade de generalização entre sinalizadores e entre bases de dados distintas.

- Experimentos independentes por base:
  - Avaliar o limite superior do desempenho em cada base.
- Transfer learning:
  - Investigar a possibilidade de treinar em uma base e testar em outra, o que seria desejável em aplicações reais, onde o modelo irá lidar com estilos variados de sinalização e sinalizadores diversos.
- Bases combinadas:
  - A partir da amostras de classes em comum entre bases, testar o modelo de forma mais assertiva, observando seu ccomportamento para dados um pouco menos controlados(e mais diversos).
- Signer dependent vs. signer independent:
  - A execução de experimentos *signer independent* ocorre quando um modelo é treinado em um grupo de sinalizadores, mas testado em outros. É uma forma de teste que melhor simula a realidade, onde um modelo teria que resolver em tempo real a interpretação de sinais de sinalizadores nunca vistos antes.

A. Setup experimental

Já no que tange ao setup experimental, foram realizados testes de hiperparâmetros diversos. Observou-se que usar poucos frames resultava em perda de informação temporal, enquanto utilizar muitos frames tornava o treinamento pesado e propenso à instabilidade. Dessa forma, um valor intermediário de 40 frames por vídeo mostrou-se mais adequado. Para o treino nas na MINDS-Libras foram usadas aproximadamente 30 épocas para a LSTM e cerca de 70 épocas para o Transformer. O melhor modelo de cada experimento era salvo com base na menor loss observada durante o treinamento e posteriormente utilizado na etapa de teste.

No processo de ajuste dos hiperparâmetros, também foram testadas diferentes taxas de aprendizado, tamanhos de batch e configurações de normalização temporal. Valores muito altos de learning rate levaram a oscilações no treinamento, enquanto taxas moderadas, como 3×10^{-4}, favoreceram convergência estável. O batch size de 32 apresentou bom equilíbrio entre estabilidade e velocidade.

No caso dos experimentos com bases combinadas, embora houvessem menos amostras por classe, as amostras eram mais diversas, pois provinham de dois domínios distintos. Para lidar com essa maior variabilidade entre os dados, foi neessário treinar ambos os modelos por 45 épocas, garantindo tempo suficiente para que as redes conseguissem estabilizar o aprendizado mesmo diante da heterogeneidade das sequências.

V. Resultados

Os resultados revelam diferenças relevantes entre a capacidade dos modelos de lidar com bases distintas. No caso do MINDS-Libras, houve convergência rápida e desempenho elevado para ambos os modelos no cenário signer dependent. O Transformer destacou-se com acurácia de 90,6%, demonstrando alta capacidade de aprendizado quando dispõe de quantidade moderada de dados e um tamanho enxuto.

Base de Dados	Modelo	Acurácia (%)	SD/SI
MINDS-Libras	LSTM	81,4	SD
MINDS-Libras	Transformer	90,6	SD
MINDS-Libras	LSTM	69,3	SI
MINDS-Libras	Transformer	82,7	SI

SI: signer independent, SD: signer dependent

Porém, no cenário signer independent, observou-se uma queda significativa nas duas arquiteturas. Esse comportamento já era esperado e reflete a dificuldade de generalização dos modelos e a dependência ao estilo individual dos sinalizadores, comentada amplamente na literatura.

Para a base V-Librasil, nenhum modelo convergiu satisfatoriamente quando treinado isoladamente. A combinação de número reduzido de amostras por classe e a grande quantidade de classes criou um cenário de alta esparsidade, difícil de ser tratado sem técnicas avançadas de aumento de dados ou modelos pré-treinados.

Já no teste de transfer learning, o número de classes era sete, correspondente a interseção de classes do MINDS-Libras e do V-Librasil. Os modelos treinados na MINDS-Libras foram testados diretamente na base do V-Librasil(em um total de 21 amostras), sem finetune, e a acurácia foi zero, o que indica incompatibilidade entre domínios, mesmo considerando que ambas as bases foram filmadas em condições semelhantes.

A partir da seleção de amostras das classes em comum das duas bases de dados, foi feito um treinamento com batches representativos de ambos os datasets.

Base de Dados	Modelo	Acurácia (%)	SD/SI
V-Librasil+Minds	LSTM	81,5	SD
V-Librasil+Minds	Transformer	83,7	SD
V-Librasil+Minds	LSTM	73,2	SI
V-Librasil+Minds	Transformer	73,3	SI

Observa-se que, no cenário de bases combinadas, o desempenho dos modelos apresenta uma redução em comparação com os experimentos realizados apenas no MINDS-Libras.

Essa queda era esperada, uma vez que o V-Librasil é significativamente mais desafiador devido ao pequeno número de amostras por classe. Ainda assim, o treinamento conjunto possibilitou ao modelo aprender características mínimas necessárias para reconhecer sinais do V-Librasil, algo que não ocorreu no experimento de transferência direta, em que a acurácia foi zero quando o modelo treinado exclusivamente no MINDS-Libras foi testado no V-Librasil. Isso reforça que, para que o modelo consiga generalizar para o domínio do V-Librasil, é indispensável incluir pelo menos algumas amostras dessa base durante o treinamento. Além disso, nota-se que o desempenho entre LSTM e Transformer torna-se muito próximo nesse cenário, indicando que, com menos dados disponíveis por classe, as vantagens do Transformer são reduzidas, e ambos os modelos passam a operar em regime semelhante de limitação de dados.

Conclusões

Os experimentos no MINDS-Libras mostraram que o Transformer se apresentou como o modelo de melhor desempenho, que foi o cenário com a maior quantidade de dados, superando o LSTM tanto no teste signer dependent quanto signer independent.

No entanto, também observou-se que ambos os modelos ainda dependem fortemente das características específicas do conjunto de treino, dado a sua ausência de acertos ao fazer o teste de transfer learning. A diferença entre as gravações das bases impediu a transferência direta de conhecimento, apesar da sua similariedade, dado que ambas são gravadas em chroma key, com visão total do sinalizador.

Além disso, nos experimentos com bases misturadas, onde havia menos dados por classe, o desempenho do Transformer aproximou-se do do LSTM, o que reforça que modelos baseados em atenção necessitam de mais dados para expressar todo o seu potencial.

Assim, a análise conjunta dos experimentos indica que embora modelos baseados em deep learning consigam alcançar desempenhos elevados em bases controladas e relativamente pequenas, sua capacidade de generalização permanece limitada, principalmente diante da variabilidade de sinalizadores e das diferenças entre condições de gravação. Ter um bom desempenho em cenários controlados não implica em uso prático, onde como estilo individual do sinalizador, velocidade de execução, enquadramento, fundo, condições climáticas e de iluminação variam. Além disso, modelos que performaram de forma consistente signer-dependent sofreram quedas de acurácia quando expostos a dados novos, o que reforça a dependência do modelo às características do conjunto de treinamento.

Dessa forma, os resultados reforçam três pontos principais. Primeiro, a necessidade de bases de dados maiores e mais diversas, tanto em termos de número de sinalizadores quanto de condições de captura. A ausência de variabilidade suficiente faz com que os modelos aprendam padrões específicos dos indivíduos ou do ambiente, em vez de características linguísticas da Libras em si. A literatura internacional [] mostra que avanços substanciais em línguas de sinais ocorreram apenas quando bases extensas e multimodais passaram a ser utilizadas, o que ainda não se observa no contexto brasileiro.

Segundo, os experimentos ilustram as limitações dos modelos atuais em cenários de transferência entre domínios. O fato de que um modelo treinado somente no MINDS-Libras não conseguiu reconhecer nenhuma classe do V-Librasil evidencia essa problema. Isso aponta para a necessidade de técnicas mais avançadas, como domain adaptation ou treinamento adversarial entre domínios para obter maiores desenvolvimentos.

Por fim, observou-se também a dificuldade de usar Transformers em contextos com poucos dados. Enquanto a LSTM se mostrou mais estável em cenários com menor quantidade de amostras, o Transformer teve um comportamento mais sensível à variabilidade e à escassez de dados, aproximando seu desempenho ao da LSTM quando o número de exemplos por classe era reduzido.

Esses resultados estão de acordo com trabalhos recentes na área, que argumentam que o grande gargalo atual não é a escolha da arquitetura, mas a inexistência de bases adequadas e suficientemente amplas para treinar sistemas verdadeiramente robustos de reconhecimento automático de Libras. Assim, o futuro da área depende não apenas de avanços em modelagem, mas principalmente do investimento na construção de bases diversas, padronizadas e representativas da comunidade surda brasileira.

References

[1] Instituto Brasileiro de Geografia e Estatística (IBGE), "Pns 2019: país tem 17,3 milhões de pessoas com algum tipo de deficiência," https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/31445-pns-2019-pais-tem-17-3-milhoes-de-pessoas-com-algum-tipo-de-deficiencia, 2019, acesso em: 19 abr. 2025.

[2] Agência Brasil, "Programa caminhos da reportagem aborda o brasil que usa libras," https://agenciabrasil.ebc.com.br/direitos-humanos/noticia/2021-09/programa-caminhos-da-reportagem-aborda-o-brasil-que-usa-libras, 2021, acesso em: 19 abr. 2025.

[3] T. M. Rezende, "Reconhecimento automático de sinais da libras: desenvolvimento da base de dados minds-libras e modelos de redes convolucionais," 2021.

[4] A. J. RODRIGUES, "V-librasil: uma base de dados com sinais na língua brasileira de sinais (libras)," Master's thesis, Universidade Federal de Pernambuco, 2021.