UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Ana Beatriz Gomes Takehara

Avaliação de transferibilidade entre modelos de redes neurais pré-treinados em línguas estrangeiras de sinais e LIBRAS

São José do Rio Preto

2025

Ana Beatriz Gomes Takehara

Avaliação de transferibilidade entre modelos de redes neurais pré-treinados em línguas estrangeiras de sinais e LIBRAS

Monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Universidade Estadual Paulista "Júlio de Mesquita Filho" - (UNESP)

Instituto de Biociências, Letras e Ciências Exatas - (IBILCE)

Departamento de Ciências de Computação e Estatística

Bacharelado em Ciência da Computação

Orientador: Lucas Correia Ribas

São José do Rio Preto

2025

T136a

Takehara, Ana Beatriz Gomes

Avaliação de transferibilidade entre modelos de redes neurais pré-treinados em línguas estrangeiras de sinais e LIBRAS / Ana Beatriz Gomes Takehara. -- São José do Rio Preto, 2025

67 p. : tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Lucas Correia Ribas

1. Língua brasileira de sinais. 2. Inteligência artificial. 3. Visão por computador. 4. Redes neurais (Computação). I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Ana Beatriz Gomes Takehara

Avaliação de transferibilidade entre modelos de redes neurais pré-treinados em línguas estrangeiras de sinais e LIBRAS

Monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Banca Examinadora:

Prof(a). Dr(a). Lucas Correia Ribas
UNESP – Câmpus de São José do Rio Preto
Orientador

Prof(a). Dr(a). Wallace Correa de Oliveira Casaca
UNESP – Câmpus de São José do Rio Preto

Prof(a). Dr(a). Rogéria Cristiane Gratão de Souza
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
03/11/025

À minha família, meu alicerce e inspiração. Obrigada por acreditarem em mim, pelo apoio incondicional e por tornarem possível cada passo desta jornada. Este trabalho é nosso.

Com amor e gratidão.

Agradecimentos

À minha mãe, pelo amor incondicional e por acreditar sempre em mim. Sua força e dedicação foram fonte de inspiração para que eu chegasse até aqui. Cada conquista minha carrega um pouco do seu amor, e esta vitória também é sua.

Ao meu pai, por me ensinar o valor do trabalho e da perseverança. Sua confiança em mim foi o combustível que me trouxe até este momento. As lições que me transmitiu ecoam em cada página deste trabalho, e minha gratidão será eterna por tudo o que representa em minha vida.

À minha irmã mais nova, por tornar minha vida mais leve, divertida e especial. Sua torcida sincera e suas palavras de carinho foram meu refúgio em momentos de dificuldade. Mesmo sendo a caçula, me ensina muito sobre a vida. Este trabalho é também para você, minha eterna parceira.

Aos meus amigos, por serem a família que escolhi. Dividiram comigo as angústias e celebraram cada pequena vitória. Entre risadas e desabafos, tornaram a caminhada mais leve e possível. O apoio e a presença de vocês foram a energia que me manteve firme ao longo dessa jornada. Minha eterna gratidão por acreditarem em mim.

Ao meu orientador, pela paciência, sabedoria e dedicação ao longo desta pesquisa. Suas orientações foram fundamentais para o desenvolvimento deste trabalho. Sua experiência e conhecimento me guiaram nos caminhos da ciência, e levo comigo admiração e gratidão por todo o tempo compartilhado.

Aos professores, por transmitirem conhecimento e despertarem minha curiosidade acadêmica. Cada aula representou um tijolo na construção da minha formação. Vocês plantaram sementes que hoje florescem neste trabalho, e serei eternamente grata por serem verdadeiros mestres da vida.

À UNESP, por ter sido minha casa acadêmica e berço de sonhos. Aqui encontrei não apenas conhecimento, mas também minha identidade profissional. Esta universidade me transformou e me preparou para o futuro. Levo comigo o orgulho e a honra de ser filha desta instituição.

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original."
(Albert Einstein)

Resumo

O reconhecimento automático de Língua Brasileira de Sinais (LIBRAS) por computador representa um desafio fundamental para a inclusão digital da comunidade surda no Brasil. Embora sistemas de reconhecimento de gestos tenham avançado significativamente, o desenvolvimento de modelos específicos para LIBRAS é limitado pela escassez de bases de dados robustas e pela complexidade inerente aos padrões espaciais e temporais dos sinais. Este trabalho investigou a viabilidade do aprendizado por transferência como estratégia para superar essas limitações, utilizando conhecimento adquirido em outras línguas de sinais para melhorar o reconhecimento em LIBRAS. Foram implementadas e comparadas três arquiteturas complementares: I3D, baseada em convoluções tridimensionais para análise de vídeos RGB; Bi-LSTM, para modelagem de dependências temporais em sequências de poses; e SL-GCN, que utiliza grafos espaço-temporais para capturar a dinâmica corporal. Cada modelo foi inicializado com pesos pré-treinados em bases de dados internacionais (WLASL, GSL, LSA64, AUTSL, DEVISIGN e INCLUDE) e ajustado para o reconhecimento em um conjunto unificado de classes em LIBRAS. Os resultados demonstraram que o aprendizado por transferência proporciona ganhos expressivos de desempenho, mesmo em cenários com dados limitados, comprovando a viabilidade da transferência de conhecimento entre línguas de sinais distintas. Este trabalho contribui para o desenvolvimento de métodos eficientes e generalizáveis para o reconhecimento automático de LIBRAS, reduzindo barreiras tecnológicas e promovendo acessibilidade e inclusão digital para a comunidade surda brasileira.

Palavras-chave: LIBRAS; Reconhecimento de Língua de Sinais; Aprendizado por Transferência; Redes Neurais Profundas; Visão Computacional; Classificação de Vídeos

Abstract

Automatic recognition of Brazilian Sign Language (LIBRAS) by computer represents a fundamental challenge for digital inclusion of the deaf community in Brazil. Although gesture recognition systems have advanced significantly, the development of models specific to LIBRAS is limited by the scarcity of robust databases and the inherent complexity of spatial and temporal patterns in sign language. This work investigated the feasibility of transfer learning as a strategy to overcome these limitations, leveraging knowledge acquired from other sign languages to improve recognition in LIBRAS. Three complementary architectures were implemented and compared: I3D, based on three-dimensional convolutions for RGB video analysis; Bi-LSTM, for modeling temporal dependencies in pose sequences; and SL-GCN, which utilizes spatio-temporal graphs to capture body dynamics. Each model was initialized with pre-trained weights from international databases (WLASL-2000, GSL, LSA64, AUTSL, DEVISIGN, and INCLUDE) and fine-tuned for recognition on a unified set of LIBRAS classes. The results demonstrated that transfer learning provides significant performance gains, even in scenarios with limited data, proving the viability of knowledge transfer between distinct sign languages. This work contributes to the development of efficient and generalizable methods for automatic LIBRAS recognition, reducing technological barriers and promoting accessibility and digital inclusion for the Brazilian deaf community.

Keywords: *LIBRAS; Sign Language Recognition; Transfer Learning; Deep Neural Networks; Computer Vision; Video Classification*

Sumário

Lista de ilustrações **11**

Lista de tabelas **12**

1 INTRODUÇÃO **14**
1.1 Objetivo **15**
1.2 Visão Geral do Método e Justificativa **16**
1.3 Estrutura deste Trabalho **16**

2 FUNDAMENTAÇÃO E REVISÃO DA LITERATURA **17**
2.1 Inteligência Artificial e Aprendizado de Máquina **17**
2.2 Redes Neurais Artificiais **18**
2.2.1 Funcionamento de um Neurônio Artificial 18
2.2.2 Modelo de Neurônio Artificial 18
2.2.3 Características das Redes Neurais Artificiais 19
2.3 Redes Neurais Recorrentes **20**
2.3.1 *Long Short-Term Memory* 20
2.3.2 Componentes da Célula LSTM 20
2.3.3 *Bidirectional LSTM* 22
2.4 Redes Neurais Convolucionais **23**
2.4.1 Arquitetura das CNNs 23
2.4.1.1 Camada Convolucional 24
2.4.1.2 Camada de *Pooling* 24
2.4.1.3 Camada Totalmente Conectada 25
2.4.2 Inflated 3D ConvNet (I3D) 25
2.4.2.1 Inflação de Arquiteturas 2D para 3D 26
2.4.2.2 Configuração Two-Stream 27
2.5 Redes Convolucionais em Grafos **27**
2.5.1 GCNs para Reconhecimento de Ações Baseado em Esqueleto 28
2.5.2 SL-GCN para Reconhecimento de Língua de Sinais 28
2.6 Arquitetura do SL-GCN **29**
2.7 Transferência de Aprendizado **31**
2.7.1 Domínio 31
2.7.2 Tarefa 32
2.7.3 Domínio Fonte e Domínio Alvo 32
2.7.4 Aprendizado por Transferência 32

2.7.4.1 Definição dos Domínios . . . 32
2.7.4.2 Definição das Tarefas . . . 33
2.7.5 Transferência Positiva e Negativa . . . 33
2.7.6 Ajuste Fino . . . 33
2.7.6.1 Estratégias de Ajuste Fino . . . 34
2.8 Trabalhos Relacionados . . . 34

3 METODOLOGIA . . . 38
3.1 Base de Dados . . . 38
3.2 Métricas de Avaliação . . . 41
3.2.1 Acurácia Top-k . . . 41
3.2.2 Matriz de Confusão . . . 41
3.3 Modelos . . . 42
3.3.1 I3D . . . 43
3.3.2 Bi-LSTM e SL-GCN . . . 44

4 RESULTADOS . . . 47
4.1 I3D . . . 47
4.2 Bi-LSTM . . . 48
4.3 SL-GCN . . . 49
4.4 Comparação entre Modelos RGB e Baseados em Poses . . . 51
4.5 Análise Comparativa de Transferência entre Línguas de Sinais . . . 52
4.6 Análise das Matrizes de Confusão . . . 53

5 CONCLUSÃO . . . 60
5.1 Limitações e Trabalhos Futuros . . . 61

REFERÊNCIAS . . . 63

Lista de ilustrações

Figura 1 – Modelo de um neurônio artificial. 19
Figura 2 – Representação esquemática da unidade LSTM, com seus componentes principais: portas de esquecimento (f_t), entrada (i_t, g_t) e saída (o_t), e a atualização do estado da célula (C_t) e do estado oculto (h_t). 22
Figura 3 – Arquitetura de uma Bi-LSTM. 22
Figura 4 – Arquitetura da rede AlexNet para classificação de imagens. 24
Figura 5 – Arquitetura do I3D (*Inflated Inception-V1*). À esquerda, a estrutura completa da rede com indicação dos campos receptivos. À direita, detalhamento de um módulo Inception inflado para 3D. 26
Figura 6 – Visão geral da arquitetura do modelo: (a) vídeo de entrada, (b) extração do grafo espacial do esqueleto, (c) processamento multi-fluxo com SL-GCN, (d) bloco SL-GCN com módulos de atenção, (e) módulo STC de atenção espacial, temporal e por canal. 30
Figura 7 – Pipeline da metodologia 38
Figura 8 – Quadro ilustrativo representando um exemplo de cada conjunto de dados utilizado: Dicionário Online Libras-Português, Libras SignBank, VLIBRASIL e MINDS-Libras. 41
Figura 9 – Exemplo de extração de pontos-chave utilizados no pré-processamento dos vídeos (MINDS-Libras Dataset). 45
Figura 10 – Matriz de confusão do modelo Bi-LSTM (WLASL-2000). 54
Figura 11 – Matriz de confusão do modelo SL-GCN (GSL). 55
Figura 12 – Matriz de confusão do modelo I3D (WLASL-1000). 56
Figura 13 – Comparação entre os sinais “Conhecer” e “Vontade” com pontos-chave destacados (MINDS-Libras Dataset). 57
Figura 14 – Comparação entre os sinais “Piano” e “Canguru” com pontos-chave destacados (Libras SignBank). 58
Figura 15 – Comparação entre os sinais “Dinheiro”, “Abraçar” e “Comer” com pontos-chave destacados (V-LIBRASIL). 59

Lista de tabelas

Tabela 1 – Distribuição das classes do *dataset*: identificador, palavra correspondente, número total de vídeos e respectivas quantidades de treino, validação e teste. 40
Tabela 2 – Hiperparâmetros utilizados no treinamento dos modelos. 43
Tabela 3 – Resultados de acurácia Top-k para a arquitetura I3D com diferentes pesos pré-treinados e *fine-tuning* completo 48
Tabela 4 – Resultados de acurácia Top-k para a arquitetura Bi-LSTM com diferentes pesos pré-treinados e *fine-tuning* completo 49
Tabela 5 – Resultados de acurácia Top-k para a arquitetura SL-GCN com diferentes pesos pré-treinados e *fine-tuning* completo 50

Lista de abreviaturas e siglas

2D	Duas dimensões
3D	Três dimensões
Adam	*Adaptive Moment Estimation*
ASL	*American Sign Language* (Língua Americana de Sinais)
Bi-LSTM	*Bidirectional Long Short-Term Memory*
CNN	*Convolutional Neural Network* (Rede Neural Convolucional)
CSL	*Chinese Sign Language* (Língua Chinesa de Sinais)
GCN	*Graph Convolutional Network* (Rede Convolucional em Grafos)
GSL	*Greek Sign Language* (Língua Grega de Sinais)
I3D	*Inflated 3D Convolutional Network*
IA	Inteligência Artificial
ISL	*Indian Sign Language* (Língua Indiana de Sinais)
LIBRAS	Língua Brasileira de Sinais
LSA	*Lengua de Señas Argentina* (Língua de Sinais Argentina)
LSTM	*Long Short-Term Memory*
ReLU	*Rectified Linear Unit* (Unidade Linear Retificada)
RGB	*Red, Green, Blue* (Vermelho, Verde e Azul)
RNA	Rede Neural Artificial
RNN	*Recurrent Neural Network* (Rede Neural Recorrente)
SL-GCN	*Skeleton-based Graph Convolutional Network* (Rede Convolucional em Grafos baseada em Esqueleto)
ST-GCN	*Spatial Temporal Graph Convolutional Network* (Rede Convolucional em Grafos Espaço-Temporal)
TSL	*Turkish Sign Language* (Língua Turca de Sinais)

1 Introdução

Conversar, expressar sentimentos e interagir com o mundo são aspectos centrais da experiência humana. No entanto, para pessoas com deficiência auditiva, essa interação pode ser comprometida. A impossibilidade de ouvir, muitas vezes, impede a participação plena em conversas e atividades sociais, resultando em sentimentos de exclusão, solidão e frustração (FAROOQ et al., 2021). Segundo a Organização Mundial da Saúde, mais de 430 milhões de indivíduos em todo o mundo, entre eles 34 milhões de crianças, necessitam de reabilitação devido à perda auditiva incapacitante, o que representa mais de 5% da população global. Estima-se que, até 2050, esse número ultrapasse 700 milhões, ou seja, uma em cada dez pessoas enfrentará dificuldades auditivas significativas (Organização Mundial da Saúde, 2025).

Nesse cenário, o desenvolvimento de sistemas capazes de reconhecer a linguagem de sinais apenas por meio dos gestos das mãos, sem a necessidade de equipamentos especializados, é fundamental para promover uma comunicação mais equitativa entre pessoas surdas e ouvintes. Embora as técnicas de visão computacional e aprendizado de máquina tenham avançado com o suporte de grandes volumes de dados coletados por dispositivos modernos, esse progresso não se estende a todas as línguas de maneira uniforme. As línguas de sinais, em especial, enfrentam a carência de bases de dados robustas, o que limita o acesso das comunidades surdas a tecnologias de interação, transcrição e tradução automatizadas (MOCIALOV; TURNER; HASTIE, 2018).

A LIBRAS exemplifica esse desafio, pois a falta de conjuntos de dados adequados compromete o avanço de modelos de reconhecimento automático de sinais e limita o desenvolvimento de aplicações que poderiam efetivamente promover a inclusão digital. Uma alternativa promissora para contornar essa limitação é o uso da transferência de aprendizado. Essa técnica consiste em aproveitar o conhecimento adquirido por modelos treinados em tarefas relacionadas, adaptando-os para novos contextos com menor necessidade de dados específicos. Modelos de redes neurais pré-treinados em outras línguas de sinais, como a ASL (SHARMA; KUMAR, 2021) e a CSL (JIANG et al., 2024), podem capturar padrões gestuais e visuais que se mostram transferíveis para a LIBRAS. Essa abordagem permite reduzir a dependência de grandes volumes de dados rotulados e diminuir os custos computacionais em comparação com o treinamento realizado do zero.

Apesar dos avanços que essa estratégia oferece, ainda existem limitações significativas. A diversidade restrita dos conjuntos de dados em LIBRAS dificulta a generalização dos modelos. Além disso, diferenças linguísticas e culturais, como estruturas gramaticais, gestos e expressões faciais, reduzem a eficácia da adaptação direta de modelos estrangeiros. Soma-se a isso a ausência de padronização nas bases de dados disponíveis, que variam quanto à qualidade

dos vídeos, formatos e critérios de anotação.

Outro ponto crítico é que a aplicação da transferência de aprendizado ao contexto da LIBRAS ainda é pouco explorada na literatura. Grande parte das pesquisas concentra-se em poucas línguas de sinais, o que limita a criação de soluções mais universais. Poucas investigações analisam a similaridade entre diferentes bases, os ajustes necessários nas arquiteturas existentes e os impactos dessas adaptações no desempenho em tarefas de reconhecimento e tradução.

Diante desse quadro, esta monografia busca investigar a transferibilidade de modelos pré-treinados em línguas estrangeiras de sinais para o reconhecimento da LIBRAS, identificando seus limites, possibilidades e condições para uma adaptação eficaz. O objetivo é contribuir para o avanço da comunicação acessível e para a promoção da inclusão digital da comunidade surda no Brasil.

1.1 Objetivo

O presente trabalho de conclusão de curso tem como objetivo geral avaliar a eficácia do aprendizado por transferência no reconhecimento automático de LIBRAS, utilizando modelos pré-treinados em outras línguas de sinais. Busca-se compreender como conhecimentos adquiridos em contextos linguísticos distintos podem ser adaptados à língua de sinais brasileira, investigando diferentes arquiteturas de redes neurais e analisando o impacto dos pesos pré-treinados no desempenho dos modelos.

Para alcançar esse objetivo geral, foram estabelecidos os seguintes objetivos específicos:

(i) Estudar trabalhos da literatura sobre reconhecimento automático de línguas de sinais, identificando arquiteturas, técnicas e abordagens consolidadas na área;

(ii) Avaliar modelos de redes neurais pré-treinados em línguas de sinais estrangeiras, analisando como pesos provenientes de contextos linguísticos diversos influenciam o reconhecimento de LIBRAS;

(iii) Comparar o desempenho de diferentes arquiteturas de redes neurais no reconhecimento de sinais, considerando distintos tipos de representação de entrada;

(iv) Examinar padrões de confusão entre sinais através da análise de matrizes de confusão, identificando características visuais e estruturais que contribuem para acertos ou erros no reconhecimento automático.

1.2 Visão Geral do Método e Justificativa

Esta monografia adota uma abordagem experimental baseada em técnicas de transferência de aprendizado e redes neurais, aplicadas à adaptação de modelos previamente desenvolvidos para outras línguas de sinais. Essa escolha metodológica permite explorar a possibilidade de reutilizar modelos treinados, considerando a escassez de dados robustos para LIBRAS.

A transferência de aprendizado é especialmente relevante neste estudo, pois possibilita que modelos treinados em grandes corpora sejam ajustados à LIBRAS mesmo com um número reduzido de exemplos. Essa estratégia reduz a necessidade de bases extensas e dispensa o treinamento completo de arquiteturas complexas, processo normalmente custoso e demorado.

As redes neurais foram selecionadas devido à sua capacidade de aprender representações complexas e hierárquicas, essenciais para o reconhecimento de sinais, que envolve variações espaciais, temporais e expressivas. Além disso, já se consolidaram como ferramentas eficazes em tarefas de visão computacional, como análise de vídeos, alinhando-se às características multimodais dos dados em LIBRAS.

Dessa forma, a metodologia proposta não apenas avalia a viabilidade técnica da adaptação, mas também contribui para a ampliação da acessibilidade e inclusão digital da comunidade surda por meio de tecnologias avançadas.

1.3 Estrutura deste Trabalho

Este estudo está organizado em cinco capítulos. O Capítulo 1 é introdutório e apresenta o contexto, a motivação e os objetivos da pesquisa. O Capítulo 2 reúne uma revisão das principais pesquisas sobre reconhecimento de línguas de sinais e aprendizado por transferência. O Capítulo 3 descreve os métodos de adaptação de modelos e a preparação dos dados. O Capítulo 4 apresenta e discute os resultados obtidos. Por fim, o Capítulo 5 sintetiza as contribuições do trabalho e aponta direções para futuras investigações.

2 Fundamentação e Revisão da Literatura

A seguir, apresentam-se os fundamentos teóricos e os trabalhos relacionados que sustentam o desenvolvimento desta pesquisa. Inicialmente, faz-se uma contextualização sobre Inteligência Artificial e Aprendizado de Máquina; em seguida, apresentam-se as Redes Neurais Artificiais e suas características fundamentais. Posteriormente, são abordadas as arquiteturas de aprendizado profundo relevantes para esta pesquisa: Redes Neurais Recorrentes, Redes Neurais Convolucionais e Redes Convolucionais em Grafos. Também são apresentados os conceitos de Transferência de Aprendizado e Ajuste Fino, técnicas essenciais para adaptar modelos pré-treinados a novos domínios. Por fim, realiza-se uma revisão dos principais trabalhos relacionados ao reconhecimento automático de línguas de sinais, abordando a evolução histórica da área, as abordagens mais recentes baseadas em aprendizado profundo e os avanços específicos no contexto da LIBRAS.

2.1 Inteligência Artificial e Aprendizado de Máquina

A Inteligência Artificial é um campo complexo e multidimensional, cuja definição exata permanece desafiadora, pois envolve dois conceitos profundamente interligados: o que entendemos por "artificial"e o que significa "inteligência". De forma geral, IA refere-se a sistemas criados por seres humanos com o objetivo de replicar propriedades e capacidades associadas à inteligência biológica. Esses sistemas conseguem simular processos como raciocínio, aprendizado e tomada de decisão, embora sua base seja totalmente tecnológica e não biológica (FETZER, 1990).

No âmbito da ciência da computação, a IA investiga como sistemas computacionais podem exibir comportamentos considerados inteligentes, ou seja, sistemas capazes de captar informações do ambiente, processá-las de forma autônoma e executar ações direcionadas por objetivos específicos. Essa definição abrange habilidades como aprender com a experiência, raciocinar logicamente e resolver problemas de forma autônoma. Mais do que executar tarefas predefinidas, a IA busca desenvolver a capacidade de adaptação e aprendizado contínuo, atributos essenciais para sua aplicação em contextos diversos e dinâmicos (ONGSULEE, 2017).

Dentre os subcampos da IA, destaca-se o Aprendizado de Máquina, que se dedica ao desenvolvimento de sistemas capazes de aprender a partir de dados, sem a necessidade de instruções explícitas para cada tarefa. Por meio de algoritmos que identificam padrões e extraem conhecimento de grandes volumes de dados, o aprendizado de máquina permite que os sistemas aprimorem progressivamente seu desempenho à medida que são expostos a novas informações (MAHESH et al., 2020).

2.2 Redes Neurais Artificiais

As RNAs são modelos computacionais baseados na organização e nos mecanismos do cérebro humano. Elas processam informações de forma não linear e paralela, mostrando-se especialmente eficazes em tarefas como reconhecimento de padrões, classificação e controle. Enquanto o cérebro executa essas funções por meio de uma complexa rede de neurônios biológicos, as RNAs utilizam unidades de processamento simplificadas, chamadas neurônios artificiais, interligadas por conexões ponderadas conhecidas como pesos sinápticos (HAYKIN, 2009).

Durante o processo de aprendizado, elas ajustam os pesos sinápticos com base em exemplos fornecidos, permitindo que a rede identifique padrões e generalize para novos dados. As RNAs também possuem capacidade de adaptação, auto-organização e armazenamento do conhecimento adquirido, conferindo-lhes flexibilidade na resolução de problemas complexos (WU; FENG, 2018).

2.2.1 Funcionamento de um Neurônio Artificial

Haykin (2009) explica que o neurônio artificial é formado por três elementos principais: sinapses, somador e função de ativação.

1. **Sinapses**: correspondem às ligações estabelecidas entre os neurônios artificiais. A cada conexão sináptica é atribuído um peso w_{kj}, que quantifica a magnitude da contribuição da entrada x_j no processamento do neurônio k. Esses pesos podem ser positivos ou negativos.

2. **Somador**: realiza a combinação linear das entradas ponderadas por seus pesos associados. O valor calculado é dado por:

$$u_k = \sum_{j=1}^{m} w_{kj} x_j$$

3. **Função de ativação**: responsável por restringir a amplitude da saída do neurônio a um intervalo definido, como $[0, 1]$ ou $[-1, 1]$. A saída do neurônio y_k é calculada por:

$$y_k = \varphi(u_k + b_k)$$

   onde b_k é o viés (*bias*), um parâmetro adicional que permite deslocar a função de ativação, aumentando a flexibilidade do modelo.

2.2.2 Modelo de Neurônio Artificial

Neste modelo, o neurônio artificial recebe um conjunto de entradas $x_1, x_2, \ldots, x_m$, cada uma com um peso sináptico correspondente $w_{k1}, w_{k2}, \ldots, w_{km}$.

Para fins de simplificação, é comum incorporar o viés como um peso adicional $w_{k0} = b_k$, associado a uma entrada constante $x_0 = +1$. Assim, a soma ponderada pode ser expressa de forma compacta como:

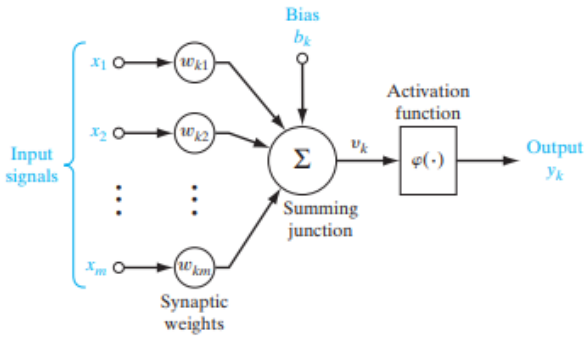
$$v_k = \sum_{j=0}^{m} w_{kj} x_j \tag{1}$$

Em seguida, essa soma é passada por uma função de ativação $\varphi(\cdot)$, que introduz não linearidade ao modelo:

$$y_k = \varphi(v_k) \tag{2}$$

A Figura 1 ilustra esquematicamente a estrutura de um neurônio artificial, mostrando as entradas, os pesos sinápticos, o somador e a função de ativação.

Figura 1 – Modelo de um neurônio artificial.



Fonte: Haykin (2009)

2.2.3 Características das Redes Neurais Artificiais

As RNAs são compostas por grandes quantidades de neurônios artificiais interconectados e organizados em camadas. Embora não ofereçam soluções exatas no sentido matemático tradicional, demonstram notável capacidade de generalização, adaptabilidade e processamento paralelo.

Quatro características fundamentais justificam o uso das redes neurais em aplicações de alta complexidade (DONGARE et al., 2012; WU; FENG, 2018):

- **Não-linearidade**: permite lidar com relações complexas e não lineares entre variáveis, simulando o comportamento do cérebro humano.
- **Distribuição da informação**: o comportamento global da rede emerge da interação entre seus neurônios, sem depender exclusivamente de uma única unidade de processamento.
- **Adaptabilidade**: a rede pode reorganizar internamente suas conexões com base na experiência, promovendo aprendizado contínuo.

- **Multiplicidade de soluções**: a rede pode atingir diferentes estados estáveis, favorecendo soluções variadas para um mesmo problema, característica relacionada à não-convexidade do espaço de soluções.

Essas propriedades tornam as RNAs especialmente eficazes em tarefas como mapeamento de funções não lineares, reconhecimento de padrões, processamento de imagens e tomada de decisão em contextos complexos.

2.3 Redes Neurais Recorrentes

As RNNs são um tipo de rede neural projetada para trabalhar com dados sequenciais, possuindo a capacidade de lembrar informações de etapas anteriores por meio de conexões recorrentes entre unidades ocultas (AL-SELWI et al., 2024). Essa característica permite que as RNNs capturem dependências temporais e identifiquem correlações entre eventos distantes na sequência de dados. Por essa razão, elas são especialmente eficazes em tarefas que dependem de contexto, como reconhecimento de fala, classificação de vídeos e processamento de linguagem natural (PERUMAL et al., 2024).

2.3.1 *Long Short-Term Memory*

As RNNs são eficazes, mas enfrentam um problema conhecido como "desaparecimento do gradiente", que dificulta o aprendizado de dependências de longo prazo em sequências extensas. Para contornar essa limitação, Hochreiter e Schmidhuber (1997) propuseram a arquitetura *long short-term memory*, uma versão especializada de RNNs. O LSTM utiliza unidades ocultas com memória interna, capazes de reter informações por períodos prolongados, o que permite à rede capturar padrões temporais de longa duração de forma mais eficiente.

2.3.2 Componentes da Célula LSTM

A arquitetura LSTM é composta por uma célula de memória e três portas principais (*gates*) que regulam o fluxo de informações (AL-SELWI et al., 2024). A seguir, cada componente será descrito, juntamente com suas funções e fórmulas correspondentes.

Porta de Esquecimento (Forget Gate). A porta de esquecimento atua como um filtro seletivo que determina quais informações do estado anterior da célula devem ser descartadas. Esta porta avalia a relevância das informações armazenadas com base na entrada atual e no estado oculto anterior, decidindo quais dados são obsoletos ou irrelevantes para o processamento da sequência atual (AL-SELWI et al., 2024).

A ativação do *forget Gate* é calculada por (SMAGULOVA; JAMES, 2019):

$$f_t = \sigma\Big(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}\Big) \tag{3}$$

Porta de Entrada (Input Gate). A porta de entrada opera em duas etapas complementares: primeiro, identifica quais novas informações merecem ser armazenadas no estado da célula, e depois determina os valores que serão efetivamente adicionados. Esta porta funciona como um mecanismo de proteção, filtrando entradas irrelevantes que poderiam contaminar ou corromper as informações importantes já armazenadas na memória da célula (AL-SELWI et al., 2024).

O cálculo do *input gate* e da informação candidata à célula é dado por (SMAGULOVA; JAMES, 2019):

$$i_t = \sigma\Big(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}\Big) \tag{4}$$

$$g_t = \tanh\Big(W^{(g)}x_t + U^{(g)}h_{t-1} + b^{(g)}\Big) \tag{5}$$

A atualização do estado da célula C_t combina a memória antiga filtrada pelo *forget gate* e a nova informação filtrada pelo *input gate*:

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t \tag{6}$$

Porta de Saída (Output Gate). A porta de saída exerce controle refinado sobre quais componentes do estado da célula influenciarão a saída final da unidade LSTM. Esta porta não apenas determina o que será transmitido para a próxima camada ou etapa temporal, mas também protege outras unidades da rede contra interferências causadas por informações de memória que podem ser relevantes para armazenamento interno, mas inadequadas para a saída no contexto atual (AL-SELWI et al., 2024).

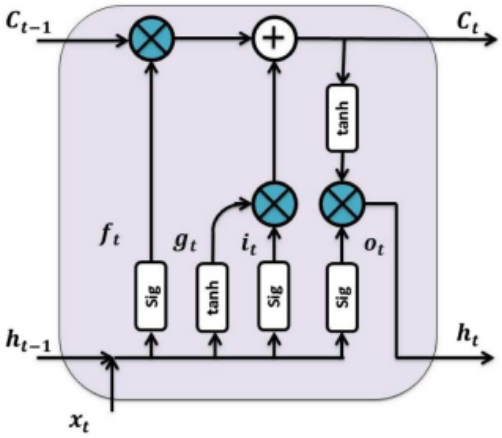
O cálculo do *output gate* e da saída da célula é dado por (SMAGULOVA; JAMES, 2019):

$$o_t = \sigma\Big(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}\Big) \tag{7}$$

$$h_t = o_t \odot \tanh(C_t) \tag{8}$$

A Figura 2 apresenta uma representação esquemática da unidade LSTM, ilustrando o funcionamento integrado de suas três portas e o processo de atualização do estado da célula e do estado oculto.

Figura 2 – Representação esquemática da unidade LSTM, com seus componentes principais: portas de esquecimento (f_t), entrada (i_t, g_t) e saída (o_t), e a atualização do estado da célula (C_t) e do estado oculto (h_t).



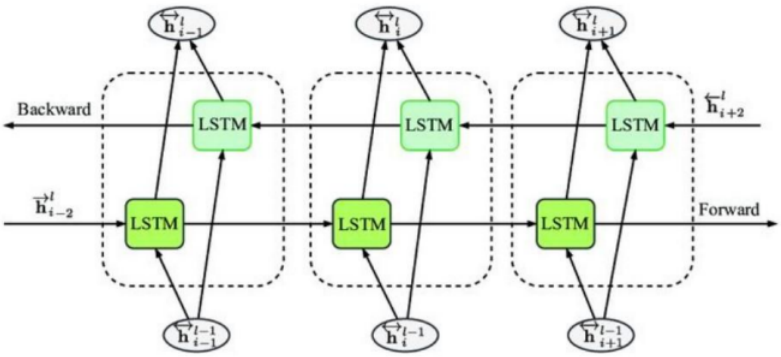
Fonte: Smagulova e James (2019)

2.3.3 *Bidirectional LSTM*

A Bi-LSTM é uma variação da arquitetura LSTM que considera tanto o contexto passado quanto o futuro em tarefas de modelagem de sequências. Diferentemente das LSTMs tradicionais, que processam os dados apenas na direção direta, a Bi-LSTM utiliza duas camadas LSTM paralelas: uma lê a sequência da esquerda para a direita e a outra da direita para a esquerda. Esse processamento bidirecional permite capturar dependências temporais de forma mais completa. Durante o treinamento, cada camada extrai características independentemente, e seus resultados são combinados para gerar a saída final, ampliando a capacidade do modelo de compreender o contexto da sequência (PERUMAL et al., 2024).

A Figura 3 ilustra essa arquitetura, mostrando as camadas LSTM processando a sequência em ambas as direções.

Figura 3 – Arquitetura de uma Bi-LSTM.



Fonte: Perumal et al. (2024)

2.4 Redes Neurais Convolucionais

As CNNs, conforme descrito por O'shea e Nash (2015), são uma arquitetura de RNA especialmente projetada para tarefas de reconhecimento de padrões em imagens. Como outras redes neurais, são compostas por neurônios artificiais que aprendem ajustando seus pesos sinápticos, porém incorporam características estruturais dos dados visuais em sua arquitetura. Essa adaptação as torna mais eficientes em aplicações visuais, ao mesmo tempo em que reduz a quantidade de parâmetros do modelo.

Diferentemente das RNAs tradicionais, as CNNs organizam seus neurônios em múltiplas dimensões (altura, largura e, quando aplicável, profundidade temporal). Os neurônios em cada camada são conectados apenas a pequenas regiões da camada anterior, denominadas campos receptivos (LECUN et al., 1998), permitindo o processamento local de padrões. Assim, os dados de entrada são progressivamente transformados através de múltiplas camadas convolucionais e de redução dimensional, extraindo características cada vez mais abstratas, até gerar uma representação condensada que pode ser utilizada para classificação ou outras tarefas.

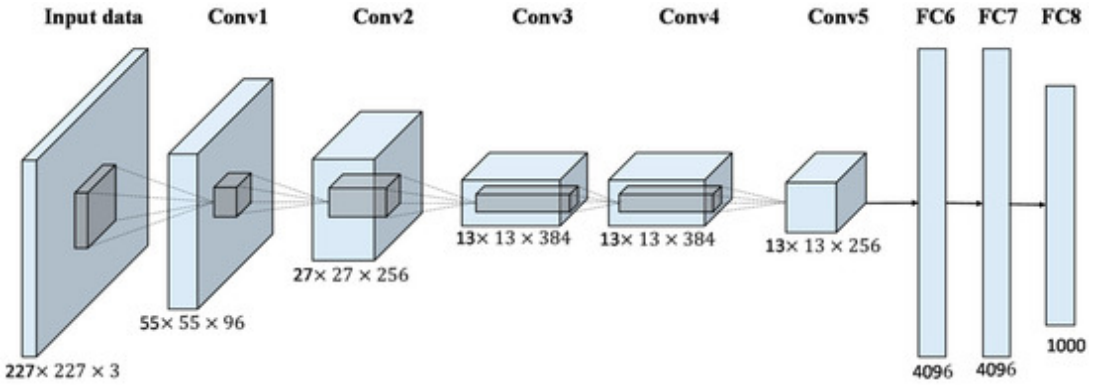
2.4.1 Arquitetura das CNNs

A arquitetura de uma CNN é composta, de modo geral, por três tipos principais de camadas (O'SHEA; NASH, 2015):

- **Camadas convolucionais**: responsáveis pela extração de características locais da imagem;
- **Camadas de pooling**: encarregadas de reduzir a dimensionalidade espacial;
- **Camadas totalmente conectadas**: produzem a saída final do modelo, geralmente associada à classificação.

A Figura 4 apresenta a arquitetura da AlexNet, uma das CNNs mais influentes na história do reconhecimento de imagens. Essa rede é composta por cinco camadas convolucionais seguidas por três camadas totalmente conectadas. As primeiras camadas realizam a extração de padrões visuais, como bordas e texturas, enquanto as camadas finais transformam essas representações em pontuações de classe para a tarefa de classificação.

Figura 4 – Arquitetura da rede AlexNet para classificação de imagens.



Fonte: Han et al. (2017)

2.4.1.1 Camada Convolucional

A camada convolucional aplica filtros (ou *kernels*) aprendíveis sobre regiões locais da imagem de entrada, gerando mapas de ativação que destacam padrões visuais específicos, como bordas ou texturas. Cada filtro gera um mapa distinto, e a sobreposição de diversos filtros forma o volume de saída da camada.

Essa operação possui três características fundamentais:

- **Conexões locais**: cada neurônio se conecta apenas a uma área limitada da entrada;
- **Compartilhamento de pesos**: o mesmo filtro é aplicado a diferentes regiões da entrada;
- **Hiperparâmetros principais**:
  - **Profundidade**: número de filtros aplicados, que determina o número de mapas de ativação;
  - **Stride**: o número de pixels pelos quais o filtro se desloca em cada passo;
  - **Zero-padding**: preenchimento das bordas com zeros, usado para controlar o tamanho da saída.

Esses hiperparâmetros influenciam diretamente o tamanho da saída da camada. O uso de filtros pequenos, aplicados de forma compartilhada, permite detectar padrões relevantes com menor número de parâmetros, o que é crucial para o processamento eficiente de imagens em alta resolução (O'SHEA; NASH, 2015; YAMASHITA et al., 2018; KHAN et al., 2020).

2.4.1.2 Camada de *Pooling*

A camada de *pooling* realiza um processo de subamostragem que reduz a resolução espacial dos mapas de ativação. A técnica mais frequente é o *max-pooling*, que extrai o valor

máximo de regiões específicas da entrada (geralmente de tamanho 2×2 com *stride* igual a 2).

Essa redução espacial contribui para:

- Reduzir a quantidade de parâmetros treináveis subsequentes e o custo computacional;
- Mitigar o risco de *overfitting*;
- Preservar as características mais relevantes.

A profundidade da saída permanece inalterada. Embora seja possível utilizar regiões maiores ou sobreposição entre janelas de pooling, isso deve ser feito com cautela, para evitar a perda de informações importantes (O'SHEA; NASH, 2015; YAMASHITA et al., 2018; KHAN et al., 2020).

2.4.1.3 Camada Totalmente Conectada

As camadas totalmente conectadas (fully connected) funcionam de maneira análoga às redes neurais artificiais tradicionais: cada neurônio é conectado a todos os neurônios da camada anterior. Essa estrutura tem como função consolidar as ativações extraídas pelas camadas anteriores e transformá-las em pontuações de classe (no caso de classificação) ou em valores contínuos (no caso de regressão).

Apesar de sua eficácia em problemas com imagens pequenas, o uso direto dessas conexões em imagens maiores pode resultar em um número extremamente alto de parâmetros. As CNNs contornam esse problema utilizando, previamente, as camadas convolucionais e de pooling, que extraem e condensam as informações visuais antes da aplicação das camadas densas (O'SHEA; NASH, 2015; YAMASHITA et al., 2018; KHAN et al., 2020).

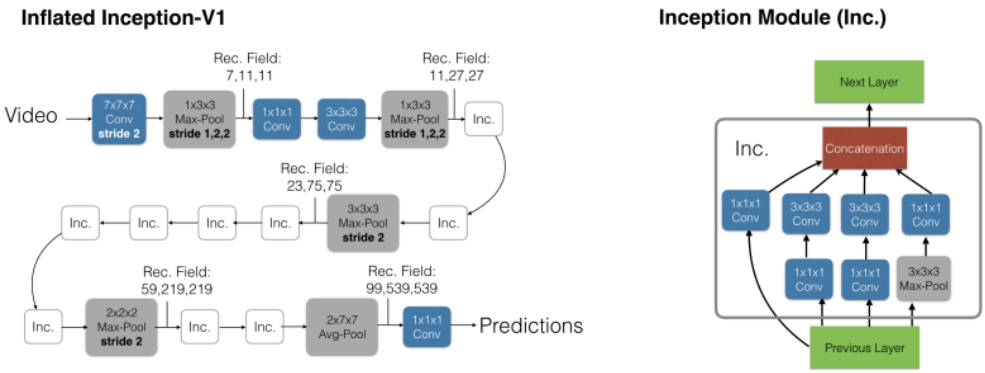
2.4.2 Inflated 3D ConvNet (I3D)

O I3D é uma arquitetura de rede neural proposta por Carreira e Zisserman (2017), projetada para reconhecimento de ações em vídeos. A arquitetura expande redes convolucionais 2D tradicionais pré-treinadas em ImageNet ao incorporar a dimensão temporal, permitindo que a rede aproveite tanto o design quanto os pesos já aprendidos em imagens estáticas para acelerar o treinamento e melhorar a detecção de padrões em sequências de vídeo.

O modelo adota uma configuração de dois fluxos (*two-stream*), conceito originalmente proposto por Simonyan e Zisserman (2014): o primeiro fluxo utiliza *frames* RGB para aprender informações espaciais sobre a aparência dos objetos, enquanto o segundo utiliza fluxo óptico para capturar informações temporais sobre o movimento. Cada fluxo realiza o reconhecimento de vídeo de forma independente e, para a classificação final, as predições são combinadas por fusão tardia (*late fusion*) (CARREIRA; ZISSERMAN, 2017).

A Figura 5 apresenta a estrutura do I3D baseado na arquitetura Inception-V1 inflada para 3D.

Figura 5 – Arquitetura do I3D (*Inflated Inception-V1*). À esquerda, a estrutura completa da rede com indicação dos campos receptivos. À direita, detalhamento de um módulo Inception inflado para 3D.



Fonte: Carreira e Zisserman (2017)

2.4.2.1 Inflação de Arquiteturas 2D para 3D

Em vez de projetar uma arquitetura espaço-temporal desde o início, o I3D converte modelos bem-sucedidos de classificação de imagens 2D em redes convolucionais 3D. Esse processo de "inflação" consiste em adicionar uma dimensão temporal a todos os filtros e *kernels* de *pooling* da arquitetura original, transformando operações bidimensionais em operações tridimensionais que processam simultaneamente informações espaciais e temporais.

O I3D utiliza como base a arquitetura Inception-v1, uma rede convolucional amplamente reconhecida por sua eficiência em classificação de imagens. Essa escolha permite herdar tanto a topologia otimizada quanto os parâmetros aprendidos no conjunto de dados ImageNet (CARREIRA; ZISSERMAN, 2017).

Inicialização de Pesos

Para aproveitar os parâmetros pré-treinados de modelos 2D, o I3D emprega uma estratégia de inicialização que replica os pesos dos filtros ao longo da dimensão temporal e os reescala adequadamente. Essa abordagem garante que, ao processar um vídeo onde todos os frames são idênticos, o modelo 3D produza ativações equivalentes às do modelo 2D original processando uma única imagem. Devido à linearidade das convoluções, essa propriedade é preservada e facilita o aproveitamento do conhecimento pré-treinado (CARREIRA; ZISSERMAN, 2017).

Crescimento do Campo Receptivo

Um aspecto importante no design do I3D é o controle do crescimento do campo

receptivo ao longo das dimensões espaciais e temporal. Um crescimento temporal inadequado pode comprometer tanto a detecção de características iniciais quanto a captura da dinâmica temporal da cena. Para equilibrar esse crescimento, o I3D utiliza configurações específicas de *stride* e tamanho de *kernel* ao longo da arquitetura, controlando quando e como a informação temporal é agregada (CARREIRA; ZISSERMAN, 2017).

2.4.2.2 Configuração Two-Stream

Embora uma ConvNet 3D seja capaz de aprender características de movimento diretamente de entradas RGB, o uso de fluxo óptico pré-computado fornece uma representação explícita do movimento que complementa as informações visuais. Experimentalmente, constatou-se que a configuração *two-stream* proporciona benefícios significativos ao desempenho do modelo.

O I3D adota, portanto, duas redes separadas:

- **RGB I3D**: Treinada com entradas RGB, capturando informações sobre aparência e contexto espacial;
- **Flow I3D**: Treinada com entradas de fluxo óptico, capturando informações sobre o movimento entre frames.

As duas redes são treinadas de forma independente e suas predições são combinadas durante a inferência. O modelo processa segmentos de vídeo durante o treinamento e agrega predições ao longo do tempo durante a avaliação (CARREIRA; ZISSERMAN, 2017).

2.5 Redes Convolucionais em Grafos

As GCNs são redes neurais projetadas para trabalhar com dados estruturados em grafos, onde cada nó pode ter um número variável de vizinhos. Proposta por Kipf e Welling (2017), essa arquitetura permite que o modelo propague e agregue informações entre os nós, combinando as características próprias de cada nó com as de seus vizinhos, capturando tanto informações locais quanto a estrutura global do grafo.

A operação fundamental de uma GCN pode ser expressa como uma propagação de camada que atualiza as representações dos nós com base em suas vizinhanças. Essa capacidade de processar estruturas irregulares torna as GCNs particularmente adequadas para modelar o esqueleto humano, onde as conexões entre juntas corporais formam naturalmente um grafo (YAN; XIONG; LIN, 2018).

2.5.1 GCNs para Reconhecimento de Ações Baseado em Esqueleto

A aplicação de GCNs para modelagem de esqueleto humano foi pioneiramente proposta por Yan, Xiong e Lin (2018) no ST-GCN (*Spatial Temporal Graph Convolutional Network*), que demonstrou a eficácia dessa abordagem para reconhecimento de ações. O ST-GCN modela o esqueleto humano como um grafo espaço-temporal, onde cada junta representa um nó e as arestas codificam tanto conexões anatômicas quanto relações temporais entre *frames* consecutivos.

Posteriormente, Shi et al. (2019) expandiram essa ideia com o 2s-AGCN (*Two-Stream Adaptive Graph Convolutional Network*), uma arquitetura multi-fluxo que processa informações de juntas (*joints*) e ossos (*bones*) separadamente, permitindo capturar tanto a configuração espacial das articulações quanto as relações geométricas entre membros conectados. Essa abordagem multi-fluxo demonstrou melhorias significativas na capacidade discriminativa do modelo.

A incorporação de mecanismos de atenção em GCNs foi explorada por Si et al. (2019), que propuseram atenção espacial e temporal para identificar quais juntas e quais momentos da sequência são mais relevantes para a classificação. Esses avanços estabeleceram as bases para arquiteturas mais sofisticadas voltadas ao reconhecimento de língua de sinais.

2.5.2 SL-GCN para Reconhecimento de Língua de Sinais

Adaptando essas arquiteturas para o domínio específico de língua de sinais, Jiang et al. (2021) propõem o SL-GCN (*Sign Language Graph Convolutional Network*), um modelo que utiliza mecanismos de atenção para modelar o esqueleto humano em um grafo espaço-temporal, permitindo capturar a dinâmica dos movimentos no reconhecimento de sinais. Nesse modelo, cada junta corporal é representada como um nó do grafo, e as conexões seguem tanto a estrutura anatômica (braços, mãos, tronco) quanto a continuidade temporal entre *frames*, formando uma representação espaço-temporal dos gestos.

A matriz de adjacência espacial é definida como:

$$A_{i,j} = \begin{cases} 1, & \text{se } d(v_i, v_j) = 1 \\ 0, & \text{caso contrário} \end{cases}$$

em que $d(v_i, v_j)$ representa a menor distância (número de nós no caminho mais curto) entre os nós v_i e v_j.

A operação de convolução em grafos espaciais, baseada na formulação de Kipf e Welling (2017), pode ser escrita como:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

onde $\tilde{A} = A + I_N$ combina a matriz de adjacência das juntas com auto-conexões, $\tilde{D}$ é a matriz diagonal de graus correspondente, $H^{(l)}$ representa as *features* da camada l, $W^{(l)}$ são os pesos treináveis e $\sigma(\cdot)$ é a função de ativação. Já a convolução temporal é implementada como uma convolução 2D com *kernel* de tamanho $k_t \times 1$, capturando as dependências entre *frames*.

Para enriquecer a representação, seguindo a abordagem multi-fluxo de Shi et al. (2019), são considerados diferentes fluxos de informação (JIANG et al., 2021): (i) as posições das juntas (*joint*), (ii) os vetores que conectam juntas vizinhas (*bone*), (iii) as variações das juntas ao longo do tempo (*joint motion*) e (iv) as variações dos ossos entre *frames* (*bone motion*).

As *features* de ossos são calculadas como:

$$v_{j,t}^{B} = (x_{j,t} - x_{i,t},\ y_{j,t} - y_{i,t},\ s_{j,t}),$$

onde $(i, j) \in H$, sendo H o conjunto das conexões naturais do corpo humano, e s representa o grau de confiança da detecção da junta.

As *features* de movimento são obtidas pela diferença entre *frames* consecutivos:

$$v_{i,t}^{JM} = (x_{i,t+1} - x_{i,t},\ y_{i,t+1} - y_{i,t},\ s_{i,t}),$$

$$v_{i,t}^{BM} = v_{i,t+1}^{B} - v_{i,t}^{B}.$$

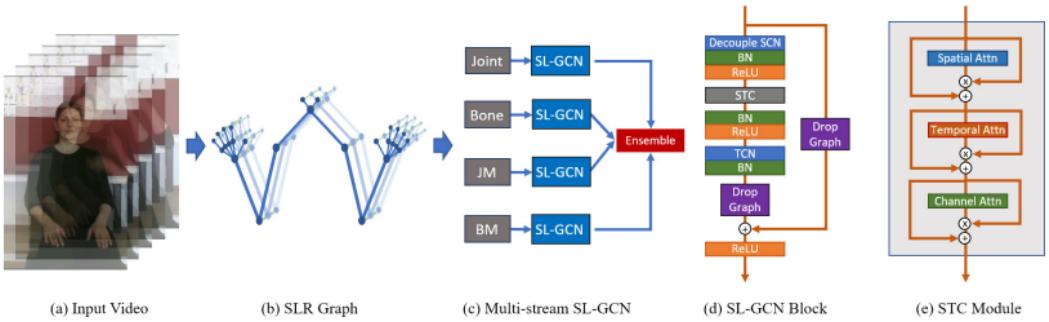
Cada fluxo é processado separadamente e suas saídas são combinadas em uma etapa de *ensemble*, resultando em uma predição mais robusta.

Além disso, o modelo incorpora módulos de atenção espacial, temporal e por canal (*STC Attention*) (SI et al., 2019; JIANG et al., 2021), que permitem priorizar regiões do corpo e momentos da sequência mais relevantes para a interpretação do sinal. Isso é especialmente importante na Língua de Sinais, em que movimentos sutis das mãos e dos braços carregam a maior parte da informação semântica. Assim, o SL-GCN consegue integrar informações locais (configuração das mãos e dedos) e globais (movimento do corpo e postura), alcançando uma representação discriminativa eficaz para o reconhecimento automático de sinais.

2.6 Arquitetura do SL-GCN

A Figura 6 apresenta a arquitetura completa do modelo proposto para reconhecimento de sinais em vídeo, organizada em cinco etapas principais:

Figura 6 – Visão geral da arquitetura do modelo: (a) vídeo de entrada, (b) extração do grafo espacial do esqueleto, (c) processamento multi-fluxo com SL-GCN, (d) bloco SL-GCN com módulos de atenção, (e) módulo STC de atenção espacial, temporal e por canal.



Fonte: Jiang et al. (2021)

(a) Vídeo de Entrada: O processo inicia-se com a aquisição de sequências de vídeo contendo a execução de sinais. Cada vídeo é decomposto em *frames* individuais, dos quais são extraídas as coordenadas espaciais das juntas corporais do sinalizante.

(b) Grafo Espacial do Esqueleto: A partir das coordenadas extraídas, constrói-se um grafo espaço-temporal onde cada nó representa uma junta corporal e as arestas codificam tanto as conexões anatômicas naturais (estrutura do esqueleto humano) quanto as relações temporais entre *frames* consecutivos (YAN; XIONG; LIN, 2018; JIANG et al., 2021). Esta representação em grafo permite capturar simultaneamente a configuração espacial do corpo e a dinâmica temporal dos movimentos.

(c) Multi-stream SL-GCN: O modelo emprega uma arquitetura multi-fluxo que processa quatro modalidades complementares de informação em paralelo: *Joint*, que corresponde às coordenadas absolutas das juntas (x, y, s), onde s representa a confiança da detecção; *Bone*, que representa os vetores que conectam juntas adjacentes e capturam relações geométricas entre membros; *Joint Motion* (JM), que corresponde às diferenças entre posições de juntas em frames consecutivos, representando velocidade e direção do movimento; *Bone Motion* (BM), que corresponde às variações temporais dos vetores de ossos, capturando a dinâmica de deformação do esqueleto. Cada fluxo é alimentado em um bloco SL-GCN independente, permitindo que o modelo aprenda representações especializadas para cada tipo de informação. Na etapa de *ensemble*, as predições dos quatro fluxos são combinadas para produzir a classificação final, agregando informações complementares e aumentando a robustez do reconhecimento (SHI et al., 2019; JIANG et al., 2021).

(d) Bloco SL-GCN: Cada bloco SL-GCN é composto por uma camada de convolução espacial desacoplada (*Decouple* SCN), um módulo de atenção STC, uma camada de convolução temporal (TCN) e um módulo DropGraph. A convolução espacial é implementada utilizando a matriz de adjacência que representa as conexões intra-corporais, enquanto

a convolução temporal é implementada como uma convolução 2D com kernel de tamanho $k_t \times 1$ que opera na dimensão temporal. O modelo proposto consiste em 10 desses blocos GCN empilhados (JIANG et al., 2021).

(e) Módulo STC: O mecanismo de atenção STC (*Spatio-Temporal-Channel Attention*) consiste em módulos de atenção espacial, temporal e por canal conectados em cascata, permitindo que o modelo adapte dinamicamente suas representações internas ao ponderar a relevância de diferentes juntas corporais, momentos temporais e canais de características para o reconhecimento de sinais (JIANG et al., 2021).

A arquitetura resultante combina a capacidade das GCNs de modelar estruturas irregulares (grafos espaço-temporais) (KIPF; WELLING, 2017) com mecanismos de atenção adaptativos (SI et al., 2019) e processamento multi-fluxo (SHI et al., 2019), resultando em um sistema robusto e discriminativo para o reconhecimento automático de Língua de Sinais (JIANG et al., 2021).

2.7 Transferência de Aprendizado

A notação e os conceitos apresentados nesta seção seguem as definições do artigo *A Survey of Transfer Learning* (WEISS; KHOSHGOFTAAR; WANG, 2016) e *A Survey on Transfer Learning* (PAN; YANG, 2010).

2.7.1 Domínio

Formalmente, um domínio D é caracterizado por dois elementos fundamentais: um espaço de características $\mathcal{X}$ e a distribuição de probabilidade marginal $P(X)$ associada a esse espaço (PAN; YANG, 2010; WEISS; KHOSHGOFTAAR; WANG, 2016). Matematicamente, essa relação pode ser expressa como:

$$D = \{\mathcal{X}, P(X)\},$$

em que $X = \{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$ corresponde ao conjunto de amostras observadas. Cada elemento $x_i \in \mathcal{X}$ consiste em um vetor contendo as características descritivas de uma determinada instância.

Em outras palavras, um domínio caracteriza não apenas o tipo de dados disponíveis (o espaço de características $\mathcal{X}$), mas também como esses dados estão distribuídos ($P(X)$). Domínios diferentes podem ter o mesmo tipo de características, mas com distribuições distintas.

2.7.2 Tarefa

Uma tarefa T é composta por um espaço de rótulos $\mathcal{Y}$ e uma função preditiva $f(\cdot)$, definida como (PAN; YANG, 2010):

$$T = \{\mathcal{Y}, f(\cdot)\},$$

onde $f : \mathcal{X} \rightarrow \mathcal{Y}$ é responsável por prever um rótulo $y_i \in \mathcal{Y}$ a partir de uma instância $x_i \in \mathcal{X}$. Do ponto de vista probabilístico, $f(\cdot)$ pode ser interpretada como uma aproximação da distribuição condicional $P(Y|X)$.

Em outras palavras, uma tarefa define o objetivo do aprendizado: determinar quais são os possíveis rótulos e como mapear os dados de entrada para esses rótulos. Tarefas diferentes podem operar sobre o mesmo domínio, mas com objetivos de predição distintos.

2.7.3 Domínio Fonte e Domínio Alvo

O *domínio fonte* D_S contém dados rotulados $\{(x_i^s, y_i^s)\}_{i=1}^{n_s}$, enquanto o *domínio alvo* D_T contém dados $\{(x_i^t, y_i^t)\}_{i=1}^{n_t}$. As tarefas correspondentes são T_S e T_T, com funções preditivas $f_S(\cdot)$ e $f_T(\cdot)$, respectivamente. Tipicamente, $0 \leq n_t \ll n_s$, ou seja, o domínio alvo possui nenhum ou significativamente menos dados rotulados que o domínio fonte (PAN; YANG, 2010).

Em outras palavras, o domínio fonte representa o conjunto de dados onde há abundância de informações rotuladas disponíveis para treinamento, enquanto o domínio alvo é aquele onde se deseja fazer predições, mas há escassez ou ausência de dados rotulados. O objetivo do transfer learning é aproveitar o conhecimento do domínio fonte para melhorar o desempenho no domínio alvo.

2.7.4 Aprendizado por Transferência

O aprendizado por transferência (*transfer learning*) tem como objetivo aprimorar o desempenho da função preditiva do domínio-alvo $f_T(\cdot)$, aproveitando o conhecimento obtido a partir de um domínio-fonte D_S e de sua respectiva tarefa T_S, onde: (WEISS; KHOSHGOFTAAR; WANG, 2016)

$$D_S \neq D_T \quad \text{ou} \quad T_S \neq T_T.$$

2.7.4.1 Definição dos Domínios

Os domínios são definidos como (WEISS; KHOSHGOFTAAR; WANG, 2016):

$$D_S = \{\mathcal{X}_S, P(X_S)\}, \quad D_T = \{\mathcal{X}_T, P(X_T)\},$$

onde $\mathcal{X}$ representa o espaço de características e $P(X)$ a distribuição dos dados.

- Se $\mathcal{X}_S \neq \mathcal{X}_T$, há transferência heterogênea, pois os espaços de características diferem entre os domínios.
- Se $\mathcal{X}_S = \mathcal{X}_T$, ocorre transferência homogênea, caracterizada por diferenças na distribuição marginal dos dados.

2.7.4.2 Definição das Tarefas

As tarefas são formalmente definidas como (WEISS; KHOSHGOFTAAR; WANG, 2016):

$$T_S = \{\mathcal{Y}_S, P(Y_S|X_S)\}, \quad T_T = \{\mathcal{Y}_T, P(Y_T|X_T)\},$$

onde $\mathcal{Y}$ é o espaço de rótulos e $P(Y|X)$ a distribuição condicional entre rótulos e características.

- Quando $\mathcal{Y}_S \neq \mathcal{Y}_T$, os domínios possuem espaços de rótulos distintos.
- Quando $P(Y_S|X_S) \neq P(Y_T|X_T)$, há diferença na distribuição condicional, mesmo que os espaços de características e rótulos sejam os mesmos.

2.7.5 Transferência Positiva e Negativa

Considere dois modelos (WEISS; KHOSHGOFTAAR; WANG, 2016):

- $f_{T1}(\cdot)$: modelo preditivo treinado apenas com D_T;
- $f_{T2}(\cdot)$: modelo preditivo treinado com um processo de transferência de aprendizado que combina D_T e D_S.

Define-se:

- **Transferência negativa:** ocorre quando o desempenho de $f_{T1}(\cdot)$ é maior que o desempenho de $f_{T2}(\cdot)$, ou seja, a transferência de conhecimento prejudica o modelo no domínio-alvo.
- **Transferência positiva:** ocorre quando o desempenho de $f_{T2}(\cdot)$ é maior que o desempenho de $f_{T1}(\cdot)$, indicando que o conhecimento transferido melhora o desempenho no domínio-alvo.

2.7.6 Ajuste Fino

O Ajuste Fino é uma das técnicas mais empregadas em aprendizado por transferência, permitindo ajustar modelos pré-treinados para tarefas específicas (YOSINSKI et al., 2014). Ele consiste em re-treinar total ou parcialmente os parâmetros do modelo, a fim de adaptá-lo ao novo domínio.

2.7.6.1 Estratégias de Ajuste Fino

Existem três estratégias principais de ajuste fino:

1. **Extração de Características (Feature Extraction):** apenas a camada de saída é treinada, mantendo as demais camadas congeladas. É a abordagem mais eficiente computacionalmente, mas tende a apresentar desempenho limitado quando os domínios são muito distintos (DAVILA; COLAN; HASEGAWA, 2024).

2. **Ajuste Fino Parcial (Partial Fine-Tuning):** algumas camadas superiores são descongeladas e ajustadas, enquanto as iniciais permanecem fixas. As primeiras camadas geralmente capturam características genéricas (como bordas e texturas), enquanto as superiores aprendem padrões específicos da tarefa (SHEN et al., 2021; DAVILA; COLAN; HASEGAWA, 2024; YOSINSKI et al., 2014).

3. **Ajuste Fino Completo (Full Fine-Tuning):** todas as camadas da rede são ajustadas durante o treinamento no domínio-alvo. Essa é a estratégia mais custosa computacionalmente, mas frequentemente proporciona o melhor desempenho quando aplicada de forma controlada (SHEN et al., 2021; DAVILA; COLAN; HASEGAWA, 2024; YOSINSKI et al., 2014).

2.8 Trabalhos Relacionados

O reconhecimento automático de línguas de sinais tem ganhado destaque na comunidade científica como uma estratégia promissora para promover a acessibilidade e a inclusão de pessoas surdas em ambientes digitais. Essa tarefa envolve a integração de diversas áreas, como visão computacional, processamento de linguagem natural, linguística e técnicas de reconhecimento de padrões (WADHAWAN; KUMAR, 2021).

Os primeiros sistemas de interpretação automática de sinais surgiram a partir dos avanços em interfaces homem-máquina baseadas em gestos. Um marco inicial ocorreu em 1987, quando Zimmerman et al. (1987) propôs uma luva sensorial equipada com sensores de flexão, magnéticos e ultrassônicos, capaz de captar a curvatura dos dedos, além da posição e da orientação da mão. Embora inovadora, essa tecnologia dependia de dispositivos intrusivos e pouco acessíveis, o que limitava sua aplicação a ambientes laboratoriais.

Nos anos seguintes, os esforços continuaram, ainda com vocabulário restrito e forte dependência de sensores vestíveis. Tamura e Kawasaki (1988) desenvolveram um sistema de processamento de imagens para reconhecimento de sinais da Língua de Sinais Japonesa, utilizando a correspondência de queremas, as menores unidades linguísticas das línguas de sinais. De forma semelhante, Charayaphan e Marble (1992) propuseram um sistema para reconhecimento de 31 sinais isolados da Língua de Sinais Americana, baseado em características

visuais extraídas de imagens. Apesar desses avanços, as abordagens iniciais enfrentavam limitações importantes, como a dependência de hardware intrusivo e falhas metodológicas que comprometiam a generalização dos resultados.

Nesse período, o conceito de decompor sinais em subunidades linguísticas começou a ganhar destaque. Kim e Waldron (1993 *apud* (COOPER et al., 2012)) estavam entre os primeiros a reconhecerem a importância de segmentar os sinais em componentes menores, como configuração da mão, orientação, posição e movimento, com o objetivo de tornar o reconhecimento mais eficiente e escalável. Estudos subsequentes, como o de Waldron e Kim (1995), desenvolveram módulos fonêmicos para representar essas subunidades e chegaram a propor a exclusão do movimento como parâmetro isolado em vocabulários reduzidos. Embora inovadora para a época, essa simplificação limitava a aplicabilidade em contextos mais complexos, nos quais o movimento é essencial para diferenciar os sinais.

Ao longo dos anos 2000, surgiram abordagens orientadas por dados para a segmentação automática dessas subunidades, utilizando sensores como acelerômetros (YIN et al., 2009) e explorando padrões de movimento nas trajetórias das mãos (KONG; RANGANATH, 2008). A partir de 2010, com o lançamento do Kinect, houve um avanço significativo na acessibilidade dos sistemas, eliminando a necessidade de dispositivos vestíveis. Trabalhos como o de Zafrulla et al. (2011) demonstraram a viabilidade do rastreamento em tempo real do esqueleto humano, permitindo o reconhecimento de sinais em vocabulários pequenos. No entanto, desafios como a generalização, a independência em relação aos usuários e a escalabilidade para léxicos maiores persistiam. Esses sistemas ainda dependiam de condições controladas e apresentavam desempenho limitado diante da variabilidade linguística e das diferenças individuais, evidenciando a necessidade de técnicas mais robustas e adaptativas.

Com o amadurecimento das tecnologias de visão computacional e o advento do aprendizado profundo, os sistemas de reconhecimento passaram a incorporar arquiteturas baseadas em redes neurais profundas, capazes de extrair automaticamente características visuais e temporais complexas dos vídeos. Esse avanço representou uma ruptura em relação aos métodos anteriores, que dependiam de características projetadas manualmente e de dispositivos físicos para coleta de dados.

As redes convolucionais (CNNs) passaram a ser amplamente utilizadas para capturar padrões espaciais em imagens ou quadros de vídeo, enquanto as redes recorrentes, como as LSTM (Long Short-Term Memory), foram empregadas para modelar dependências temporais dos sinais. Essa combinação tornou o reconhecimento mais robusto e eficiente, especialmente para vocabulários médios e grandes. Por exemplo, Huang e Chouvatut (2024) exploraram essa abordagem híbrida (CNN-LSTM) para o reconhecimento de sinais isolados em vídeos da LSA, estabelecendo um novo patamar de desempenho.

Mais recentemente, arquiteturas baseadas em *Transformers* têm ganhado destaque. Modelos como o *Sign Language Transformer* (SLT) (CAMGOZ et al., 2020) e o trabalho

de Yin e Read (2020) com STMC-Transformer demonstraram que mecanismos de atenção são altamente eficazes para capturar relações espaciais e temporais entre os elementos dos sinais. Esses modelos também se mostraram promissores em tarefas mais complexas, como a tradução automática de vídeos de línguas de sinais para texto escrito, marcando a transição do reconhecimento de sinais isolados para a compreensão de sentenças contínuas.

Paralelamente, ferramentas como o MediaPipe e o OpenPose facilitaram a extração em tempo real de esqueletos corporais, impulsionando abordagens baseadas em *pose estimation*. Essas técnicas passaram a ser utilizadas para reduzir a complexidade computacional dos modelos e aumentar sua capacidade de generalização para diferentes usuários e contextos.

Nesse cenário, a transferência de aprendizado entre diferentes línguas de sinais surgiu como uma estratégia promissora para superar a escassez de dados rotulados em línguas menos documentadas. A ideia central é que modelos treinados em línguas amplamente estudadas possam ser adaptados para outras línguas, como a LIBRAS com um volume reduzido de dados locais para ajuste (*fine-tuning*).

Bird, Ekárt e Faria (2020) aplicaram uma abordagem multimodal com imagens RGB e dados do *Leap Motion*, obtendo bons resultados ao transferir pesos de redes treinadas em Língua de Sinais Britânica (BSL) para ASL. Já Farhadi, Forsyth e White (2007) mostraram que modelos treinados com dados sintéticos (avatares) podem ser aplicados a vídeos reais por meio de representações baseadas em similaridades fonológicas e estruturais. Esses trabalhos reforçam a eficácia do *transfer learning* em reconhecimento de sinais, desde que representações adequadas sejam utilizadas.

No contexto brasileiro, alguns estudos recentes têm avançado no reconhecimento e na tradução automática da LIBRAS, especialmente na construção de bases de dados e na adaptação de modelos de aprendizado profundo. A dissertação de Rodrigues (2021) apresenta a criação do V-LIBRASIL, um repositório robusto de vídeos em LIBRAS alinhado lexicalmente à base ASLLVD. Foram coletados 4.089 vídeos com 1.364 termos, gravados por três sinalizadores em estúdio com fundo cromakey. O trabalho documenta detalhadamente o protocolo de gravação, o pré-processamento, a rotulagem e a disponibilização pública da base, estabelecendo um referencial importante para experimentos em transferência de aprendizado na LIBRAS.

O estudo de Sarmento e Ponti (2023) adota uma abordagem *cross-dataset*, integrando diferentes fontes de dados em LIBRAS para simular cenários mais realistas. Os autores comparam dois métodos de extração de características, CNNs pré-treinadas e *landmarks* corporais, combinados com LSTM para classificação. O trabalho destaca a importância da curadoria e da diversidade dos dados no treinamento, oferecendo diretrizes relevantes para estudos de generalização e transferência entre diferentes conjuntos de sinais.

O estudo de Estevam e Menotti (2023) apresenta uma abordagem para tradução

automática de Libras para português que dispensa a necessidade de anotação manual de sinais individuais. Para viabilizar a pesquisa, foi desenvolvido um dataset original com mais de mil vídeos extensos, obtidos a partir da tradução bíblica em Libras das Testemunhas de Jeová. Empregando *Transformers* em uma arquitetura sequence-to-sequence, o trabalho contribui de forma relevante para o avanço das investigações sobre tradução de línguas de sinais no Brasil.

Prestes (2024) utiliza a biblioteca MediaPipe Holistic para extrair pontos de referência corporais a partir de vídeos da base V-LIBRASIL, combinando essas representações com modelos LSTM para a classificação de sinais isolados. Essa abordagem baseada em *landmarks* alcançou acurácia média de 82,77%, mostrando-se promissora diante da complexidade dos sinais da LIBRAS.

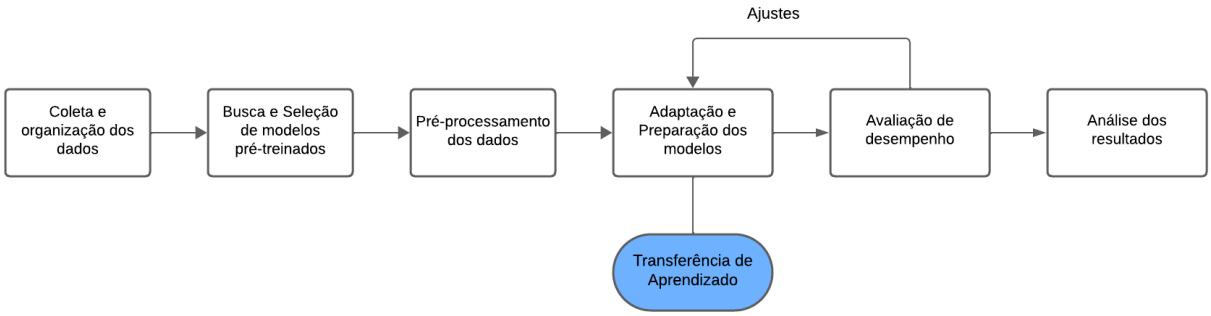
Mais recentemente, Delucis et al. (2025) desenvolveram o conjunto de dados MALTA-LIBRAS, com maior variabilidade de sinalizadores, ambientes e condições de gravação. O estudo investigou três arquiteturas baseadas em *Transformers* (VideoMAE, TimeSformer e ViViT) em diferentes configurações: pré-treinamento em reconhecimento de ações, técnicas de aumento de dados e transferência de conhecimento entre ASL, Língua de Sinais Russa (RSL) e LIBRAS. Os resultados mostraram que modelos pré-treinados atingiram 29% de acurácia, enquanto o aumento de dados elevou esse valor para 33,6%. A transferência direta entre línguas de sinais apresentou ganhos modestos de apenas 2,7%, reforçando a necessidade de estratégias de adaptação específicas por domínio e evidenciando que a diversidade dos dados é tão crucial quanto o volume.

Esses trabalhos evidenciam o avanço da pesquisa sobre o reconhecimento e a tradução automática da LIBRAS, promovendo o reuso, a integração e a adaptação de recursos computacionais para o contexto brasileiro. Ainda assim, estudos focados especificamente na adaptação sistemática de modelos estrangeiros para a LIBRAS permanecem limitados. Diante desse cenário, a presente pesquisa busca preencher essa lacuna, avaliando a transferibilidade de modelos pré-treinados por meio de experimentos com sinais organizados em classes e empregando estratégias de *fine-tuning*. O objetivo é identificar quais abordagens oferecem melhor desempenho em cenários com poucos dados rotulados, contribuindo para o desenvolvimento de soluções tecnológicas mais acessíveis, escaláveis e adequadas às especificidades da LIBRAS, fortalecendo a inclusão digital da comunidade surda brasileira.

3 Metodologia

Este capítulo apresenta os procedimentos metodológicos adotados nesta monografia, conforme ilustrado no pipeline da Figura 7. São descritas as etapas de coleta e organização dos dados, o pré-processamento dos vídeos, a configuração e adaptação dos modelos, bem como os processos de treinamento e validação. Também são apresentados os critérios de análise e as métricas utilizadas para examinar o desempenho dos modelos, permitindo uma avaliação sistemática e precisa dos resultados obtidos.

Figura 7 – Pipeline da metodologia



Fonte: Elaborada pela Autora

3.1 Base de Dados

Para a realização dos experimentos, foram utilizados quatro conjuntos de dados públicos de vídeos em LIBRAS. De forma isolada, cada repositório apresentava um número reduzido de amostras por sinal, o que inviabilizava sua aplicação como conjunto independente. Dessa forma, optou-se pela combinação dos quatro, resultando em um *dataset* unificado.

a) V-LIBRASIL

O V-LIBRASIL, desenvolvido no contexto de uma dissertação de mestrado (RODRIGUES, 2021), é uma base de vídeos de sinais em LIBRAS com o objetivo de reduzir a escassez de repositórios amplos e bem estruturados para a língua. A base é composta por 1.364 termos, expressões e sintagmas, articulados por três sinalizadores (identificados como articuladores 1, 2 e 3), totalizando 4.089 instâncias. Os vídeos apresentam boa qualidade visual, com fundo neutro e enquadramento padronizado, além de conterem, na posição superior esquerda, um rótulo que indica o significado do sinal.

b) Libras SignBank

O Libras SignBank, desenvolvido na Universidade Federal de Santa Catarina (PIZZIO et al., 2021), disponibiliza vídeos de sinais em LIBRAS, contando atualmente com cerca de 3.093 glossas. Cada sinal é apresentado por apenas um intérprete. Os vídeos estão em formato .mp4, possuem fundo azul neutro e são organizados alfabeticamente, facilitando a consulta.

c) MINDS-Libras Dataset

O MINDS-Libras Dataset, construído como parte de uma tese de doutorado (REZENDE, 2021) para reconhecimento de sinais em LIBRAS, é composto por 20 sinais, cada um executado 5 vezes por 12 diferentes sinalizadores, totalizando 1.200 vídeos. As gravações têm resolução de 1920×1080 pixels, formato .mp4 e fundo preparado com técnica de *Chroma Key*, garantindo contraste adequado entre sinalizador e ambiente.

d) Dicionário Online Libras-Português

O Dicionário Online Libras-Português, desenvolvido pelo projeto Inovar +, da Universidade Federal de Viçosa (GEDIEL; OLIVEIRA; RODRIGUES, 2021; INOVAR+, 2017), é uma ferramenta digital criada para auxiliar no ensino e aprendizagem da Língua Brasileira de Sinais. Cada vídeo apresenta um articulador por palavra e possui fundo neutro, garantindo melhor visibilidade do sinal.

Integração dos Conjuntos de Dados

Para a realização dos experimentos, os quatro conjuntos de dados foram integrados a partir da seleção de classes comuns presentes em todas ou na maioria das bases, priorizando palavras que possibilitassem reunir o maior número de amostras e ampliar a representatividade do *dataset*. Como resultado, foi construído um conjunto unificado contendo 45 classes. Para cada uma delas, foram incorporadas, em média, 4 a 6 instâncias de vídeo provenientes das diferentes bases, buscando equilibrar a distribuição entre elas e minimizar possíveis vieses.

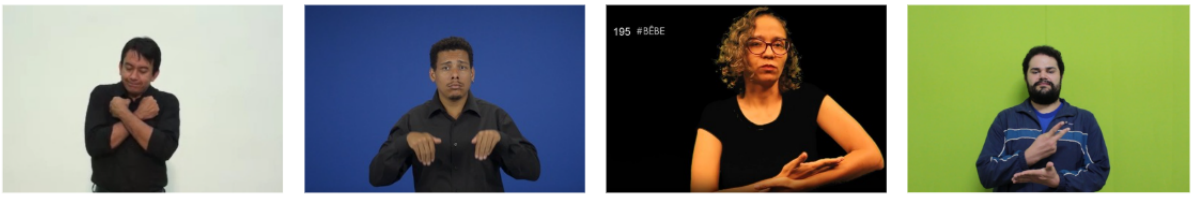
Embora correspondam às mesmas palavras, foram observadas variações na execução dos gestos entre diferentes intérpretes e bases de dados, refletindo a diversidade natural da LIBRAS. Essas diferenças tornam o conjunto mais rico e representativo, mas também aumentam a complexidade da tarefa de reconhecimento. Dessa forma, o *dataset* unificado apresenta maior heterogeneidade e robustez ao reunir sinais de diferentes origens, produzidos por múltiplos comunicadores e com distintas formas de execução.

A Tabela 1 apresenta a distribuição detalhada das instâncias, enquanto a Figura 8 mostra exemplos de vídeos provenientes de cada *dataset* original.

Tabela 1 – Distribuição das classes do *dataset*: identificador, palavra correspondente, número total de vídeos e respectivas quantidades de treino, validação e teste.

ID	**Classe (Palavra)**	**Total de Vídeos**	**Treino**	**Validação**	**Teste**
0	Abraçar	5	3	1	1
1	Acontecer	6	4	1	1
2	Amarelo	5	3	1	1
3	América	5	3	1	1
4	Amigo	5	3	1	1
5	Ano	4	2	1	1
6	Aproveitar	4	2	1	1
7	Bala	5	3	1	1
8	Banco	4	2	1	1
9	Banheiro	5	3	1	1
10	Barulho	5	3	1	1
11	Bebê	4	2	1	1
12	Bicicleta	5	3	1	1
13	Canguru	5	3	1	1
14	Carro	5	3	1	1
15	Casa	5	3	1	1
16	Cavalo	5	3	1	1
17	Chocolate	5	3	1	1
18	Cinco	5	3	1	1
19	Comer	5	3	1	1
20	Conhecer	5	3	1	1
21	Cor	4	2	1	1
22	Corpo	4	2	1	1
23	Desculpa	4	2	1	1
24	Dinheiro	4	2	1	1
25	Espelho	5	3	1	1
26	Esquina	5	3	1	1
27	Família	5	3	1	1
28	Flor	5	3	1	1
29	Irmã	5	3	1	1
30	Maçã	5	3	1	1
31	Macaco	4	2	1	1
32	Medo	5	3	1	1
33	Não	4	2	1	1
34	Pessoa	5	3	1	1
35	Piano	5	3	1	1
36	Porco	4	2	1	1
37	Quarta	5	3	1	1
38	Respirar	4	2	1	1
39	Ruim	5	3	1	1
40	Sapo	5	3	1	1
41	Sim	4	2	1	1
42	Sol	4	2	1	1
43	Vacina	5	3	1	1
44	Vontade	5	3	1	1

Figura 8 – Quadro ilustrativo representando um exemplo de cada conjunto de dados utilizado: Dicionário Online Libras-Português, Libras SignBank, VLIBRASIL e MINDS-Libras.



Fonte: Adaptado de (GEDIEL; OLIVEIRA; RODRIGUES, 2021; INOVAR+, 2017), (PIZZIO et al., 2021), (RODRIGUES, 2021), (REZENDE, 2021)

3.2 Métricas de Avaliação

Para avaliar o desempenho dos modelos de forma abrangente, foram utilizadas métricas complementares à acurácia tradicional, permitindo uma análise mais meticulosa e consistente dos resultados obtidos.

3.2.1 Acurácia Top-k

A acurácia Top-k é uma métrica que considera uma previsão correta sempre que o rótulo verdadeiro está entre as k opções mais prováveis indicadas pelo modelo, diferentemente da acurácia convencional, que avalia apenas a previsão principal.

No contexto do reconhecimento automático de sinais, a acurácia Top-k é especialmente relevante devido ao grande número de gestos e à semelhança visual entre eles, que diferem apenas por detalhes sutis. Nesses casos, o modelo pode confundir alternativas próximas, ainda que a classe correta esteja entre as previsões mais prováveis. A Top-k fornece uma avaliação mais realista da capacidade do modelo de restringir suas previsões a um conjunto reduzido de sinais plausíveis, evidenciando sua utilidade em aplicações práticas (SARMENTO, 2023).

3.2.2 Matriz de Confusão

A matriz de confusão é uma ferramenta amplamente utilizada na avaliação de modelos de classificação, pois permite analisar de forma estruturada os acertos e erros do classificador. Ela é representada em formato matricial, em que as linhas correspondem aos rótulos reais e as colunas, aos rótulos previstos pelo modelo. Cada elemento M_{ij} da matriz indica quantas instâncias pertencentes ao rótulo real i foram atribuídas ao rótulo j pelo classificador. Dessa forma, a diagonal principal concentra os acertos, enquanto os elementos fora da diagonal representam os erros de classificação (CARVALHO et al., 2011).

A matriz de confusão permite evidenciar quais classes são mais suscetíveis a erros de classificação, oferecendo uma visão detalhada sobre o comportamento do modelo. Em tarefas envolvendo múltiplas categorias ou sinais visualmente similares, como no reconhecimento de LIBRAS, essa análise é fundamental tanto para a interpretação do desempenho quanto para a definição de estratégias de otimização do modelo.

3.3 Modelos

Para avaliar o desempenho no reconhecimento de sinais em LIBRAS, foram implementados três modelos distintos baseados em duas abordagens de extração de características. O primeiro modelo explora a aparência visual holística dos vídeos (*frames* RGB), aprendendo a reconhecer os sinais a partir de todas as informações visuais disponíveis, como a configuração das mãos, as cores e o fundo. Os outros dois modelos concentram-se na pose humana 2D, desconsiderando a aparência visual e utilizando apenas a geometria corporal: a entrada consiste em sequências temporais das coordenadas (X, Y) dos pontos chave (*keypoints*) do corpo, permitindo que a rede aprenda exclusivamente a cinemática e a trajetória dos movimentos característicos da LIBRAS.

O processo de *fine-tuning* foi estruturado de forma padronizada para todas as arquiteturas, compreendendo os seguintes procedimentos:

- **Modificação da camada de saída:** A camada final de classificação de cada modelo foi substituída e reinicializada para refletir o número de classes do novo *dataset* utilizado.
- **Treinamento da rede:** Todas as camadas da arquitetura foram mantidas treináveis (*fine-tuning* completo), permitindo que os parâmetros pré-treinados se adaptassem às características específicas do dataset de LIBRAS.

Essa estratégia buscou, ao mesmo tempo, preservar o conhecimento prévio dos modelos sobre dinâmicas de movimento em geral e promover sua adaptação às particularidades visuais e temporais da língua de sinais brasileira.

A Tabela 2 apresenta os hiperparâmetros utilizados no treinamento das arquiteturas I3D, Bi-LSTM e SL-GCN. Para todos os modelos, foi adotado o otimizador Adam com o objetivo de garantir convergência estável durante o treinamento. Os parâmetros especificados incluem a taxa de aprendizado, o tamanho do lote (*batch size*), o número máximo de quadros considerados por amostra de vídeo, a semente aleatória para garantir a reprodutibilidade dos experimentos e o número de *workers* empregados no carregamento paralelo dos dados. A definição desses hiperparâmetros foi realizada considerando as limitações computacionais disponíveis e as características específicas de cada arquitetura, buscando um equilíbrio entre

desempenho e eficiência durante o processo de treinamento.

Tabela 2 – Hiperparâmetros utilizados no treinamento dos modelos.

Hiperparâmetro	**I3D**	**LSTM**	**SL-GCN**
Learning Rate	0,001	0,001	0,0005
Batch Size	8	64	32
Otimizador	Adam	Adam	Adam
Max Frames	64	32	32
Seed	0	42	0
Workers	4	2	4

3.3.1 I3D

O primeiro modelo selecionado para o fine-tuning baseou-se na arquitetura I3D, detalhada no Capítulo 2. A implementação adotada seguiu o trabalho de (LI et al., 2020), empregando o fluxo RGB da arquitetura I3D, sem a incorporação de fluxo óptico.

No estudo utilizado como referência, o conjunto de dados WLASL (*Word-Level American Sign Language*) foi originalmente subdividido pelos autores com o objetivo de permitir a avaliação de modelos de reconhecimento de sinais em diferentes escalas de vocabulário. O *dataset* completo foi dividido em quatro subconjuntos organizados de forma hierárquica, de modo que cada subconjunto maior continha todos os sinais do subconjunto menor anterior acrescidos de sinais adicionais:

- **WLASL100:** 100 sinais
- **WLASL300:** 300 sinais
- **WLASL1000:** 1.000 sinais
- **WLASL2000:** 2.000 sinais

Para a adaptação ao reconhecimento de sinais em LIBRAS, o modelo I3D foi inicializado com pesos pré-treinados nos quatro subconjuntos do WLASL. Essa estratégia de *transfer learning* permitiu que a rede preservasse o conhecimento prévio sobre padrões de movimento e dinâmicas temporais adquiridos em ASL.

Para adequar os dados de vídeo às especificações da arquitetura neural utilizada, foi desenvolvido um pipeline de pré-processamento descrito a seguir.

Leitura e Captura dos Frames. O processo inicia com a leitura dos vídeos no formato .mp4 utilizando a biblioteca OpenCV. Para cada vídeo, captura-se uma sequência fixa de 64 *frames*. Essa padronização garante que todos os vídeos possuam a mesma dimensão temporal, facilitando o processamento em lote pelo modelo.

Redimensionamento e Normalização. Após a captura, cada *frame* individual passa por normalização e redimensionamento:

- **Redimensionamento Fixo:** Todos os *frames* são redimensionados para 256×256 pixels, padronizando as dimensões espaciais das entradas do modelo. Essa etapa não preserva a proporção original, mas garante uniformidade entre as amostras.
- **Normalização de Pixels:** Os valores de pixel são normalizados para o intervalo $[-1, 1]$ com o objetivo de melhorar a estabilidade numérica e acelerar o processo de treinamento.

Padronização da Quantidade de Frames. Como medida de consistência, verifica-se que todos os vídeos resultantes possuem exatamente 64 *frames*. Caso contrário, vídeos mais curtos são completados com a repetição do último *frame*, enquanto vídeos mais longos são truncados até atingir esse limite.

Transformações e Aumento de Dados. Com os *frames* no formato padronizado, aplicam-se transformações diferenciadas para as fases de treinamento e validação:

- **Recorte (Crop):** Durante o treinamento, aplica-se um recorte aleatório (`RandomCrop`) de 224×224 pixels nos *frames* de 256×256, servindo como técnica de aumento de dados. Para a validação e teste, utiliza-se um recorte central (`CenterCrop`) de 224×224 pixels, garantindo consistência na avaliação.
- **Aumento de Dados:** Durante o treinamento, aplica-se espelhamento horizontal aleatório (`RandomHorizontalFlip`) para tornar o modelo mais robusto a variações de orientação dos gestos.

3.3.2 Bi-LSTM e SL-GCN

Os dois modelos restantes têm como referência o trabalho de (SELVARAJ et al., 2021). Nesse estudo, foram avaliados quatro modelos distintos em sete bases de dados (INCLUDE - Indiana (SRIDHAR et al., 2020), AUTSL - Turca (SINCAN; KELES, 2020), GSL - Grega (ADALOGLOU et al., 2021), DEVISIGN - Chinesa (CHAI; WANG; CHEN, 2014), CSL - Chinesa (HUANG et al., 2018), LSA64 - Argentina (RONCHETTI et al., 2016) e WLASL2000 - Americana (LI et al., 2020)), abrangendo um total de seis línguas de sinais, sendo o chinês considerado em mais de um *dataset*. Entre os quatro modelos analisados, os que serviram de base para a aplicação de *transfer learning* nesta monografia foram o Bi-LSTM e o SL-GCN, ambos descritos em detalhe no Capítulo 2.

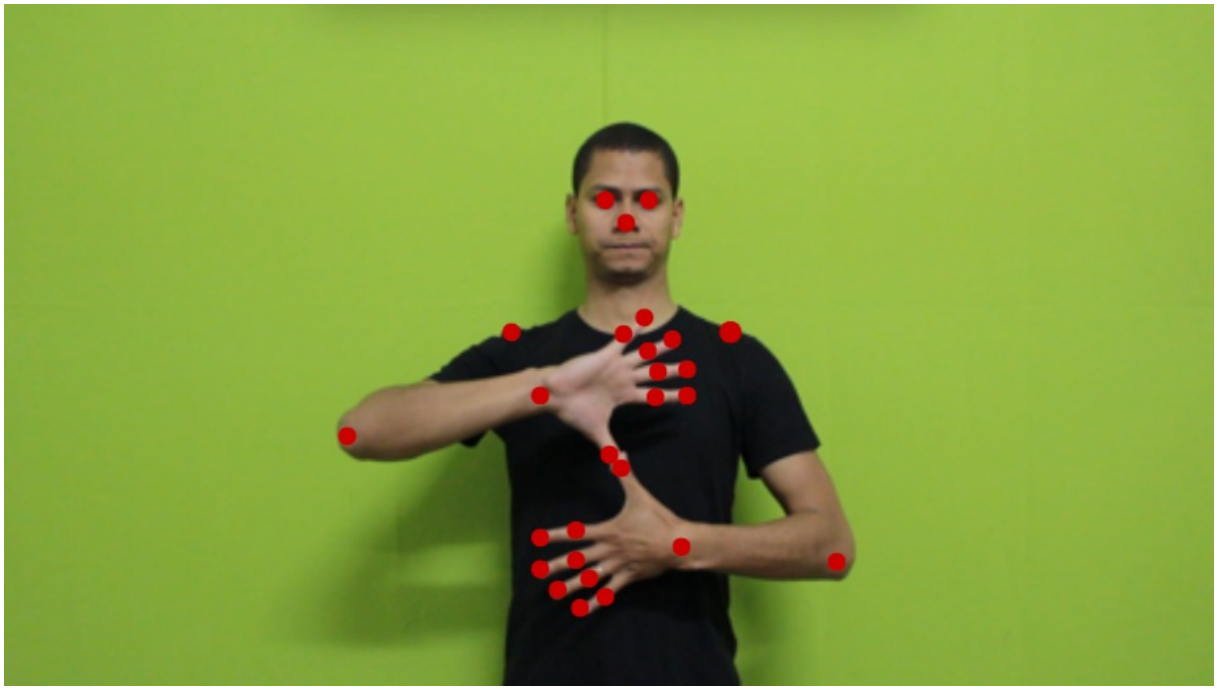
Diferentemente do modelo I3D, que processa diretamente os *frames* dos vídeos em formato de imagem, os modelos baseados em Bi-LSTM e SL-GCN utilizam como entrada sequências numéricas de pontos-chave (*keypoints*) extraídos dos vídeos. O objetivo desse pipeline de pré-processamento é transformar os vídeos brutos em representações temporais

compactas e padronizadas, contendo apenas as informações mais relevantes para a modelagem dos gestos.

Extração de Pontos-Chave. A primeira etapa consiste na conversão de cada vídeo em uma sequência de pontos-chave, realizada com a biblioteca `MediaPipe Holistic`. Essa ferramenta permite a detecção simultânea da pose do corpo, das mãos e da face. Para este trabalho, foram selecionadas as seguintes regiões de interesse: ombros e cotovelos (corpo), punhos, articulações e pontas dos dedos (mãos), além do nariz e dos olhos (face). A Figura 9 ilustra os *keypoints* extraídos para utilização nas etapas posteriores do processo de reconhecimento.

Em cada *frame*, as coordenadas bidimensionais (x, y) desses pontos são concatenadas em um vetor único, representando a pose correspondente ao instante analisado. Ao longo do tempo, os vetores formam uma sequência que descreve a evolução do gesto. Em casos de falhas de detecção, as posições não identificadas são preenchidas com $(0.0, 0.0)$, garantindo consistência no formato dos dados.

Figura 9 – Exemplo de extração de pontos-chave utilizados no pré-processamento dos vídeos (MINDS-Libras Dataset).



Fonte: Adaptado de (REZENDE, 2021).

Normalização e Aumentação de Dados. Após a extração, os arquivos de *keypoints* passam por duas transformações fundamentais:

- **Normalização Min-Max**: As coordenadas (x, y) de cada vídeo são escaladas individualmente para o intervalo $[0, 1]$, de acordo com os valores mínimos e máximos observados.

Esse procedimento confere robustez ao modelo frente a variações de escala e de posicionamento do indivíduo em relação à câmera.

- **Aumento de Dados**: Foi adicionada uma pequena perturbação de ruído gaussiano às coordenadas normalizadas dos pontos-chave. Essa técnica de regularização tem como objetivo aumentar a robustez do modelo e mitigar o *overfitting*, favorecendo uma melhor capacidade de generalização diante de amostras não vistas durante o treinamento.

Padronização do Comprimento das Sequências. Como os vídeos possuem durações distintas, as sequências resultantes apresentam comprimentos variáveis. Para padronizar as entradas:

- Sequências com mais de 32 quadros são truncadas, preservando apenas os primeiros;
- Sequências menores que 32 quadros são completadas com *padding*, preenchendo os quadros ausentes com valores nulos.

Esse processo garante que todas as sequências tenham exatamente o mesmo comprimento, requisito essencial para o processamento em lote pelos modelos Bi-LSTM e SL-GCN.

4 Resultados

Este capítulo apresenta e discute os resultados experimentais obtidos na avaliação de três arquiteturas de redes neurais aplicadas ao reconhecimento automático de sinais em LIBRAS: I3D, Bi-LSTM e SL-GCN. A análise foi conduzida considerando o treinamento completo do zero (*from scratch*) e o aprendizado por transferência a partir de diversos *datasets* de línguas de sinais internacionais. Os resultados são reportados em termos de acurácia Top-k (k = 1, 5, 10), permitindo avaliar não apenas a capacidade de predição direta dos modelos, mas também sua confiança nas hipóteses alternativas. Além da análise quantitativa de desempenho, este capítulo inclui uma comparação entre as abordagens baseadas em RGB e em poses, uma investigação sobre a transferibilidade de conhecimento entre diferentes línguas de sinais e uma análise detalhada das matrizes de confusão, visando identificar padrões de erro e similaridades visuais entre sinais que desafiam os sistemas de reconhecimento automático.

4.1 I3D

A Tabela 3 apresenta os resultados de acurácia Top-k para a arquitetura I3D considerando diferentes estratégias de inicialização de pesos, incluindo treinamento do zero (*from scratch*) e aprendizado por transferência a partir de variantes do *dataset* WLASL com diferentes volumes de dados (100, 300, 1000 e 2000 classes). O modelo treinado sem pré-treinamento obteve desempenho extremamente baixo, com Top-1 de apenas 11,11%, Top-5 de 28,89% e Top-10 de 53,33%, evidenciando a dificuldade da arquitetura I3D em aprender representações discriminativas a partir de um *dataset* limitado sem conhecimento prévio.

Ao empregar pesos pré-treinados do WLASL, observa-se um aumento expressivo na acurácia em todas as configurações, confirmando o potencial do aprendizado por transferência. O modelo com WLASL-1000 apresentou o melhor desempenho geral, atingindo 86,67% de acurácia Top-1 e 93,33% de Top-5, além de compartilhar o melhor resultado de Top-10 (97,78%) com WLASL-100 e WLASL-2000, destacando-se como a configuração mais eficaz.

Um aspecto interessante é que WLASL-100, apesar de ter o menor volume de dados pré-treinados, alcançou acurácia Top-1 de 73,33%, Top-5 de 88,89% e Top-10 de 97,78%, superando configurações com volumes intermediários como WLASL-300 (64,44% Top-1, 84,44% Top-5, 91,11% Top-10). Por outro lado, WLASL-2000 apresentou desempenho inferior em Top-1 (66,67%) comparado ao WLASL-1000, embora tenha mantido excelente acurácia Top-5 (93,33%) e Top-10 (97,78%). Isso sugere que o volume e a qualidade do *dataset* pré-treinado influenciam fortemente a transferência de conhecimento, mas que existe um ponto ótimo em que a representação pré-treinada se adapta melhor ao contexto do *dataset* alvo. Além disso, evidencia-se que nem sempre *datasets* maiores garantem melhor desempenho, podendo ocorrer

problemas de sobreajuste ou incompatibilidade de domínio.

Em síntese, os resultados demonstram que a I3D depende fortemente de aprendizado por transferência para obter desempenho competitivo no reconhecimento de sinais em LIBRAS. A escolha do tamanho e da diversidade do *dataset* pré-treinado é determinante, sendo crucial encontrar um equilíbrio entre volume e compatibilidade das representações aprendidas para maximizar a transferência de conhecimento.

Tabela 3 – Resultados de acurácia Top-k para a arquitetura I3D com diferentes pesos pré-treinados e *fine-tuning* completo

Pesos Pré-treinados	**Língua de Sinais**	**Top-1**	**Top-5**	**Top-10**
from scratch	-	0.1111	0.2889	0.5333
WLASL-100	Língua Americana de Sinais	0.7333	0.8889	**0.9778**
WLASL-300	Língua Americana de Sinais	0.6444	0.8444	0.9111
WLASL-1000	Língua Americana de Sinais	**0.8667**	**0.9333**	**0.9778**
WLASL-2000	Língua Americana de Sinais	0.6667	**0.9333**	**0.9778**

4.2 Bi-LSTM

A Tabela 4 apresenta os resultados de acurácia Top-k para a arquitetura Bi-LSTM considerando diferentes estratégias de inicialização de pesos: treinamento completo do zero (*from scratch*) e aprendizado por transferência a partir de seis *datasets* distintos de línguas de sinais. O modelo treinado sem pré-treinamento alcançou acurácia Top-1 de 31,11%, estabelecendo a linha de base para comparação. Os resultados evidenciam que a maioria dos modelos com pesos pré-treinados supera essa configuração inicial, comprovando o potencial do aprendizado por transferência para o reconhecimento de LIBRAS. Entretanto, verifica-se variação considerável no desempenho dependendo da fonte dos pesos utilizados.

Entre os modelos pré-treinados, destacam-se diferenças substanciais conforme a origem dos pesos. O WLASL-2000 (Língua Americana de Sinais) obteve a maior acurácia Top-1, com 55,56%, seguido pelo LSA64 (Língua Argentina de Sinais), com 46,67%, e pelo DEVISIGN (Língua Chinesa de Sinais), com 37,78%. Por outro lado, o AUTSL (Língua Turca de Sinais) e o GSL (Língua Grega de Sinais) apresentaram acurácia de apenas 33,33%, enquanto o INCLUDE (Língua Indiana de Sinais) atingiu 35,56% de acurácia Top-1. Essa variação indica que a eficácia da transferência de conhecimento não se explica apenas pela proximidade geográfica entre as línguas de sinais, mas possivelmente por similaridades estruturais e gestuais, além da qualidade e abrangência dos *datasets* de origem. O excelente desempenho tanto do WLASL-2000 quanto do LSA64 pode ser atribuído a múltiplos fatores complementares: o volume e diversidade de seus respectivos *datasets*, potenciais características gestuais e estruturais compartilhadas com a LIBRAS, e a qualidade das representações capturadas durante o

pré-treinamento. A diferença nos padrões de desempenho — com WLASL apresentando melhor Top-1 e LSA64 superior em Top-5 e Top-10 — sugere que cada *dataset* captura aspectos distintos da variabilidade gestual, o que merece investigação futura para compreender quais propriedades linguísticas e computacionais favorecem diferentes tipos de transferência.

As métricas Top-5 e Top-10 apresentam valores substancialmente superiores ao Top-1 em todos os modelos avaliados. Para o melhor modelo (WLASL-2000), as acurácias Top-5 e Top-10 atingem 68,89% e 82,22%, respectivamente, contra 55,56% do Top-1. Destaca-se também o desempenho do LSA64, que alcançou as maiores acurácias Top-5 (75,56%) e Top-10 (84,44%), superando inclusive o WLASL-2000 nessas métricas. Esse comportamento revela que, mesmo quando a predição de maior confiança não corresponde ao sinal correto, o modelo frequentemente inclui a classe correta entre suas principais alternativas.

Em síntese, os resultados demonstram que o aprendizado por transferência a partir de pesos pré-treinados em outras línguas de sinais constitui uma estratégia eficaz para o reconhecimento de LIBRAS com arquitetura Bi-LSTM, superando consistentemente o treinamento do zero. A variação de desempenho entre diferentes fontes de pré-treinamento reforça que o sucesso da transferência depende de múltiplos fatores, incluindo similaridades estruturais entre as línguas de sinais, características gestuais compartilhadas e a riqueza dos *datasets* utilizados, não se limitando apenas à proximidade geográfica ou familiar entre as línguas.

Tabela 4 – Resultados de acurácia Top-k para a arquitetura Bi-LSTM com diferentes pesos pré-treinados e *fine-tuning* completo

Pesos Pré-treinados	**Língua de Sinais**	**Top-1**	**Top-5**	**Top-10**
from scratch	-	0.3111	0.6889	0.7778
GSL	Língua Grega de Sinais	0.3333	0.5556	0.6889
AUTSL	Língua Turca de Sinais	0.3333	0.6889	0.7111
INCLUDE	Língua Indiana de Sinais	0.3556	0.5333	0.6444
DEVISIGN	Língua Chinesa de Sinais	0.3778	0.6667	0.8222
LSA64	Língua Argentina de Sinais	0.4667	**0.7556**	**0.8444**
WLASL-2000	Língua Americana de Sinais	**0.5556**	0.6889	0.8222

4.3 SL-GCN

A arquitetura SL-GCN apresenta resultados que evidenciam a complexidade do aprendizado por transferência em modelos baseados em grafos espaço-temporais, conforme demonstrado na Tabela 5. A variabilidade observada entre os diferentes conjuntos de pesos pré-treinados sugere que a eficácia da transferência de conhecimento depende de fatores específicos relacionados à compatibilidade das representações espaciais capturadas, à diversidade dos *datasets* de origem e às características estruturais dos sinais, mais do que simplesmente à proximidade linguística ou geográfica entre as línguas.

Os resultados de acurácia Top-k revelam desempenhos heterogêneos entre os modelos avaliados. O modelo treinado do zero (*from scratch*) estabeleceu a linha de base com acurácia Top-1 de 37,78%. Entre as configurações com pré-treinamento, três *datasets* — LSA (Língua Argentina de Sinais), AUTSL (Língua Turca de Sinais) e GSL (Língua Grega de Sinais) — alcançaram o melhor resultado, todos atingindo 53,33% de acurácia Top-1, representando melhoria expressiva de aproximadamente 41% em relação à linha de base. O DEVISIGN (Língua Chinesa de Sinais) e o WLASL (Língua Americana de Sinais) apresentaram desempenho equivalente, ambos atingindo 51,11% de acurácia Top-1. O INCLUDE (Língua Indiana de Sinais) obteve ganho mais modesto, com 46,67%. Esses resultados indicam que a SL-GCN possui requisitos particulares quanto à natureza e estrutura dos dados pré-treinados para que a transferência de conhecimento seja efetiva.

As métricas de classificação expandida revelam padrão consistente em todas as configurações testadas. Os valores de Top-5 e Top-10 superam substancialmente os resultados de Top-1, indicando que a SL-GCN consegue extrair características discriminativas relevantes da estrutura esquelética representada como grafo espaço-temporal. Mesmo quando a predição de maior confiança não corresponde ao sinal correto, o modelo frequentemente posiciona a classe correta entre as hipóteses de maior probabilidade. Notavelmente, o GSL apresentou o melhor desempenho em Top-5 (77,78%), enquanto o DEVISIGN alcançou o maior valor de Top-10 (88,89%). Em resumo, o aprendizado por transferência mostra-se benéfico para a arquitetura SL-GCN no contexto de reconhecimento de LIBRAS, embora os ganhos variem consideravelmente dependendo da fonte de pré-treinamento. A seleção criteriosa dos pesos iniciais emerge como fator determinante para o sucesso da transferência, sendo essencial considerar não apenas o volume ou a qualidade geral do *dataset*, mas especialmente sua compatibilidade com a natureza das representações baseadas em grafos espaço-temporais que caracterizam este modelo. Os resultados sugerem que determinados *datasets*, mesmo de línguas geograficamente distantes, podem fornecer representações mais transferíveis do que outros, reforçando a necessidade de experimentação sistemática ao aplicar aprendizado por transferência em arquiteturas baseadas em GCN.

Tabela 5 – Resultados de acurácia Top-k para a arquitetura SL-GCN com diferentes pesos pré-treinados e *fine-tuning* completo

Pesos Pré-treinados	**Língua de Sinais**	**Top-1**	**Top-5**	**Top-10**
from scratch	-	0.3778	0.6444	0.8000
INCLUDE	Língua Indiana de Sinais	0.4667	0.7556	0.8000
WLASL	Língua Americana de Sinais	0.5111	0.6889	0.8000
DEVISIGN	Língua Chinesa de Sinais	0.5111	0.7556	**0.8889**
LSA64	Língua Argentina de Sinais	**0.5333**	0.7111	0.7778
AUTSL	Língua Turca de Sinais	**0.5333**	0.7333	0.8222
GSL	Língua Grega de Sinais	**0.5333**	**0.7778**	0.8444

4.4 Comparação entre Modelos RGB e Baseados em Poses

A análise comparativa entre os resultados obtidos pelas arquiteturas baseadas em *RGB* (I3D) e aquelas fundamentadas em representações de *poses* (Bi-LSTM e SL-GCN) evidencia comportamentos distintos quanto ao impacto do aprendizado por transferência no desempenho final, embora ambos os tipos de modelo se beneficiem dessa estratégia.

O modelo *I3D* apresentou uma elevação extremamente acentuada na acurácia após o uso de pesos pré-treinados, passando de apenas 11,11% de *Top-1* no treinamento do zero para 86,67% quando inicializado com pesos provenientes do *WLASL-1000*. Esse salto expressivo demonstra o quanto modelos de natureza visual, que operam diretamente sobre informações de textura, cor e movimento, dependem fortemente de conhecimento prévio para extrair representações discriminativas robustas, sobretudo em contextos onde o conjunto de dados de destino é limitado.

Essa dependência pode ser explicada pela alta complexidade do *I3D*, refletida em seu grande número de parâmetros e na profundidade de sua estrutura convolucional tridimensional. Modelos profundos como o *I3D* possuem elevada capacidade de memorização, conseguindo ajustar-se perfeitamente até mesmo a rótulos aleatórios dos dados de treinamento (ZHANG et al., 2017), o que indica que, na ausência de dados suficientes, tais modelos podem sobreajustar aos padrões específicos do conjunto de treino, prejudicando a generalização. Nesse contexto, o aprendizado por transferência atua como um fator crucial ao fornecer ao modelo uma base semântica rica que facilita a adaptação às novas classes de sinais. Os pesos pré-treinados carregam conhecimento sobre características visuais fundamentais, como detecção de bordas, texturas, movimentos e padrões espaço-temporais, aprendidos a partir de grandes volumes de dados, os quais podem ser reutilizados e refinados para o contexto específico da *LIBRAS*, reduzindo significativamente a quantidade de dados necessários para atingir desempenho competitivo.

Os modelos baseados em *poses*, Bi-LSTM e SL-GCN, apresentam, em contraste, incrementos mais modestos de desempenho com o uso de pré-treinamento. No caso do Bi-LSTM, a acurácia *Top-1* evoluiu de 31,11% no treinamento do zero para 55,56% com pesos do *WLASL-2000*, enquanto o *SL-GCN* passou de 26,67% para 53,33% na mesma métrica. Esses aumentos, embora significativos, são consideravelmente mais suaves que os observados no *I3D*. Tal comportamento decorre da natureza mais compacta e estruturada das representações de *poses*, que descrevem a dinâmica corporal por meio de coordenadas articulares, eliminando variações visuais irrelevantes como iluminação, textura e fundo (LI et al., 2025). Por operarem em um espaço de características mais abstrato e de menor dimensionalidade, esses modelos demandam menos dados para aprender relações temporais e espaciais relevantes, tornando-se intrinsecamente menos dependentes de grandes volumes de pré-treinamento. Ademais, a simplicidade relativa de suas arquiteturas, em comparação com o *I3D*, reduz o risco de sobreajuste e favorece uma adaptação mais estável, ainda que com ganhos mais moderados. Essa diferença

é consistente com a literatura, que aponta os efeitos adversos da alta dimensionalidade em modelos complexos e destaca a importância de representações mais compactas para mitigar o sobreajuste e aprimorar a generalização (VERLEYSEN; FRANÇOIS, 2005).

A discrepância entre os ganhos de acurácia observados nos modelos baseados em *RGB* e nos modelos baseados em *poses* reflete, portanto, a diferença fundamental de complexidade e natureza das representações aprendidas. Enquanto modelos visuais profundos, como o I3D, beneficiam-se de forma dramática do aprendizado por transferência devido à sua alta capacidade de parametrização e dependência de padrões visuais complexos, modelos estruturais, como Bi-LSTM e SL-GCN, apresentam comportamento mais robusto e estável, alcançando melhorias progressivas e consistentes. Esses resultados reforçam a importância de alinhar a estratégia de pré-treinamento à complexidade intrínseca do modelo e à natureza dos dados de entrada, uma vez que o impacto do conhecimento prévio tende a ser proporcional à expressividade e à profundidade da arquitetura utilizada.

4.5 Análise Comparativa de Transferência entre Línguas de Sinais

A análise dos resultados apresentados nas Tabelas 4 e 5 revela que não existe uma língua de sinais universalmente superior para a transferência de aprendizado, uma vez que o desempenho varia conforme a arquitetura e a métrica considerada.

Na arquitetura Bi-LSTM, os pesos pré-treinados em WLASL-2000 (Língua Americana de Sinais) alcançaram a maior acurácia *Top-1* (55,56%), enquanto a LSA64 (Língua Argentina de Sinais) se destacou nas métricas *Top-5* (75,56%) e *Top-10* (84,44%).

Para a arquitetura SL-GCN, três conjuntos de pesos pré-treinados — LSA64 (Língua Argentina de Sinais), AUTSL (Língua Turca de Sinais) e GSL (Língua Grega de Sinais) — compartilharam o melhor desempenho em *Top-1* (53,33%). Quanto às métricas expandidas, o GSL apresentou os melhores resultados em *Top-5* (77,78%), enquanto o DEVISIGN alcançou o maior valor de *Top-10* (88,89%).

Apesar dessa variação entre as origens dos pesos, a LSA64 demonstrou desempenho consistente e competitivo em ambas as arquiteturas, figurando entre os melhores resultados em múltiplas métricas. No caso da Bi-LSTM, a LSA64 obteve a segunda melhor *Top-1* (46,67%) e liderou as métricas *Top-5* e *Top-10*, evidenciando boa capacidade de generalização. Já na SL-GCN, manteve desempenho competitivo, com *Top-1* de 53,33% (empatada com AUTSL e GSL), *Top-5* de 71,11% e *Top-10* de 77,78%.

Essa consistência sugere que o conjunto LSA64 captura características visuais e temporais dos sinais que apresentam boa transferibilidade para a LIBRAS. As razões específicas para esse desempenho ainda requerem investigação mais aprofundada, podendo estar relacionadas a similaridades gestuais, estruturais, complexidade do vocabulário ou características

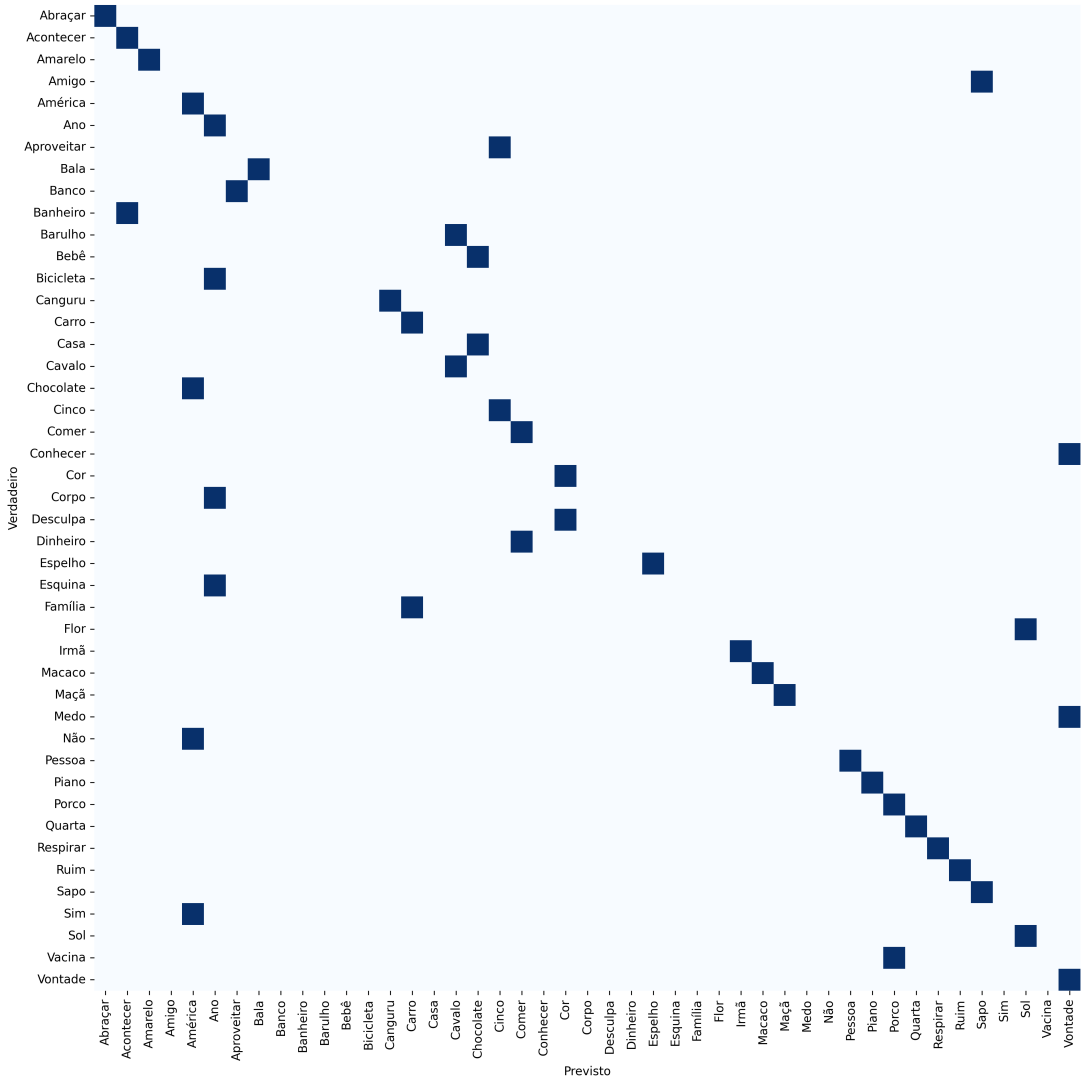
intrínsecas dos datasets. Estudos linguísticos comparativos entre LIBRAS e LSA (FISCHER et al., 2022) documentam que, embora sejam línguas historicamente e geneticamente distintas, ambas compartilham alguns processos de formação de sinais, como o uso de iconicidade baseada em referentes visuais e a incorporação de empréstimos linguísticos através de inicialização. Tais semelhanças estruturais podem refletir-se em features visuais compartilhadas extraídas por modelos de pose estimation, embora essa hipótese necessite de validação empírica em trabalhos futuros.

Dessa forma, a LSA64 se apresenta como uma escolha equilibrada quando não há conhecimento prévio sobre qual arquitetura ou métrica será mais relevante para a aplicação. Assim, a seleção dos pesos pré-treinados deve considerar não apenas a arquitetura da rede neural, mas também os objetivos específicos de desempenho do sistema de reconhecimento de sinais.

4.6 Análise das Matrizes de Confusão

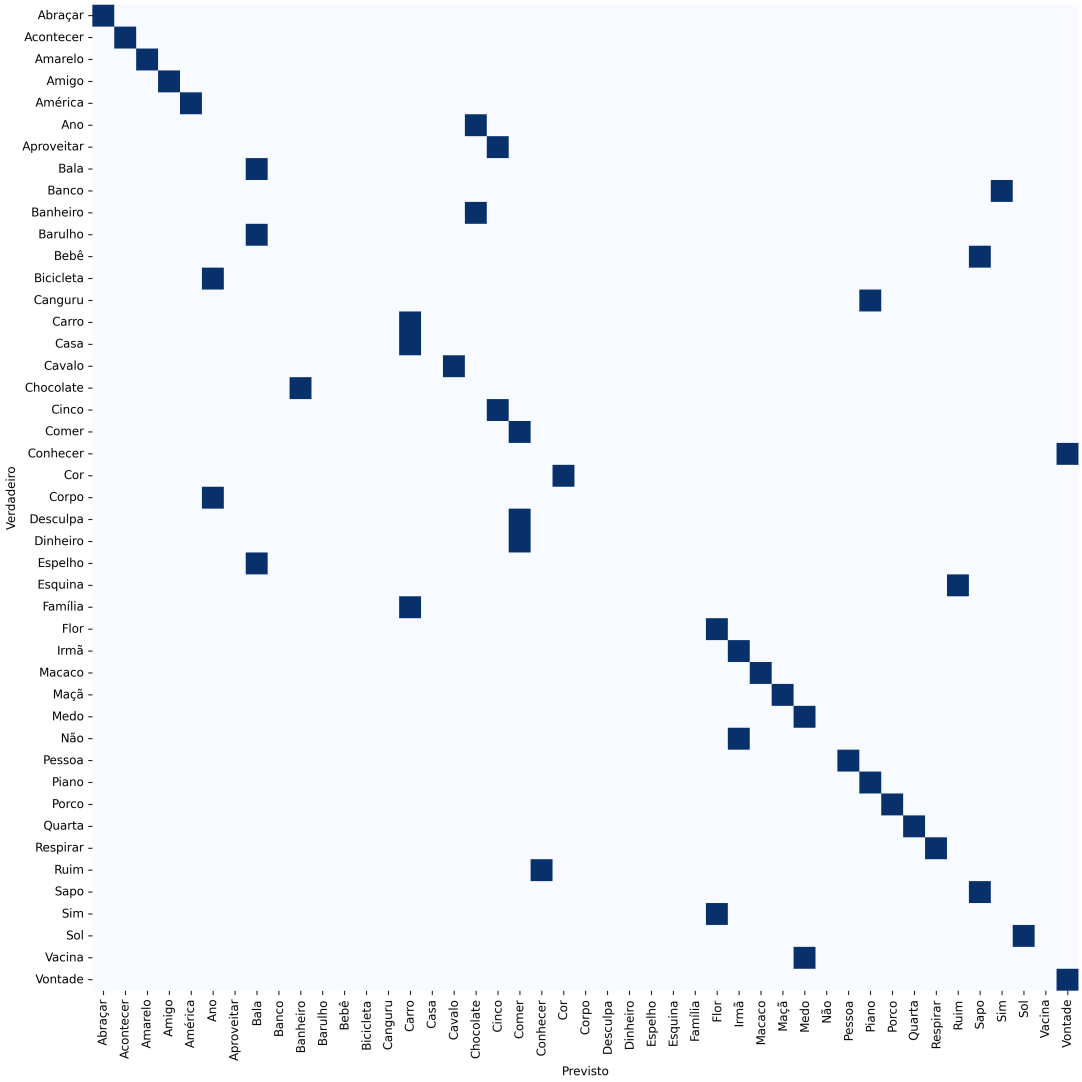
Nesta seção, são apresentadas as matrizes de confusão obtidas a partir dos experimentos realizados com os modelos Bi-LSTM, SL-GCN e I3D no reconhecimento de sinais em LIBRAS. As Figuras 10, 11 e 12 apresentam as matrizes correspondentes ao melhor resultado alcançado no top-1 por cada modelo durante os experimentos, permitindo visualizar de forma detalhada a relação entre os rótulos corretos e os rótulos previstos, evidenciando tanto os acertos quanto as confusões cometidas por cada arquitetura. Essa forma de representação gráfica é particularmente útil, pois destaca quais sinais foram corretamente reconhecidos e quais apresentaram maior dificuldade de classificação, revelando padrões que não seriam facilmente perceptíveis apenas por métricas globais de acurácia.

Figura 10 – Matriz de confusão do modelo Bi-LSTM (WLASL-2000).



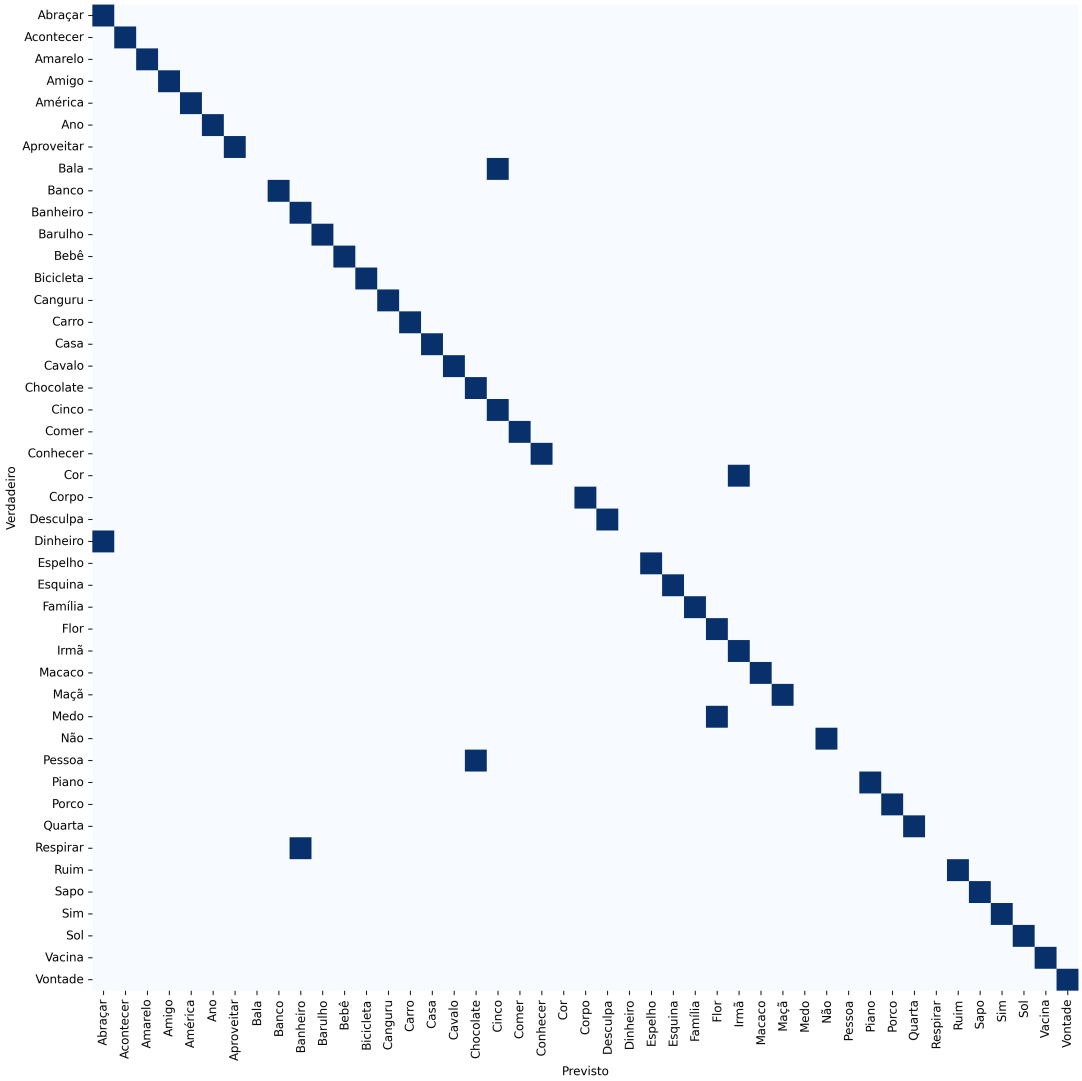
Fonte: Elaborada pela Autora.

Figura 11 – Matriz de confusão do modelo SL-GCN (GSL).



Fonte: Elaborada pela Autora.

Figura 12 – Matriz de confusão do modelo I3D (WLASL-1000).

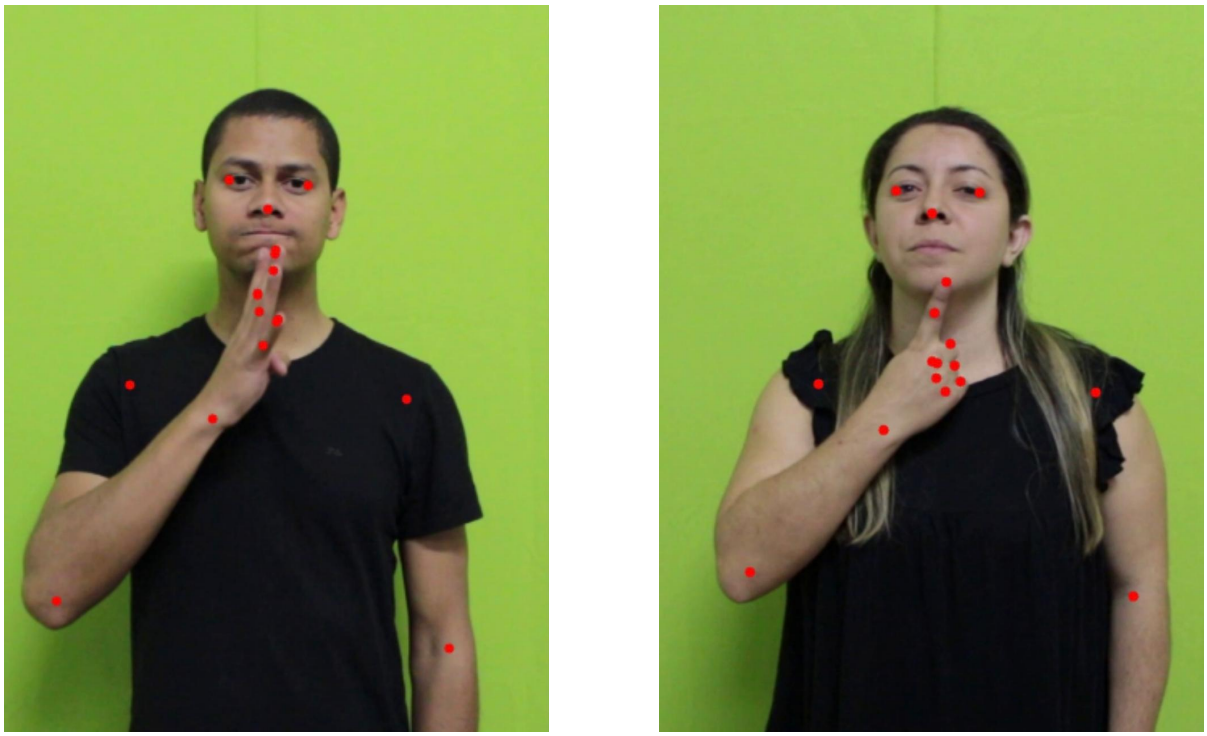


Fonte: Elaborada pela Autora.

Os resultados mostram que os três modelos tiveram desempenho consistente em várias classes. Sinais como *Abraçar*, *Acontecer*, *Amarelo*, *América*, *Carro*, *Cavalo*, *Cinco*, *Comer*, *Irmã*, *Maçã*, *Macaco*, *Piano*, *Porco*, *Quarta*, *Sapo*, *Sol* e *Vontade* foram classificados corretamente por todos os modelos, indicando que essas classes possuem características visuais distintivas e pouco ambíguas, facilitando seu reconhecimento independentemente da arquitetura utilizada.

Por outro lado, a análise das matrizes de confusão revelou padrões recorrentes de erro entre sinais visualmente semelhantes. Observou-se, por exemplo, que os modelos *Bi-LSTM* e *SL-GCN* classificaram o sinal *Conhecer* incorretamente como *Vontade*. A Figura 13 mostra esses dois sinais, em que ambos envolvem a mão posicionada próxima ao queixo, com contato direto na mesma região da face. Essa semelhança espacial, somada à configuração semelhante da mão, pode induzir confusões em modelos de reconhecimento automático, especialmente quando estes se baseiam em pontos de referência corporais para distinguir os gestos.

Figura 13 – Comparação entre os sinais "Conhecer" e "Vontade" com pontos-chave destacados (MINDS-Libras Dataset).

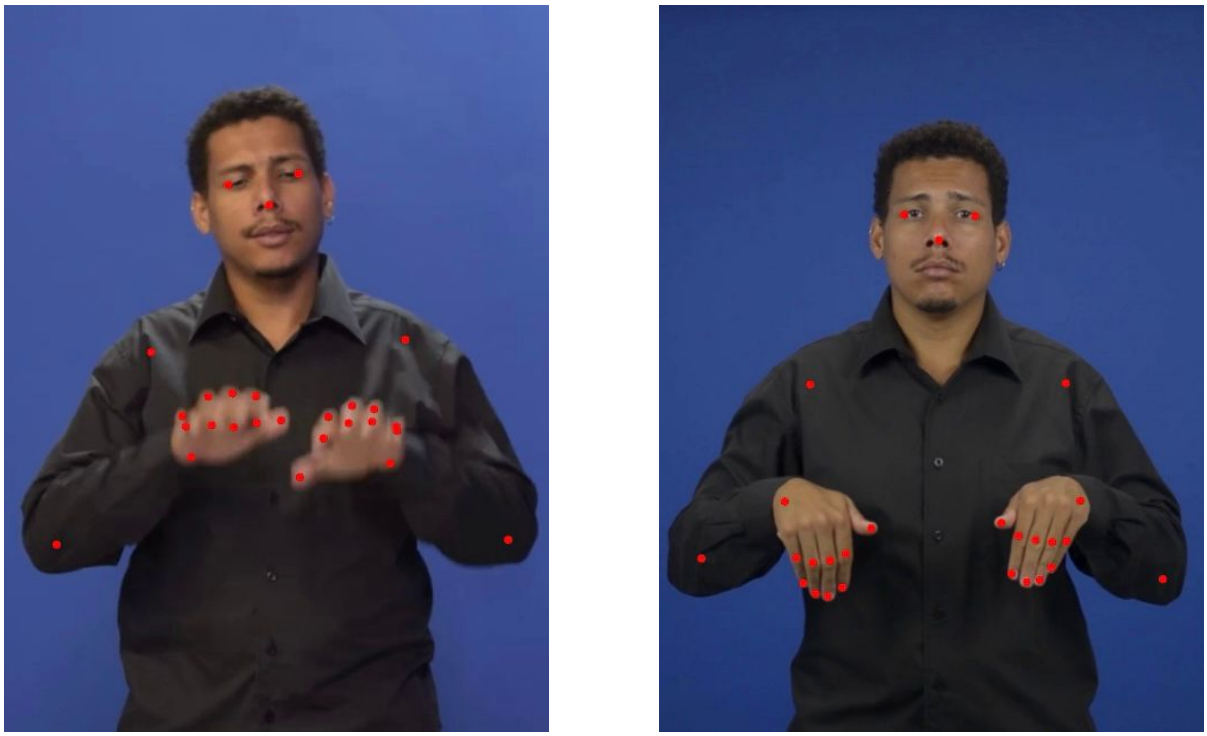


(a) Sinal: Conhecer (b) Sinal: Vontade

Fonte: Adaptado de (REZENDE, 2021).

Outro exemplo relevante de confusão visual ocorre entre os sinais *Piano* e *Canguru*, identificado pelo modelo SL-GCN. A Figura 14 ilustra esses dois sinais, nos quais as mãos estão posicionadas à frente do corpo, voltadas para baixo, com configuração de dedos semelhantes. A principal diferença está no movimento característico de cada sinal: em *Piano*, há uma alternância rítmica dos dedos que remete ao ato de tocar o instrumento, enquanto em *Canguru* o gesto se mantém mais estático, evocando o movimento das patas do animal. Para sistemas automáticos, essa proximidade visual pode gerar confusões, já que a distinção depende mais da dinâmica temporal do movimento do que da posição estática das mãos. Esse caso ilustra como gestos semelhantes em configuração manual e localização podem se diferenciar apenas pela trajetória e ritmo, desafiando especialmente modelos com menor sensibilidade ao aspecto temporal.

Figura 14 – Comparação entre os sinais "Piano" e "Canguru" com pontos-chave destacados (Libras SignBank).



(a) Sinal: Piano (b) Sinal: Canguru

Fonte: Adaptado de (PIZZIO et al., 2021).

Contudo, nem todas as confusões observadas apresentam justificativa visual evidente. A Figura 15 mostra o sinal *Dinheiro*, que o modelo Bi-LSTM classificou incorretamente como *Comer*, enquanto o modelo I3D o confundiu com *Abraçar*. Esses três sinais possuem características estruturais bastante distintas: *Dinheiro* é realizado com uma mão próxima ao rosto e movimento de esfregar os dedos; *Abraçar* envolve os braços cruzados à frente do corpo; e *Comer* apresenta a mão posicionada em frente à boca com movimento característico. Tais diferenças tornam os sinais facilmente distinguíveis para usuários fluentes de LIBRAS, sugerindo que os erros podem estar relacionados a limitações na captura ou processamento das características temporais e espaciais pelos modelos.

Figura 15 – Comparação entre os sinais “Dinheiro”, “Abraçar” e “Comer” com pontos-chave destacados (V-LIBRASIL).

(a) Sinal: Dinheiro (b) Sinal: Abraçar (c) Sinal: Comer

Fonte: Adaptado de (RODRIGUES, 2021).

Embora os sinais analisados sejam claramente distintos para usuários fluentes de LIBRAS, os padrões capturados pelos algoritmos podem não ser suficientemente discriminativos, resultando em erros recorrentes observados nas matrizes de confusão. Esse comportamento evidencia a importância de analisar não apenas os acertos dos modelos, mas também os padrões de confusão entre sinais, sejam eles visualmente semelhantes ou não. Tais análises revelam limitações importantes na capacidade de generalização dos modelos e contribuem para a identificação de aspectos que podem ser aprimorados em trabalhos futuros.

5 Conclusão

Este estudo investigou a aplicação do aprendizado por transferência para o reconhecimento automático de LIBRAS, comparando três arquiteturas distintas: I3D, Bi-LSTM e SL-GCN. O trabalho utilizou pesos pré-treinados em múltiplas línguas de sinais para treinar modelos em um *dataset* de vocabulário restrito de LIBRAS, explorando como conhecimentos adquiridos em diferentes contextos linguísticos podem ser adaptados à nossa língua de sinais nacional.

Os experimentos demonstraram que o aprendizado por transferência é uma estratégia essencial para o reconhecimento de sinais em LIBRAS, especialmente em cenários com limitação de dados. Todos os modelos apresentaram ganhos significativos de desempenho ao serem inicializados com pesos pré-treinados em outras línguas de sinais, comprovando que representações gestuais aprendidas em diferentes contextos podem ser reutilizadas de forma eficaz. Além disso, os experimentos revelaram que a complexidade da arquitetura influencia diretamente o impacto e a necessidade dos pesos pré-treinados.

Entre as três arquiteturas avaliadas, o modelo I3D apresentou desempenho substancialmente superior às demais, confirmando a eficácia das redes convolucionais 3D na captura de padrões espaço-temporais característicos da comunicação gestual. O SL-GCN e o Bi-LSTM apresentaram desempenhos comparáveis entre si, com o SL-GCN demonstrando maior consistência e melhor desempenho na maioria das fontes de pré-treinamento avaliadas, enquanto o Bi-LSTM alcançou desempenho ligeiramente superior especificamente com pesos do WLASL-2000. Esses resultados indicam que tanto representações baseadas em grafos esqueléticos quanto modelagem temporal sequencial são estratégias viáveis para o reconhecimento de LIBRAS, com a escolha entre elas dependendo da disponibilidade de pesos pré-treinados e das características específicas da aplicação pretendida.

A análise qualitativa das matrizes de confusão revelou padrões coerentes: sinais visualmente distintos foram reconhecidos de forma confiável pela maioria dos modelos, enquanto sinais com movimentos, configurações de mão ou localizações semelhantes apresentaram confusões recorrentes. Esse comportamento evidencia a dificuldade dos modelos em discriminar sutilezas visuais, mas também destaca que as representações aprendidas capturam informações relevantes sobre a estrutura dos sinais. Um resultado promissor foi observado nas métricas *Top-5* e *Top-10*, que apresentaram valores significativamente superiores ao *Top-1*, indicando que o rótulo correto frequentemente aparece entre as principais previsões. Esse padrão reforça o potencial de estratégias complementares, como o uso de contexto linguístico, atenção temporal e integração de informações não manuais, para refinar a desambiguação entre sinais semelhantes.

Do ponto de vista teórico, este trabalho demonstra a viabilidade da transferência de aprendizado entre línguas de sinais estruturalmente distintas. Os resultados comprovam que é possível transferir conhecimento entre línguas de sinais independentemente de proximidade geográfica, raízes linguísticas comuns ou similaridades culturais. Essa descoberta amplia as possibilidades de pesquisa na área, indicando que *datasets* e modelos desenvolvidos para qualquer língua de sinais podem beneficiar investigações em outras. Além disso, o estudo evidencia que a complexidade da arquitetura influencia o quanto os pesos pré-treinados são necessários e benéficos para o desempenho final, contribuindo para o entendimento teórico sobre como projetar e otimizar sistemas de reconhecimento de línguas de sinais.

Sob a perspectiva prática, os resultados indicam que sistemas de reconhecimento de LIBRAS podem ser desenvolvidos com recursos computacionais moderados, aproveitando *datasets* públicos e modelos pré-treinados em outras línguas de sinais. Essa abordagem reduz substancialmente as barreiras de entrada para pesquisa e desenvolvimento, viabilizando ferramentas educacionais, sistemas de legendagem assistida e interfaces de comunicação voltadas à inclusão da comunidade surda. Embora desafios técnicos persistam — como ampliação de vocabulário, aumento da diversidade dos dados e reconhecimento contínuo —, os avanços apresentados demonstram que sistemas práticos e acessíveis são alcançáveis no médio prazo.

5.1 Limitações e Trabalhos Futuros

Este trabalho apresenta limitações que devem ser consideradas na interpretação dos resultados. O *dataset* utilizado possui tamanho reduzido, tanto em número de classes (45 sinais) quanto em amostras por classe, o que não reflete a riqueza e complexidade da LIBRAS, composta por milhares de sinais. Além disso, a homogeneidade dos vídeos — quanto aos sinalizantes, ambientes e condições de captura — pode ter levado à superestimação do desempenho, já que os modelos foram avaliados em cenários controlados. Em aplicações reais, espera-se maior variabilidade nas condições de gravação, nas características físicas dos sinalizantes e na execução dos sinais.

Outra limitação metodológica é o uso de vídeos segmentados contendo sinais isolados. Esse cenário simplificado não contempla o desafio do reconhecimento contínuo, no qual o sistema deve detectar automaticamente os limites temporais entre sinais, distinguir movimentos de transição e manter coerência linguística ao longo da sequência. Assim, os resultados obtidos representam um passo inicial, mas ainda distante das condições reais de uso.

Para superar essas limitações, recomenda-se expandir significativamente o *dataset* de LIBRAS, incorporando um vocabulário mais amplo e aumentando o número de amostras por classe. É fundamental também diversificar as condições de captura, incluindo diferentes sinalizantes, ambientes, iluminações e ângulos de câmera, de modo a garantir maior representatividade e robustez dos modelos. Adicionalmente, sugere-se investigar mecanismos

mais sofisticados para discriminar sinais visualmente semelhantes, considerando o contexto das sequências gestuais e incorporando informações complementares que auxiliem na distinção entre sinais com características próximas.

Um avanço essencial para futuras aplicações é a adaptação das arquiteturas ao reconhecimento contínuo, com mecanismos capazes de identificar automaticamente os limites entre sinais consecutivos e processar fluxos de vídeo em tempo real. A validação dos modelos em cenários reais, com sinalizantes nativos e situações naturais de comunicação, é igualmente necessária para garantir a robustez e a utilidade prática das soluções propostas.

Em síntese, apesar das limitações identificadas e dos desafios que ainda precisam ser superados, este trabalho fornece evidências sólidas de que o aprendizado por transferência é uma estratégia viável e promissora para o reconhecimento automático de LIBRAS. Os resultados obtidos, aliados às direções futuras delineadas, estabelecem uma base consistente para o avanço contínuo da área. O progresso nessa direção representa não apenas um avanço técnico, mas uma contribuição concreta para a inclusão social, promovendo maior autonomia, acessibilidade e participação da comunidade surda na sociedade.

Referências

ADALOGLOU, N. et al. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE transactions on multimedia*, IEEE, v. 24, p. 1750–1762, 2021.

AL-SELWI, S. M. et al. Rnn-lstm: From applications to modeling techniques and beyond—systematic review. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, v. 36, n. 5, p. 102068, 2024. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S1319157824001575⟩.

BIRD, J. J.; EKáRT, A.; FARIA, D. R. British sign language recognition via late fusion of computer vision and leap motion with transfer learning to american sign language. *Sensors*, v. 20, n. 18, 2020. ISSN 1424-8220. Disponível em: ⟨https://www.mdpi.com/1424-8220/20/18/5151⟩.

CAMGOZ, N. C. et al. Sign language transformers: Joint end-to-end sign language recognition and translation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2020. p. 10023–10033.

CARREIRA, J.; ZISSERMAN, A. Quo vadis, action recognition? a new model and the kinetics dataset. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2017. p. 6299–6308.

CARVALHO, A. et al. Inteligência artificial–uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, v. 2, p. 45, 2011.

CHAI, X.; WANG, H.; CHEN, X. The devisign large vocabulary of chinese sign language database and baseline evaluations. In: INSTITUTE OF COMPUTING TECHNOLOGY. *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)*. [S.l.], 2014.

CHARAYAPHAN, C.; MARBLE, A. Image processing system for interpreting motion in american sign language. *Journal of Biomedical Engineering*, v. 14, n. 5, p. 419–425, 1992. ISSN 0141-5425. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/0141542592900883⟩.

COOPER, H. et al. Sign language recognition using sub-units. *The Journal of Machine Learning Research*, JMLR. org, v. 13, n. 1, p. 2205–2231, 2012.

DAVILA, A.; COLAN, J.; HASEGAWA, Y. Comparison of fine-tuning strategies for transfer learning in medical image classification. *Image and Vision Computing*, Elsevier, v. 146, p. 105012, 2024.

DELUCIS, M. M. et al. *Isolated sign language recognition in LIBRAS*. Dissertação (Mestrado), 2025. Escola Politécnica. Disponível em: ⟨https://tede2.pucrs.br/tede2/handle/tede/11629⟩.

DONGARE, A. et al. Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, Citeseer, v. 2, n. 1, p. 189–194, 2012.

ESTEVAM, W.; MENOTTI, D. Towards a realistic libras to portuguese translation. In: IEEE. *2023 36th Conference on Graphics, Patterns and Images (SIBGRAPI)*. 2023. Disponível em: ⟨https://ieeexplore.ieee.org/document/10347035⟩.

FARHADI, A.; FORSYTH, D.; WHITE, R. Transfer learning in sign language. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2007. p. 1–8.

FAROOQ, U. et al. Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications*, Springer, 2021. Published online: 18 May 2021. ⟨https://link.springer.com/article/10.1007/s00521-021-06079-3⟩.

FETZER, J. H. What is artificial intelligence? In: _____. *Artificial Intelligence: Its Scope and Limits*. Dordrecht: Springer Netherlands, 1990. p. 3–27. ISBN 978-94-009-1900-6. Disponível em: ⟨https://doi.org/10.1007/978-94-009-1900-6_1⟩.

FISCHER, C. et al. Estudio comparativo de señas que designan colores en libras y en lsa. *Quintú Quimün: revista de lingüística*, Universidad Nacional del Comahue, n. 6, p. 51, 2022.

GEDIEL, A. L. B.; OLIVEIRA, B. S. B. de; RODRIGUES, L. M. Dicionário online bilíngue libras/português: reflexões sobre a formação de professores do ensino superior. *Fórum Linguístico*, v. 18, n. 3, p. 6495–6513, 2021.

HAN, X. et al. Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing*, v. 9, n. 8, 2017. ISSN 2072-4292. Disponível em: ⟨https://www.mdpi.com/2072-4292/9/8/848⟩.

HAYKIN, S. *Neural networks and learning machines, 3/E*. [S.l.]: Pearson Education India, 2009.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997.

HUANG, J.; CHOUVATUT, V. Video-based sign language recognition via resnet and lstm network. *Journal of Imaging*, v. 10, n. 6, 2024. ISSN 2313-433X. Disponível em: ⟨https://www.mdpi.com/2313-433X/10/6/149⟩.

HUANG, J. et al. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 29, n. 9, p. 2822–2832, 2018.

INOVAR+, U. F. de V. . . P. *Dicionário de Libras*. 2017. ⟨https://sistemas.cead.ufv.br/capes/dicionario/⟩. Acesso em: 15 maio 2025.

JIANG, S. et al. Skeleton aware multi-modal sign language recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2021. p. 3413–3423.

JIANG, X. et al. A survey on chinese sign language recognition: From traditional methods to artificial intelligence. *CMES - Computer Modeling in Engineering and Sciences*, v. 140, n. 1, p. 1–40, 2024. ISSN 1526-1492. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S1526149224001577⟩.

KHAN, A. et al. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, Springer, v. 53, n. 8, p. 5455–5516, 2020.

KIPF, T. N.; WELLING, M. Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2017.

KONG, W.; RANGANATH, S. Automatic hand trajectory segmentation and phoneme transcription for sign language. In: *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*. [S.l.: s.n.], 2008. p. 1–6.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998.

LI, C. et al. Semantic information guided multimodal skeleton-based action recognition. *Information Fusion*, Elsevier, p. 103289, 2025.

LI, D. et al. 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*. [S.l.: s.n.], 2020. p. 1459–1469.

MAHESH, B. et al. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, n. 1, p. 381–386, 2020.

MOCIALOV, B.; TURNER, G.; HASTIE, H. Transfer learning for british sign language modelling. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*. Santa Fe, New Mexico, USA, 2018. p. 101–110. ⟨https://aclanthology.org/W18-3911.pdf⟩.

ONGSULEE, P. Artificial intelligence, machine learning and deep learning. In: *2017 15th International Conference on ICT and Knowledge Engineering (ICTKE)*. [S.l.: s.n.], 2017. p. 1–6.

Organização Mundial da Saúde. *Deafness and hearing loss*. 2025. Acesso em: 21 abr. 2025. Disponível em: ⟨https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss⟩.

O'SHEA, K.; NASH, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 22, n. 10, p. 1345–1359, 2010.

PERUMAL, T. et al. A comprehensive overview and comparative analysis on deep learning models. *Journal on Artificial Intelligence*, Tech Science Press, v. 6, n. 1, p. 301–360, 2024. ISSN 2579-003X. Disponível em: ⟨http://dx.doi.org/10.32604/jai.2024.054314⟩.

PIZZIO, A. et al. Signbank da libras. *Fórum Linguístico*, v. 17, p. 5475–5487, 02 2021. Disponível em: ⟨https://periodicos.ufsc.br/index.php/forum/article/view/77342⟩.

PRESTES, W. P. Um sistema de reconhecimento de sinais isolados de libras utilizando mediapipe holistic e lstm. Universidade Estadual Paulista (Unesp), 2024.

REZENDE, T. M. Reconhecimento automático de sinais da libras: desenvolvimento da base de dados minds-libras e modelos de redes convolucionais. 2021. Disponível em: ⟨https://hdl.handle.net/1843/39785⟩.

RODRIGUES, A. J. *V-LIBRASIL: uma base de dados com sinais na Língua Brasileira de Sinais (Libras)*. Dissertação (Dissertação de Mestrado) — Universidade Federal de Pernambuco (UFPE), Recife, Brasil, 2021. Disponível em: ⟨https://repositorio.ufpe.br/handle/123456789/43491⟩.

RONCHETTI, F. et al. Lsa64: A dataset of argentinian sign language. *XX II Congreso Argentino de Ciencias de la Computación (CACIC)*, 2016.

SARMENTO, A. H. d. A. *Integração de Datasets de Vídeo para Tradução Automática da LIBRAS com Aprendizado Profundo*. Tese (Doutorado) — Universidade de SÃ£ o Paulo, 2023.

SARMENTO, M. d. A.; PONTI, M. A cross-dataset study on the brazilian sign language translation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. [s.n.], 2023. p. 3752–3761. Disponível em: ⟨https://openaccess.thecvf.com/content/ICCV2023W/CLVL/html/de_Avellar_Sarmento_A_Cross-Dataset_Study_on_the_Brazilian_Sign_Language_Translation_ICCVW_2023_paper.html⟩.

SELVARAJ, P. et al. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. *arXiv preprint arXiv:2110.05877*, 2021.

SHARMA, S.; KUMAR, K. Asl-3dcnn: American sign language recognition technique using 3-d convolutional neural networks. *Multimedia Tools and Applications*, Springer Science+Business Media, LLC, part of Springer Nature, Department of Computer Science and Engineering, National Institute of Technology Uttarakhand, Srinagar, Garhwal, India, v. 80, n. 26, p. 26319–26331, 2021. Received: 12 June 2020 / Revised: 8 February 2021 / Accepted: 25 February 2021 / Published online: 1 May 2021. Disponível em: ⟨https://link.springer.com/article/10.1007/s11042-021-10897-2⟩.

SHEN, Z. et al. Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2021. v. 35, n. 11, p. 9594–9602.

SHI, L. et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 12026–12035.

SI, C. et al. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 1227–1236.

SIMONYAN, K.; ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, v. 27, 2014.

SINCAN, O. M.; KELES, H. Y. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, v. 8, p. 181340–181355, 2020.

SMAGULOVA, K.; JAMES, A. P. A survey on lstm memristive neural network architectures and applications. *The European Physical Journal Special Topics*, v. 228, n. 10, p. 2313–2324, 2019. ISSN 1951-6401. Disponível em: ⟨https://doi.org/10.1140/epjst/e2019-900046-x⟩.

SRIDHAR, A. et al. Include: A large scale dataset for indian sign language recognition. In: . [S.l.]: Association for Computing Machinery, 2020. (MM '20). ISBN 9781450379885.

TAMURA, S.; KAWASAKI, S. Recognition of sign language motion images. *Pattern Recognition*, v. 21, n. 4, p. 343–353, 1988. ISSN 0031-3203. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/0031320388900489⟩.

VERLEYSEN, M.; FRANÇOIS, D. The curse of dimensionality in data mining and time series prediction. In: CABESTANY, J.; PRIETO, A.; SANDOVAL, F. (Ed.). *Computational Intelligence and Bioinspired Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 758–770. ISBN 978-3-540-32106-4.

WADHAWAN, A.; KUMAR, P. Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, Springer, v. 28, n. 2, p. 785–813, 2021. Received 14 May 2019; Accepted 08 Dec 2019; Published 17 Dec 2019. Disponível em: ⟨https://doi.org/10.1007/s11831-019-09384-2⟩.

WALDRON, M. B.; KIM, S. Isolated asl sign recognition system for deaf persons. *IEEE Transactions on rehabilitation engineering*, IEEE, v. 3, n. 3, p. 261–271, 1995.

WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. *Journal of Big data*, Springer, v. 3, p. 1–40, 2016.

WU, Y.-c.; FENG, J.-w. Development and application of artificial neural network. *Wireless Personal Communications*, Springer, v. 102, p. 1645–1656, 2018.

YAMASHITA, R. et al. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, Springer, v. 9, n. 4, p. 611–629, 2018.

YAN, S.; XIONG, Y.; LIN, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: *Thirty-second AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2018.

YIN, K.; READ, J. *Better Sign Language Translation with STMC-Transformer*. 2020. Disponível em: ⟨https://arxiv.org/abs/2004.00588⟩.

YIN, P. et al. Learning the basic units in american sign language using discriminative segmental feature selection. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2009. p. 4757–4760.

YOSINSKI, J. et al. How transferable are features in deep neural networks? In: *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2014. v. 27, p. 3320–3328.

ZAFRULLA, Z. et al. American sign language recognition with the kinect. In: *Proceedings of the 13th international conference on multimodal interfaces*. [S.l.: s.n.], 2011. p. 279–286.

ZHANG, C. et al. *Understanding deep learning requires rethinking generalization*. 2017. Disponível em: ⟨https://arxiv.org/abs/1611.03530⟩.

ZIMMERMAN, T. et al. A hand gesture interface device. In: *Proceeding CHI 87: Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics*. [S.l.: s.n.], 1987. p. 189–192. ⟨https://dl.acm.org/doi/10.1145/29933.275628⟩.