

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA E INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM ENGENHARIA COMPUTACIONAL

Pedro de Araújo Bhering Bittencourt

**Classificação de sinais dinâmicos de Libras relacionados ao contexto hospitalar
utilizando redes neurais recorrentes**

Juiz de Fora
2026

Pedro de Araújo Bhering Bittencourt

**Classificação de sinais dinâmicos de Libras relacionados ao contexto hospitalar
utilizando redes neurais recorrentes**

Monografia apresentada ao Curso de Graduação em Engenharia Computacional da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Engenharia Computacional.

Orientador: Prof. Dr. Luiz Maurílio da Silva Maciel

Juiz de Fora
2026

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

A. B. Bittencourt, Pedro de.

Classificação de sinais dinâmicos de Libras relacionados ao contexto hospitalar utilizando redes neurais recorrentes / Pedro de Araújo Bhering Bittencourt. – 2026.

78 f. : il.

Orientador: Luiz Maurílio da Silva Maciel

Trabalho de Conclusão de Curso – Universidade Federal de Juiz de Fora, Faculdade de Engenharia e Instituto de Ciências Exatas. Bacharelado em Engenharia Computacional, 2026.

1. Aprendizado de máquina. 2. Inclusão social. 3. Visão computacional.
4. Língua de sinais. I. Maciel, Luiz Maurílio da Silva, orient. II. Título.

Pedro de Araújo Bhering Bittencourt

**Classificação de sinais dinâmicos de Libras relacionados ao contexto hospitalar
utilizando redes neurais recorrentes**

Monografia apresentada ao Curso de Graduação em Engenharia Computacional da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Engenharia Computacional.

Aprovada em 13 de Janeiro de 2026

BANCA EXAMINADORA

Prof. Dr. Luiz Maurílio da Silva Maciel - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Marcelo Bernardes Vieira
Universidade Federal de Juiz de Fora

Prof. Dr. Saulo Moraes Villela
Universidade Federal de Juiz de Fora

Dedico este trabalho a todos que estiveram ao meu lado durante minha jornada, especialmente aos meus pais, madrastra e irmãos, bem como meus amigos. O apoio, a confiança e o incentivo de cada um foram essenciais para que esta conquista se tornasse possível.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por ter me concedido inúmeras oportunidades ao longo desta trajetória, permitindo-me aproveitá-las da melhor forma possível. Expresso minha profunda gratidão à minha família, que nunca mediu esforços para me apoiar e auxiliar em cada etapa dessa jornada. Ao meu pai, agradeço pela transmissão de valores e princípios que moldaram não apenas meu percurso acadêmico, mas também minha vida. À minha mãe, sou imensamente grato pelo apoio constante em todos os momentos e pela contribuição direta no desenvolvimento deste trabalho, participando das gravações do conjunto de dados adicional aqui apresentado.

Aos meus amigos Gustavo Coelho e João Pedro Banhato, que me acompanharam durante a graduação e me auxiliaram no processo de gravação do conjunto de dados apresentado neste trabalho. Agradeço também aos meus amigos Ricardo Silva, Lucas Santana e Yan Fonseca pelo companheirismo ao longo das disciplinas da graduação.

Aos professores Flávia Bastos e Rodrigo Weber, que possibilitaram o início da pesquisa do tema apresentado neste trabalho durante a disciplina Trabalho Multidisciplinar.

Ao professor orientador Luiz Maurílio, expresso minha sincera gratidão por todas as oportunidades oferecidas ao longo desta etapa acadêmica, pela paciência, pelas orientações e pela oportunidade de ser seu orientando. Sou grato não apenas pelo apoio à escolha do tema deste trabalho, mas também pelas revisões criteriosas e pelos incentivos constantes durante todo o desenvolvimento do TCC.

Gostaria de expressar meu sincero agradecimento ao Grupo de Estudos Linguísticos da Libras (GELLI) da Faculdade de Letras da Universidade Federal de Juiz de Fora. Agradeço à Profa. Hadassa Rodrigues Santos, Profa. Aline Gracia Rodero Takahira, Prof. Lucas Alves Mendes e aos bolsistas Deivin Vasconcelos dos Santos, Maria Vitória Paulino Quitella, Samira Lúcio da Silva, Ronaldo Adriano Guimarães, Maria Eduarda Castellões Senna e Mirella de Oliveira Pena Araújo, que participaram da construção do conjunto de dados e forneceram sugestões fundamentais para o desenvolvimento deste trabalho.

À Universidade Federal de Juiz de Fora, pela oferta de ensino público de qualidade e pela contribuição significativa à minha formação acadêmica.

Aos meus amigos de fora da Universidade Federal de Juiz de Fora, expresso também meu agradecimento por todo o apoio e incentivo ao longo dessa trajetória.

Por fim, estendo meus agradecimentos a todos aqueles que, mesmo não mencionados individualmente, apoiaram e contribuíram direta ou indiretamente para a realização deste trabalho. O incentivo de cada um foi fundamental para a concretização deste importante marco da minha vida.

“Artificial intelligence is not a substitute for human intelligence; it is a tool to amplify human creativity and ingenuity.” - (Fei-Fei Li)

RESUMO

A comunicação limitada entre profissionais de saúde e pacientes surdos ainda representa um obstáculo significativo para a garantia de um atendimento inclusivo e equitativo, especialmente em contextos clínicos nos quais a rapidez e a precisão das informações são essenciais. Nesse cenário, este trabalho propõe um método para a classificação de dez sinais dinâmicos da Língua Brasileira de Sinais (Libras) relacionados ao domínio da saúde, abrangendo desde a construção de um conjunto de dados até o desenvolvimento e a avaliação de modelos de aprendizado de máquina. O conjunto de dados elaborado contou com a participação de profissionais fluentes em Libras, assegurando qualidade linguística e diversidade de execução. Foram gravados 492 vídeos correspondentes a dez sinais comumente utilizados em contextos hospitalares, executados por nove indivíduos sob condições controladas, com variações de vestimenta, aparência e repetição dos gestos, de modo a ampliar a variabilidade das amostras. O pré-processamento incluiu a seleção de quadros por amostragem probabilística e a extração de marcos tridimensionais das mãos e da pose corporal, os quais correspondem a pontos de referência anatômicos que descrevem a configuração e o movimento do corpo ao longo do tempo. Esses marcos foram organizados em sequências temporais, preservando a dinâmica dos gestos. A partir desses dados, foram investigadas diferentes abordagens de classificação baseadas em modelos de redes neurais com camadas de memória de longo e curto prazo (*Long Short-Term Memory* – LSTM), capazes de modelar as dependências temporais presentes nas sequências de movimentos dos sinais. Inicialmente, o problema foi formulado com dez classes. A análise das matrizes de confusão revelou confusões pontuais entre pares específicos de sinais, o que motivou a adoção de uma estratégia hierárquica baseada em modelos especialistas, na qual o problema foi reorganizado em sete classes, complementadas por classificadores dedicados aos pares conflituosos. Essa estratégia resultou em melhorias significativas na acurácia e na estabilidade dos resultados, alcançando média de 92,44% nos experimentos. A avaliação em um conjunto de dados de teste adicional confirmou a boa capacidade de generalização do sistema, com média de acurácia de 98,92% e baixo desvio padrão. Os resultados indicam que a combinação de um conjunto de dados bem estruturado, técnicas adequadas de pré-processamento e uma arquitetura hierárquica de modelos especialistas constitui uma solução promissora para o reconhecimento de sinais de Libras, contribuindo para a redução de barreiras comunicacionais e com potencial aplicação em sistemas de apoio à comunicação em ambientes de saúde.

Palavras-chave: Aprendizado de máquina; Inclusão social; Visão computacional; Língua de sinais.

ABSTRACT

Limited communication between healthcare professionals and deaf patients still represents a significant obstacle to ensuring inclusive and equitable care, especially in clinical contexts in which the speed and accuracy of information are essential. In this scenario, this work proposes a method for the classification of ten dynamic signs of Brazilian Sign Language (Libras) related to the healthcare domain, encompassing the construction of a dataset as well as the development and evaluation of machine learning models. The constructed dataset involved the participation of professionals fluent in Libras, ensuring linguistic quality and diversity of execution. A total of 492 videos were recorded, corresponding to ten signs commonly used in hospital contexts, performed by nine individuals under controlled conditions, with variations in clothing, appearance, and gesture repetition in order to increase sample variability. The preprocessing stage included frame selection through probabilistic sampling and the extraction of three-dimensional hand and body pose landmarks, which correspond to anatomical reference points that describe the configuration and movement of the body over time. These landmarks were organized into temporal sequences, preserving the dynamics of the gestures. Based on this data, different classification approaches were investigated using neural network models with Long Short-Term Memory (LSTM) layers, capable of modeling the temporal dependencies present in the sequences of sign movements. Initially, the problem was formulated with ten classes. The analysis of confusion matrices revealed specific confusions between pairs of signs, which motivated the adoption of a hierarchical strategy based on specialist models, in which the problem was reorganized into seven classes, complemented by classifiers dedicated to the conflicting pairs. This strategy resulted in significant improvements in accuracy and result stability, achieving an average accuracy of 92.44% in the experiments. Evaluation on an additional test dataset confirmed the system's good generalization capability, with an average accuracy of 98.92% and a low standard deviation. The results indicate that the combination of a well-structured dataset, appropriate preprocessing techniques, and a hierarchical architecture of specialist models constitutes a promising solution for Libras sign recognition, contributing to the reduction of communication barriers and with potential application in healthcare communication support systems.

Keywords: Machine learning; Social inclusion; Computer vision; Sign language.

LISTA DE FIGURAS

Figura 1 – Marcos faciais.	18
Figura 2 – Marcos da mão.	18
Figura 3 – Marcos de pose.	19
Figura 4 – Operações realizadas no <i>perceptron</i>	20
Figura 5 – Problema de classificação XOR em 2D.	21
Figura 6 – Representação da estrutura de uma rede MLP	22
Figura 7 – Representação esquemática de uma rede neural recorrente (RNN).	24
Figura 8 – Expansão temporal de uma RNN evidenciando o compartilhamento de parâmetros.	24
Figura 9 – Diagrama de uma célula LSTM.	26
Figura 10 – Exemplo de enquadramento respeitando as configurações definidas.	33
Figura 11 – Histograma dos índices dos quadros amostrados utilizando distribuição normal.	36
Figura 12 – Marcos da face, pose e mãos estimados pelo <i>MediaPipe Holistic</i>	41
Figura 13 – Diagrama de uma célula LSTM com <i>dropout</i> recorrente.	42
Figura 14 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 14) para o problema de 10 classes.	51
Figura 15 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) para o problema de 7 classes.	56
Figura 16 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) para o problema de classificação dos sinais “Bronquite” e “Infarto”.	60
Figura 17 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 13) para o problema de classificação dos sinais “Câimbra” e “Fratura”.	64
Figura 18 – Gráfico das posições das mãos ao longo da execução do sinal com identificação do início do movimento.	66
Figura 19 – Gráfico da energia do movimento calculada ao longo da execução do movimento com início do movimento e limiar identificados.	66
Figura 20 – Matriz de confusão geral da validação cruzada para o problema de classificação dos sinais “Alergia” e “Frieira”.	68
Figura 21 – Matriz de confusão geral da validação cruzada utilizando estratégia de modelos especialistas.	72

LISTA DE TABELAS

Tabela 1 – Distribuição do número de vídeos por sinal no conjunto de dados de Libras construído neste trabalho.	32
Tabela 2 – Relação do número de parâmetros pelo tipo de camada.	43
Tabela 3 – Parâmetros avaliados no treinamento dos modelos.	45
Tabela 4 – Estrutura de uma matriz de confusão para um problema de classificação binária.	46
Tabela 5 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de 10 classes.	48
Tabela 6 – Resultados da acurácia, geral e por classe, em porcentagem, para cada <i>fold</i> da validação cruzada para o conjunto que obteve melhor acurácia geral (conjunto 14) do problema de 10 classes.	49
Tabela 7 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 14) do problema de 10 classes.	50
Tabela 8 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de 7 classes.	53
Tabela 9 – Resultados da acurácia geral e por classe, em porcentagem, para cada <i>fold</i> da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) do problema de 7 classes.	54
Tabela 10 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 20) do problema de 7 classes.	55
Tabela 11 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de classificação dos sinais “Bronquite” e “Infarto”.	58
Tabela 12 – Resultados da acurácia geral e por classe, em porcentagem para cada <i>fold</i> da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) do problema de classificação dos sinais “Bronquite” e “Infarto”. . .	59
Tabela 13 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 20) do problema de classificação dos sinais “Bronquite” e “Infarto”.	60
Tabela 14 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de classificação dos sinais “Câimbra” e “Fratura”.	61
Tabela 15 – Resultados da acurácia geral e por classe, em porcentagem, para cada <i>fold</i> da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 13) do problema de classificação dos sinais “Câimbra” e “Fratura”. . . .	62

Tabela 16 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 13) do problema de classificação dos sinais “Câimbra” e “Fratura”.	63
Tabela 17 – Resultados da acurácia, em porcentagem, e estatísticas por classe para cada <i>fold</i> da validação cruzada para o problema de classificação dos sinais “Alergia” e “Frieira”.	67
Tabela 18 – Avaliação de desempenho por classe para o problema de classificação dos sinais “Alergia” e “Frieira”.	68
Tabela 19 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada utilizando estratégia de modelos especialistas.	69
Tabela 20 – Resultados da acurácia geral e por classe, em porcentagem para cada <i>fold</i> da validação cruzada utilizando estratégia de modelos especialistas.	70
Tabela 21 – Avaliação de desempenho por classe utilizando estratégia de modelos especialistas.	71
Tabela 22 – Acurácia geral obtida em cada um dos 20 sorteios do conjunto de teste adicional.	74

LISTA DE ABREVIATURAS E SIGLAS

CMC	Carpometacarpianas
DIP	Interfalangeana distal
FPS	<i>Frames per second</i>
GELLI	Grupo de Estudos Linguísticos da Libras
HD	<i>High Definition</i>
IP	Interfalangeana
Libras	Língua Brasileira de Sinais
LSTM	<i>Long Short-Term Memory</i>
MCP	Metacarpolangeanas
MLP	<i>Multilayer Perceptron</i>
PIP	Interfalangeana proximal
ReLU	<i>Rectified Linear Unit</i>
ROI	<i>Region of Interest</i>
RNN	<i>Recurrent Neural Network</i>
UFJF	Universidade Federal de Juiz de Fora
ViT	<i>Vision Transformer</i>
YOLO	<i>You Only Look Once</i>

SUMÁRIO

1	INTRODUÇÃO	13
1.1	DEFINIÇÃO DO PROBLEMA	15
1.2	OBJETIVOS	15
2	FUNDAMENTAÇÃO TEÓRICA	17
2.1	LIBRAS	17
2.2	MARCOS CORPORAIS	17
2.3	APRENDIZADO EM REDES NEURAIIS	19
2.4	REDES NEURAIIS RECORRENTES	23
3	TRABALHOS RELACIONADOS	28
3.1	CONJUNTOS DE DADOS	28
3.2	RECONHECIMENTO DE LÍNGUAS DE SINAIS	29
4	CONJUNTO DE DADOS	31
4.1	DESCRIÇÃO GERAL	31
4.2	AQUISIÇÃO DOS VÍDEOS	32
4.3	CONJUNTO DE DADOS DE TESTE ADICIONAL	34
5	MÉTODO PROPOSTO	35
5.1	PRÉ-PROCESSAMENTO DOS DADOS	35
5.1.1	Seleção de quadros	35
5.1.2	Aumento de Dados	37
5.2	PROCESSAMENTO E EXTRAÇÃO DOS MARCOS CORPORAIS	38
5.3	MODELO DE REDE NEURAL	41
6	EXPERIMENTOS E RESULTADOS	44
6.1	AMBIENTE COMPUTACIONAL	44
6.2	PROTOCOLO DE TREINAMENTO	44
6.3	CRITÉRIOS DE AVALIAÇÃO	45
6.4	RESULTADOS	48
6.4.1	Problema de classificação de 10 classes	48
6.4.2	Problema de classificação de 7 classes	52
6.4.3	Problemas de classificação utilizando especialistas	57
6.4.4	Integração do modelo de 7 classes com os modelos especialistas	69
6.4.5	Avaliação dos modelos no conjunto de dados de teste adicional	73
7	CONCLUSÃO	75
	REFERÊNCIAS	76

1 INTRODUÇÃO

A Língua Brasileira de Sinais (Libras) desempenha um papel de fundamental relevância na sociedade, uma vez que contribui para a redução das barreiras comunicativas enfrentadas pela comunidade surda e, conseqüentemente, promove maior equidade nas interações sociais.

Conforme estabelece a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência), em seu artigo 1º:

É instituída a Lei Brasileira de Inclusão da Pessoa com Deficiência, destinada a assegurar e a promover, em condições de igualdade, o exercício dos direitos e das liberdades fundamentais por pessoa com deficiência, visando à sua inclusão social e cidadania. (BRASIL, 2015, p. 1).

Esse marco legal representa um instrumento significativo para a inclusão social das pessoas com deficiência, incluindo aquelas com deficiência auditiva, ao garantir o reconhecimento de sua cidadania plena. Os direitos e liberdades ali assegurados podem ser traduzidos como o acesso adequado a serviços de saúde, educação, transporte, cultura, lazer, além de outros serviços essenciais para a vida em sociedade. Nesse contexto, a difusão e valorização da língua de sinais revelam-se como fatores fundamentais tanto para a garantia do direito à comunicação quanto para o avanço de uma sociedade mais inclusiva, justa e democrática.

No campo da saúde, contudo, a efetivação desses direitos depende diretamente da possibilidade de comunicação clara e acessível entre profissionais e pacientes surdos, o que ainda representa um desafio significativo no contexto brasileiro. Nesse cenário, a escassez de profissionais da saúde e de recursos adequados para viabilizar a comunicação por meio da Libras permanece evidente. Um estudo conduzido com 240 estudantes de Medicina revelou que 82,9% deles não possuem habilidade comunicativa em Libras, embora 95,8% reconheçam a relevância da disciplina (OLIVEIRA et al., 2022). No que se refere à estrutura curricular, 55,2% dos participantes defenderam que a disciplina responsável por ensinar a Língua Brasileira de Sinais fosse obrigatória e, deste público, 75,7% demonstraram preferência pela modalidade presencial. Entre as principais justificativas para não cursar a disciplina, destacaram-se sua natureza optativa (41,7%) e a falta de tempo (33,3%).

Tais resultados evidenciam uma lacuna significativa na formação médica, a qual compromete a qualidade do atendimento e perpetua barreiras à inclusão da população surda. Esse cenário contraria o disposto pela Lei nº13.146/2015 (BRASIL, 2015) que assegura, em seu artigo 24, o acesso das pessoas com deficiência aos serviços de saúde, públicos e privados, mediante tecnologias assistivas e recursos de comunicação. Com isso,

observa-se que o baixo conhecimento em Libras pelos profissionais de saúde contribui para a manutenção das desigualdades vivenciadas pela comunidade surda no Brasil.

Diante desse cenário, tecnologias capazes de facilitar a comunicação entre profissionais de saúde e pacientes surdos tornam-se ainda mais relevantes, especialmente aquelas baseadas no reconhecimento automático de sinais. Para a tarefa de reconhecimento automático de sinais, é comum que os pesquisadores construam suas próprias bases de dados, já que ainda são escassos os repositórios públicos de Libras com variabilidade suficiente. Diversos trabalhos vêm seguindo essa abordagem (ROCHA et al., 2020; COSTA et al., 2017; REZENDE; ALMEIDA; GUIMARÃES, 2021), nos quais os autores não apenas organizam conjuntos de dados, mas também implementam modelos de aprendizado de máquina e redes neurais para a classificação dos sinais coletados.

Embora esses trabalhos representem avanços importantes, é preciso reconhecer algumas limitações recorrentes: a maioria das bases foi construída em ambientes controlados, com fundo uniforme, iluminação padronizada e número reduzido de participantes. Essas condições tendem a favorecer altos índices de acurácia nos experimentos, mas reduzem a variabilidade intra e intersujeitos. Essa limitação compromete a capacidade de generalização dos modelos quando aplicados em contextos reais e heterogêneos, como hospitais, escolas ou ambientes públicos.

Perante esse cenário, abordagens mais recentes têm explorado alternativas que visam superar tais limitações. A estimação de marcos corporais (mãos, pose e face), reduz a complexidade dos dados brutos ao selecionar apenas pontos anatômicos relevantes para a descrição do gesto. Esses marcos correspondem a coordenadas de pontos específicos do corpo, como articulações, extremidades dos dedos ou contornos do rosto, que capturam a forma e o movimento do corpo de forma compacta e significativa. Essa forma de representação contribui para lidar melhor com variações de ambiente e de sinalizadores.

Além disso, a representação por meio de marcos integra-se naturalmente a modelos de aprendizado sequencial, como as redes de memória de longo e curto prazo (*Long Short-Term Memory* – LSTM), que conseguem armazenar e recuperar dependências tanto de longo quanto de curto prazo, característica essencial para lidar com a natureza temporal e contínua dos sinais em Libras. Em diversos estudos (SAMAAN et al., 2022; GANDHI; SHAH; CHANDANE, 2022; HUANG; CHOUVATUT, 2024) essa arquitetura mostrou-se particularmente eficaz na interpretação de gestos dinâmicos, uma vez que o significado de um sinal não está em um único quadro isolado, mas no encadeamento de movimentos ao longo do tempo.

Diante da escassez de iniciativas voltadas à criação de bases de dados em Libras na área da saúde, este trabalho propõe um procedimento sistemático e estruturado para a coleta de sinais específicos desse domínio. A metodologia baseia-se no registro controlado de gestos executados por profissionais especializados em Libras, assegurando diversidade e

precisão nas amostras e permitindo a expansão gradual do conjunto de dados conforme novas necessidades surgirem.

O conjunto de dados proposto contempla 10 sinais dinâmicos de relevância imediata para a triagem hospitalar, correspondentes ao primeiro atendimento, onde a comunicação deve ser rápida e precisa.

Dessa forma, este trabalho propõe a criação de uma base de dados com sinais de Libras relacionados ao domínio da saúde, bem como o desenvolvimento e avaliação de modelos de classificação capazes de reconhecer automaticamente esses gestos, contribuindo para a redução de barreiras comunicacionais e para a inclusão da comunidade surda em ambientes clínicos.

1.1 DEFINIÇÃO DO PROBLEMA

O problema resolvido neste trabalho consiste em, dado um vídeo, determinar qual o sinal de Libras no contexto da saúde, dentre os 10 presentes no conjunto de dados, foi realizado. É interessante destacar que, de modo geral, vídeos podem apresentar variações de iluminação, posição e ângulo do indivíduo presente, plano de fundo e outras peculiaridades que podem tornar o processo mais desafiador. Desta forma, este trabalho utiliza representações baseadas em marcos corporais combinadas a redes neurais LSTM para desenvolver modelos capazes de mitigar esses problemas e reconhecer corretamente os sinais, independentemente do cenário ou do indivíduo que os realiza.

1.2 OBJETIVOS

Este trabalho tem como objetivo principal construir um conjunto de dados composto por vídeos de dez sinais de Libras no contexto da saúde, bem como desenvolver modelos de redes neurais capazes de realizar a classificação do sinal em execução. Desse modo, os objetivos específicos são:

- Estudar o funcionamento e o desenvolvimento de modelos de redes neurais aplicados à classificação de movimentos em vídeos;
- Realizar a gravação dos vídeos contendo sinais em Libras de forma controlada e padronizada, assegurando diversidade de gestos e participantes;
- Aplicar procedimentos de pré-processamento e organização dos vídeos, incluindo segmentação, rotulagem e estruturação do conjunto de dados para uso em modelos de aprendizado profundo;
- Implementar e treinar diferentes modelos de classificação baseados em redes neurais, ajustando seus parâmetros a partir de resultados experimentais;

- Avaliar o desempenho dos modelos segundo critérios previamente definidos, adotando estratégias de teste como a validação cruzada.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são abordados os conceitos fundamentais para o estudo e desenvolvimento deste trabalho. A Seção 2.1 apresenta alguns princípios da Libras. A Seção 2.2 detalha alguns conceitos sobre os marcos de mãos, pose e face. A Seção 2.3 apresenta os conceitos por trás do funcionamento de redes neurais e como ocorre o processo de aprendizagem. Por fim, a Seção 2.4 explora de maneira mais específica as redes neurais recorrentes, com destaque para a arquitetura LSTM.

2.1 LIBRAS

As línguas de sinais são línguas naturais de modalidade gesto-visual, estruturadas por parâmetros fonológicos como configuração e orientação das mãos, movimento, locação no espaço e aspectos não manuais. A Língua Brasileira de Sinais (Libras) compartilha essa organização linguística, apresentando complexidade comparável à das línguas orais-auditivas (QUADROS; KARNOPP, 2007).

No contexto específico das línguas de sinais, destaca-se ainda a distinção entre sinais estáticos e dinâmicos. Os sinais estáticos caracterizam-se por uma configuração única das mãos ou por uma postura específica do corpo ou do rosto, podendo ser reconhecidos a partir de uma única imagem. Em contrapartida, os sinais dinâmicos envolvem movimento e variação temporal, o que exige a análise de sequências de imagens, isto é, vídeos, para sua correta identificação. Nesse contexto, o desenvolvimento de soluções tecnológicas voltadas à tradução de línguas de sinais tem se mostrado fundamental para a ampliação da inclusão da comunidade surda em diferentes esferas sociais (GALA, 2023).

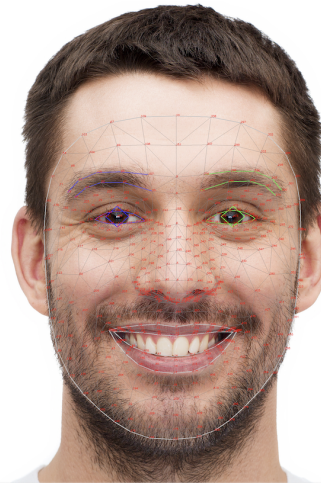
2.2 MARCOS CORPORAIS

A estimação de marcos corporais consiste na identificação de pontos-chave anatômicos que permitem representar geometricamente os movimentos do corpo humano ao longo do tempo. No reconhecimento de sinais, essa representação é especialmente relevante por descrever de forma compacta a dinâmica dos gestos manuais, expressões faciais e posturas corporais, reduzindo a influência de características visuais irrelevantes presentes nas imagens.

Ao abstrair a cena para um conjunto de coordenadas espaciais, o uso de marcos corporais favorece a modelagem do movimento e contribui para maior robustez frente a variações de iluminação, fundo e aparência dos indivíduos.

Os marcos corporais podem ser organizados em três grupos principais: face, mãos e pose corporal. A Figura 1 apresenta os pontos distribuídos ao longo da face do indivíduo, totalizando 468, os quais permitem capturar expressões faciais relevantes.

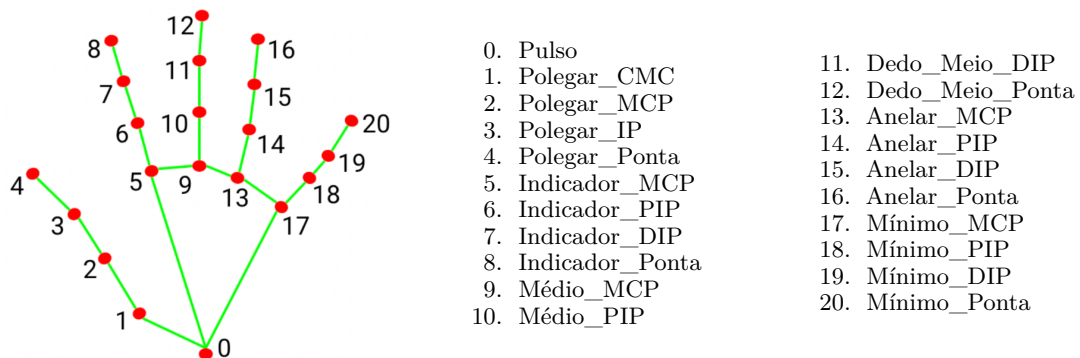
Figura 1 – Marcos faciais.



Fonte: MediaPipe Face (2025).

A Figura 2 mostra os 21 marcos referentes a uma das mãos, acompanhados do significado de cada ponto. Vale ressaltar que, caso sejam detectadas duas mãos na cena, o total de pontos estimados chega a 42.

Figura 2 – Marcos da mão. Os termos abreviados são articulações e seus nomes são: carpo-metacarpianas (CMC), metacarpolangeanas (MCP), interfalangeana (IP), interfalangeana proximal (PIP), interfalangeana distal (DIP).



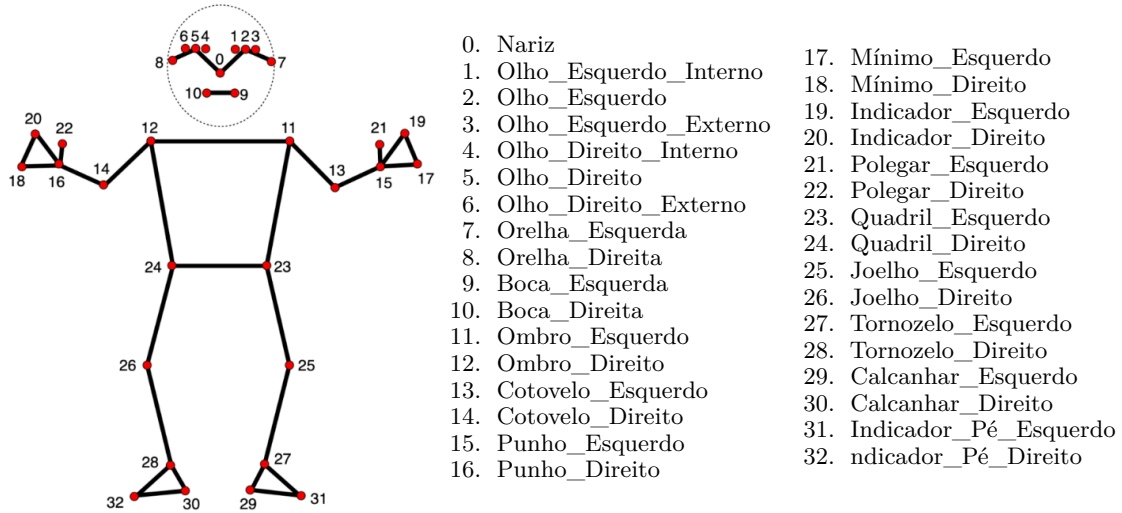
Fonte: Adaptada de MediaPipe Hands (2025).

Analisando a Figura 2, percebe-se a identificação de pontos específicos das articulações, o que torna possível analisar tanto a forma quanto o movimento dos dedos. Esse recurso é especialmente relevante no contexto do reconhecimento de línguas de sinais.

Por fim, a Figura 3 ilustra os marcos associados à pose corporal, totalizando 33 pontos de referência. Esses pontos abrangem tanto a região da face quanto membros superiores e inferiores. Essa representação fornece uma visão estrutural completa da postura

do indivíduo, permitindo a análise de movimentos corporais em diferentes contextos, como esportes ou reconhecimento de gestos.

Figura 3 – Marcos de pose.



Fonte: Adaptada de MediaPipe Pose (2025).

2.3 APRENDIZADO EM REDES NEURAIS

Rosenblatt (1958) implementou o primeiro modelo de rede neural chamado *perceptron*, o qual tinha o objetivo de reproduzir o funcionamento de um neurônio biológico. Diferentemente das redes neurais profundas, o *perceptron* consiste em um único neurônio, que realiza combinações lineares sobre os sinais de entrada, expressas pela seguinte equação:

$$z = \sum_{k=1}^n w_k x_k + w_{n+1}, \quad (2.1)$$

que, em notação vetorial, pode ser representada como:

$$z = \mathbf{w}^T \mathbf{x} + b. \quad (2.2)$$

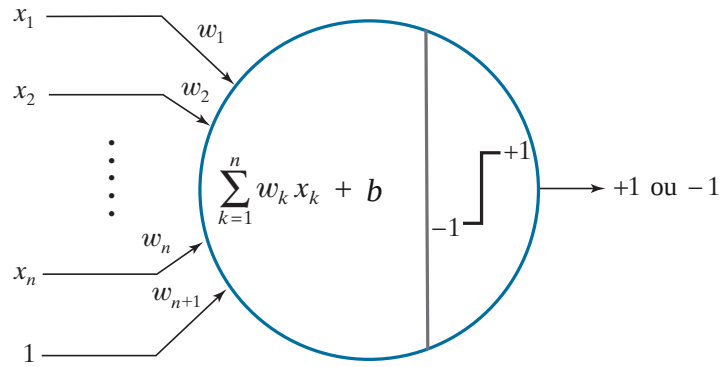
Neste contexto, o termo z é denominado potencial de ativação, $\mathbf{w}$ representa o vetor de pesos, $\mathbf{x}$ é o vetor de entradas, e b é o viés (*bias*), que difere dos pesos por ser multiplicado sempre pela constante 1.

Após o cálculo do potencial de ativação, o valor resultante é processado por uma função de ativação, gerando a saída do *perceptron*:

$$a = \begin{cases} +1, & \text{se } z > 0 \\ -1, & \text{se } z \leq 0 \end{cases}. \quad (2.3)$$

Dessa forma, o *perceptron* consegue classificar entradas em duas categorias, $+1$ ou -1 . A Figura 4 ilustra o funcionamento do *perceptron*, considerando um vetor de entrada $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$, um vetor de pesos $\mathbf{w} = [w_1, w_2, \dots, w_n]^T$, e b é o *bias*, que é somado separadamente e não incluído no vetor de entrada.

Figura 4 – Operações realizadas no *perceptron*.



Fonte: Adaptada de Gonzalez e Woods (2018).

Uma vez definida a estrutura do *perceptron*, o próximo passo é compreender como seus parâmetros são ajustados durante o treinamento. Esse processo consiste em ajustar os pesos e o *bias* de forma a minimizar a função de perda (*loss function*) que quantifica a diferença entre a saída desejada r e a saída prevista pelo modelo. Uma das maneiras mais comuns de medir essa diferença é o erro quadrático médio (*mean squared error* - MSE), definido como:

$$E(\mathbf{w}, b) = \frac{1}{2}(r - (\mathbf{w}^T \mathbf{x} + b))^2, \quad (2.4)$$

onde E representa a magnitude do erro, $\mathbf{w}$ é o vetor de pesos a ser aprendido e $\mathbf{x}$ é um padrão do conjunto de treinamento. Para encontrar o valor mínimo da Equação (2.4), utiliza-se o algoritmo de descida do gradiente iterativo, cuja equação de atualização é dada por:

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \alpha \left[\frac{\partial E(\mathbf{w}, b)}{\partial \mathbf{w}} \right]_{\mathbf{w}=\mathbf{w}(k)}, \quad (2.5)$$

onde $\alpha > 0$ representa a taxa de aprendizagem. A derivada da Equação (2.4) é dada por:

$$\frac{\partial E(\mathbf{w}, b)}{\partial \mathbf{w}} = -(r - (\mathbf{w}^T \mathbf{x} + b))\mathbf{x}. \quad (2.6)$$

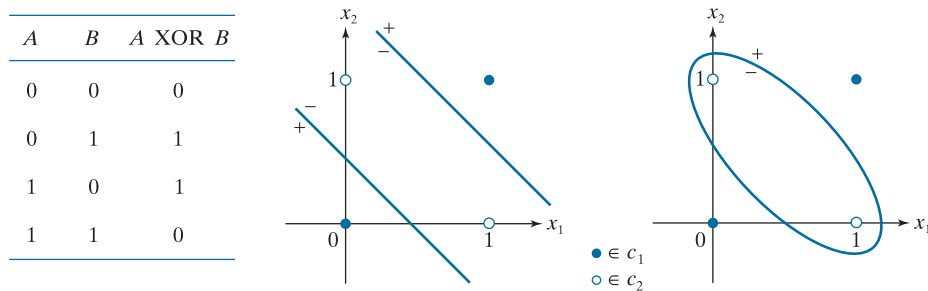
Substituindo a Equação (2.6) na Equação (2.5), obtém-se a regra de atualização dos pesos:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \alpha [r(k) - (\mathbf{w}^T(k)\mathbf{x}(k) + b(k))] \mathbf{x}(k), \quad (2.7)$$

na qual os termos são conhecidos ou obtidos com facilidade.

Porém, o *perceptron* está limitado a resolver problemas linearmente separáveis. Para ilustrar esta limitação, considere o problema da porta lógica XOR em que não existe uma única reta capaz de separar corretamente as saídas verdadeiras das falsas no espaço bidimensional de entrada. Conforme mostra a Figura 5, não é possível separar corretamente as saídas verdadeiras das falsas com uma única reta, evidenciando a limitação do *perceptron*.

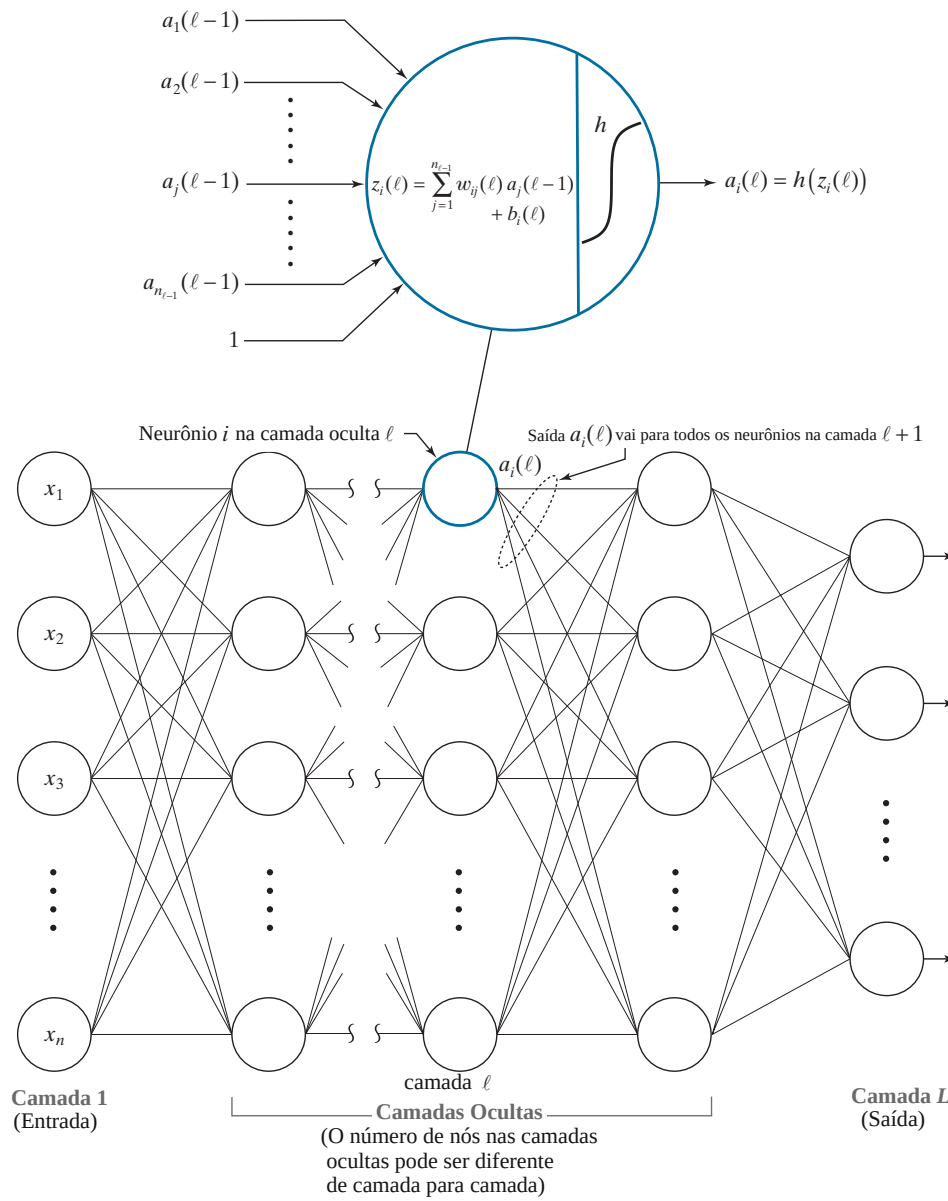
Figura 5 – Problema de classificação XOR em 2D.



Fonte: Gonzalez e Woods (2018).

Esta constatação evidencia a necessidade de arquiteturas mais elaboradas. Nesse contexto, o *perceptron multicamadas* (*multilayer perceptron* – MLP) surgiu, incorporando camadas intermediárias de neurônios, denominadas camadas ocultas, entre a entrada e a saída da rede. A Figura 6 ilustra a estrutura típica de uma rede MLP, caracterizada por múltiplas camadas totalmente conectadas, onde todos os neurônios de uma camada se conectam a todos os neurônios da camada seguinte.

Figura 6 – Representação da estrutura de uma rede MLP.



Fonte: Adaptada de Gonzalez e Woods (2018).

No MLP, cada neurônio i da camada l realiza o seguinte cálculo:

$$z_i(l) = \sum_{j=1}^{n_{l-1}} w_{ij}(l) a_j(l-1) + b_i(l) , \quad (2.8)$$

onde $w_{ij}(l)$ representa o peso da conexão, $a_j(l-1)$, a saída do neurônio j da camada anterior e $b_i(l)$, o *bias*. O resultado é então processado por uma função de ativação:

$$a_i(l) = h(z_i(l)) , \quad (2.9)$$

produzindo a saída do neurônio.

Diferentemente do *perceptron*, o MLP utiliza funções de ativação não-lineares, como a sigmoide, permitindo que a rede aprenda representações não-lineares dos dados. Arquiteturas mais modernas costumam utilizar funções como a *ReLU* (*Rectified Linear Unit*), definida por $h(z) = \max(0, z)$ que acelera o treinamento e melhora o desempenho em redes profundas.

O processo de treinamento da rede é realizado por meio do algoritmo de retropropagação (*backpropagation*), que ajusta os pesos e *biases* de modo a minimizar a função de perda, utilizando o método de descida do gradiente. Esse processo é dividido em duas etapas principais: passagem para frente (*forward pass*) e passagem para trás (*backward pass*). Na primeira, os dados são propagados da entrada até a saída através das Equações (2.8) e (2.9). Já na segunda, o erro calculado na saída é retropropagado, ajustando os pesos com base no método de descida do gradiente:

$$w_{ij}(l) = w_{ij}(l) - \alpha \frac{\partial E}{\partial w_{ij}(l)} , \quad (2.10)$$

onde α é a taxa de aprendizagem. A derivada parcial é calculada utilizando a regra da cadeia, permitindo que o erro se propague das camadas finais para as iniciais. A atualização dos *biases* segue princípio similar:

$$b_i(l) = b_i(l) - \alpha \frac{\partial E}{\partial b_i(l)} . \quad (2.11)$$

Este processo iterativo permite que o MLP ajuste seus parâmetros gradualmente, aprendendo representações hierárquicas não lineares dos dados de entrada.

2.4 REDES NEURAIS RECORRENTES

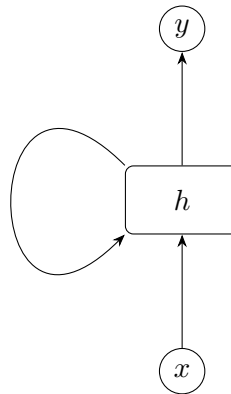
Redes neurais recorrentes (*recurrent neural networks* – RNNs) constituem uma arquitetura de aprendizado profundo voltada para o processamento de informações sequenciais. Diferentemente das redes neurais tradicionais, como MLP, que tratam cada entrada de forma independente, as RNNs mantêm conexões recorrentes que permitem a reutilização de estados anteriores durante o processamento de novos dados (MIENYE; SWART; OBAIDO, 2024). Esse mecanismo confere à rede a capacidade de modelar dependências temporais e contextuais, característica essencial em aplicações como análise de linguagem natural, reconhecimento de voz e previsão de séries temporais (MIENYE; SWART; OBAIDO, 2024).

Em essência, o diferencial das RNNs está em sua habilidade de incorporar a noção de ordem e continuidade na representação dos dados, aproximando o processamento computacional de fenômenos que evoluem no tempo. Cada neurônio recorrente mantém um estado oculto (h_t), que sintetiza a informação recebida até o instante t . Esse estado é

atualizado a cada nova entrada, combinando a entrada atual (x_t) com o estado anterior (h_{t-1}), permitindo que a rede “lembre” informações de passos temporais anteriores.

A Figura 7 apresenta uma representação simplificada de uma RNN. Nela, é possível observar o fluxo de informação entre a entrada x , o estado oculto h e a saída y . A seta curva retornando para h ilustra a recorrência, ou seja, a forma como o estado oculto de um passo de tempo influencia o próximo.

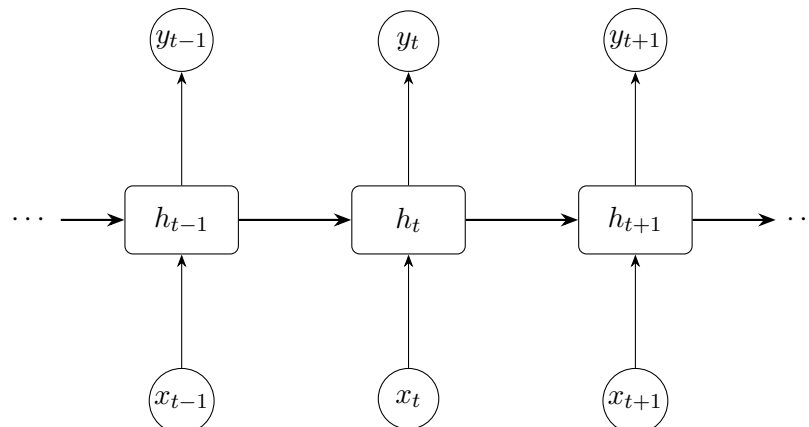
Figura 7 – Representação esquemática de uma rede neural recorrente (RNN).



Fonte: Adaptada de Feng et al. (2017).

Para visualizar o comportamento temporal da RNN de forma mais explícita, podemos expandir a rede em etapas de tempo, como mostrado na Figura 8. Essa forma expandida demonstra o compartilhamento de parâmetros entre camadas ao longo do tempo: cada bloco h_t recebe a entrada atual x_t e o estado anterior h_{t-1} , gerando a saída y_t . Essa estrutura evidencia a capacidade da RNN de propagar informações ao longo da sequência.

Figura 8 – Expansão temporal de uma RNN evidenciando o compartilhamento de parâmetros.



Fonte: Adaptada de Feng et al. (2017).

Com relação ao treinamento desse tipo de rede neural, ele segue o mesmo princípio

explicado anteriormente, ou seja, busca-se minimizar uma função de perda por meio do método de descida do gradiente e do processo de retropropagação do erro. Entretanto, devido à natureza recorrente dessas redes, a propagação dos gradientes ao longo de muitas etapas de tempo pode causar a chamada dissipação (ou desaparecimento) do gradiente, fenômeno no qual os valores dos gradientes tornam-se extremamente pequenos à medida que retrocedem pelas camadas temporais. Como consequência, a atualização dos pesos se torna ineficaz, comprometendo o aprendizado da rede. Para ser possível contornar esse problema, novas arquiteturas foram desenvolvidas e uma de destaque é, justamente, a memória de curto e longo prazo (*long short-term memory* – LSTM).

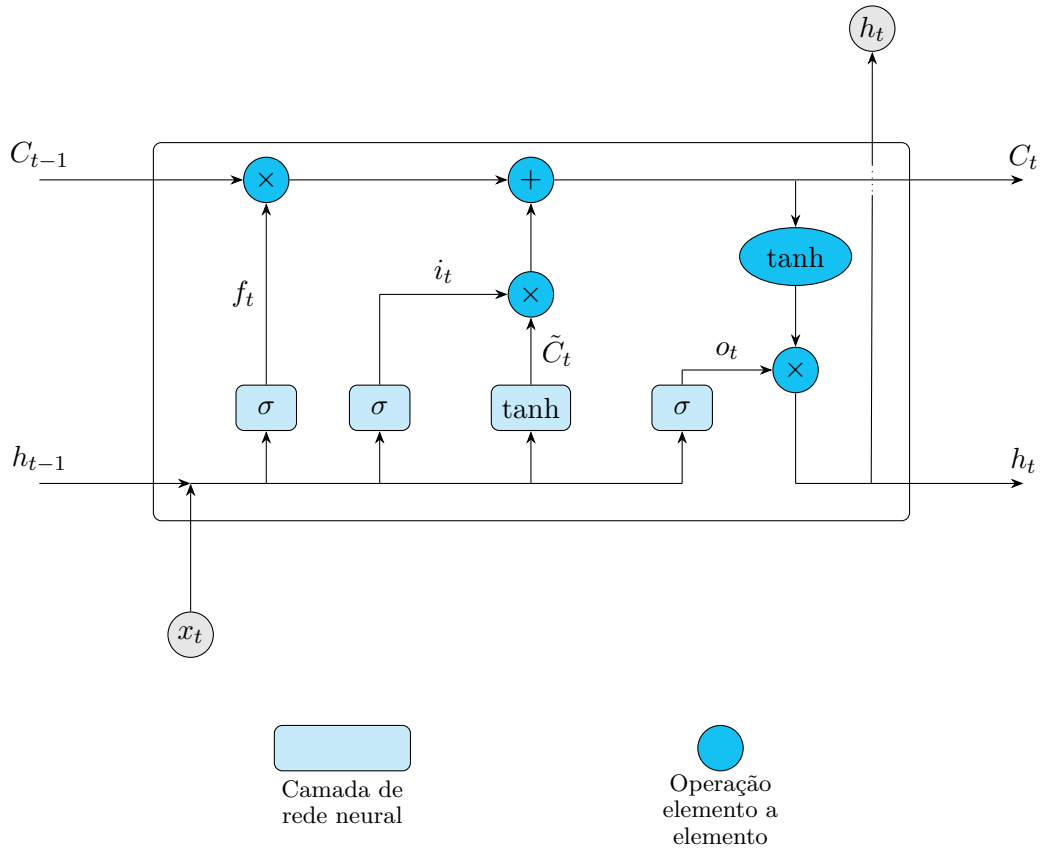
Proposta por Hochreiter e Schmidhuber (1997), a LSTM é capaz de aprender dependências de curto e longo prazo, como o próprio nome sugere. Isso é feito por meio da utilização de unidades de memória especializadas, conhecidas como células de memória (*memory cells*).

Essas células são composta por um estado interno (C_t) e por um conjunto de portões (*gates*) que controlam o fluxo de informações que entram e saem da célula. No modelo original, a LSTM possuía duas portas principais: o portão de entrada e o portão de saída, que determinavam, respectivamente, quando novas informações deveriam ser armazenadas e quando o conteúdo da célula deveria influenciar a saída da rede (HOCHREITER; SCHMIDHUBER, 1997).

Posteriormente, Gers, Schmidhuber e Cummins (1999) ampliaram o modelo original ao introduzir o portão de esquecimento (*forget gate*), responsável por determinar quanto da memória anterior deve ser mantida ou esquecida. Essa modificação tornou a arquitetura mais flexível e estável, permitindo que a rede aprenda de forma autônoma a “lembrar” ou “esquecer” informações conforme a necessidade da tarefa.

A Figura 9 demonstra a arquitetura de uma célula LSTM. Por meio dela é possível notar o compartilhamento de informações entre as células como o recebimento de h_{t-1} e C_{t-1} , provenientes da célula anterior, bem como a distribuição de h_t e C_t para a célula posterior. Em particular, a estrutura dos portões, as equações matemáticas e a interpretação de suas etapas foram fundamentadas na apresentação intuitiva e detalhada de Olah (2015).

Figura 9 – Diagrama de uma célula LSTM.



Fonte: Adaptada de Olah (2015).

No funcionamento de uma célula LSTM, ocorre, em primeiro lugar, a decisão de quais informações deverão ser mantidas ou descartadas. Isso ocorre no que foi definido como portão de esquecimento, composto por uma camada sigmoide. Essa camada analisa os valores de saída do instante anterior (h_{t-1}) e a entrada atual (x_t), produzindo um conjunto de números entre 0 e 1 para cada elemento do estado de célula anterior (C_{t-1}), conforme expresso por:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) , \quad (2.12)$$

em que valores próximos de 1 indicam que a informação deve ser mantida, enquanto valores próximos de 0 determinam que ela deve ser descartada.

Após a etapa de esquecimento, o próximo passo consiste em definir quais novas informações serão armazenadas no estado da célula. Esse processo ocorre em duas partes. Primeiramente, uma camada sigmoide, denominada portão de entrada, determina quais componentes do estado da célula serão atualizados:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) . \quad (2.13)$$

Em seguida, uma camada com função de ativação tangente hiperbólica gera um vetor de valores candidatos $\tilde{C}_t$, que representam novas informações a serem potencialmente incorporadas ao estado da célula, conforme expresso por:

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) . \quad (2.14)$$

Com essas informações, o estado da célula é atualizado a partir da combinação entre o conteúdo previamente armazenado e os novos valores candidatos. O estado anterior C_{t-1} é filtrado pelo vetor de esquecimento f_t , descartando as informações consideradas irrelevantes, enquanto o resultado do portão de entrada i_t é multiplicado pelos valores candidatos $\tilde{C}_t$. A soma dessas duas contribuições resulta no novo estado da célula:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t . \quad (2.15)$$

Por fim, a rede precisa decidir qual será a saída no instante de tempo atual. Essa etapa é realizada pelo portão de saída. Inicialmente, uma camada sigmoide define quais componentes do estado da célula serão enviados como saída:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) . \quad (2.16)$$

Em seguida, o estado da célula C_t é submetido à função tangente hiperbólica e o resultado é combinado com o vetor do portão de saída, produzindo o novo estado oculto h_t :

$$h_t = o_t \odot \tanh(C_t) . \quad (2.17)$$

Nas equações apresentadas, os termos W_f , W_i , W_c e W_o representam as matrizes de pesos associadas a cada uma das etapas que ocorrem na célula LSTM (esquecimento, entrada, candidato e saída, respectivamente), enquanto b_f , b_i , b_c e b_o correspondem aos vetores de *bias* adicionados em cada operação. Por fim, a multiplicação elemento a elemento (ou produto de *Hadamard*), denotada por $\odot$, combina as ativações dos portões com os respectivos vetores de estado.

Com isso, essa estrutura permite que a LSTM mantenha e atualize informações relevantes ao longo do tempo, mitigando o problema do desaparecimento do gradiente comumente observado em redes recorrentes tradicionais, conforme comentado anteriormente. E, dessa forma, são eficazes para tarefas que envolvam sequências longas, como a que é abordada neste trabalho, a classificação de sinais de Libras.

3 TRABALHOS RELACIONADOS

Neste capítulo são detalhados alguns trabalhos utilizados como referências para os tópicos que foram abordados neste trabalho: desenvolvimento de um conjunto de dados com vídeos de execuções de sinais de Libras e o reconhecimento de linguagens de sinais.

3.1 CONJUNTOS DE DADOS

Diversas bases de dados são utilizadas na área de reconhecimento de língua de sinais. A utilização das mesmas é de fundamental importância para que sistemas computacionais, principalmente de aprendizado de máquina (*machine learning*) ou aprendizado profundo (*deep learning*), sejam treinados e testados.

Dentre os conjuntos existentes, LSA64 (RONCHETTI et al., 2023) se destaca por abordar sinais dinâmicos. É composto por 64 sinais registrados em 3200 vídeos, capturados por 10 indivíduos, os quais realizaram 5 vídeos cada, voltados para a Língua de Sinais Argentina. Ademais, é interessante destacar que, nesse trabalho, foram utilizadas luvas fluorescentes de modo a facilitar as etapas de rastreamento e segmentação manuais, sem o uso de sensores. Além disso, os indivíduos utilizaram roupas pretas e os vídeos foram gravados em um fundo branco.

Outro trabalho de destaque é o *American Sign Language Lexicon Video Dataset* (ASLLVD) desenvolvido por Neidle, Thangali e Sclaroff (2012). Sua importância vem do fato de, atualmente, ser o maior em número de classes disponível, apresentando mais de 3300 sinais da língua de sinais americana. No entanto, apresenta algumas limitações, como apenas 6 indivíduos para a realização dos sinais e uma média de apenas 3 vídeos por sinal.

Considerando o contexto brasileiro, destacam-se dois conjuntos de dados: o MINDS-Libras *dataset*, desenvolvido por Rezende, Almeida e GUIMARÃES (2021), e o LIBRAS-UFOP, proposto por Cerna et al. (2021). O MINDS-Libras *dataset* reúne 20 sinais em Libras, selecionados por um especialista com base em critérios como configuração das mãos, localização, movimento, orientação da palma e expressões faciais. Ao todo, 12 participantes realizaram cada sinal cinco vezes, resultando em 1.158 sequências de vídeo (um pouco abaixo das 1.200 esperadas devido à ausência de alguns registros). Essa base é relevante por priorizar diversidade de gestos e incluir participantes distintos.

Já o conjunto de dados LIBRAS-UFOP contempla 56 sinais diferentes, executados por 5 pessoas. Os autores organizaram os sinais em quatro grupos, levando em consideração semelhanças nos pontos de articulação, configurações de mão, movimentos e expressões faciais. Um diferencial dessa base é a presença de arquivos auxiliares que delimitam com precisão os quadros de início e fim de cada execução, o que possibilita uma segmentação mais rigorosa dos sinais nos experimentos. Apesar de contar com um número reduzido de participantes, trata-se de um conjunto mais amplo em termos de vocabulário e com

anotações detalhadas que favorecem pesquisas avançadas em reconhecimento automático.

3.2 RECONHECIMENTO DE LÍNGUAS DE SINAIS

Nos últimos anos, o avanço do aprendizado profundo impulsionou significativamente as pesquisas em reconhecimento de línguas de sinais, especialmente em vídeos. Entre as abordagens mais promissoras estão a extração de marcos corporais — mãos, rosto e pose — e a utilização de arquiteturas de modelagem temporal, como as redes LSTM.

Huang e Chouvatut (2024) propuseram um *pipeline* baseado em vídeo que combina uma rede ResNet para extração de características visuais e uma camada LSTM para capturar dependências temporais, obtendo resultados competitivos ao avaliar o método no conjunto de dados LSA64. Embora eficiente, esse estudo ainda depende fortemente de características globais de imagem, sem explorar explicitamente a utilização de marcos corporais.

Gandhi, Shah e Chandane (2022) integraram a biblioteca *MediaPipe*¹ para a extração automática de marcos de mãos, face e pose. Esses vetores de coordenadas foram utilizados exclusivamente para a tarefa de reconhecimento de sete sinais da Língua Americana de Sinais, processados por uma rede LSTM. Além disso, o estudo também avaliou a detecção de emoções por meio de um módulo independente baseado em imagens, não em marcos. Os autores destacam a robustez da representação baseada em marcos corporais frente a variações de iluminação e plano de fundo, alcançando acurácia de 98,95% na classificação dos sinais.

Samaan et al. (2022) expandiram essa linha de pesquisa ao comparar diferentes arquiteturas recorrentes (RNN-GRU, LSTM e Bi-LSTM) utilizando o conjunto de dados DSL10, composto por 10 sinais comuns do cotidiano, como saudações e expressões básicas. Os resultados mostraram que, embora as GRUs sejam mais leves em termos computacionais, as LSTMs (especialmente as Bi-LSTMs) alcançam melhor desempenho em cenários de maior complexidade temporal, quando combinadas aos 1662 marcos de mãos, face e pose extraídos pelo *MediaPipe*.

Em um contexto mais aplicado ao cenário brasileiro, Sarmiento (2023) apresentou uma dissertação voltada à classificação de vídeos em Libras. Embora tenha empregado convoluções para a extração de características visuais em imagens, a autora enfatiza a importância de um estágio sequencial baseado em LSTM para o reconhecimento de gestos. Os resultados indicaram que o melhor desempenho foi obtido quando se utilizaram exclusivamente os marcos corporais combinados a redes LSTM, superando as abordagens que integravam imagens com LSTM ou imagens, marcos corporais e LSTM de forma conjunta.

¹ <<https://ai.google.dev/edge/mediapipe/solutions/guide?hl=pt-br>>.

Outros estudos, como o de Al-Mohimeed et al. (2021), empregaram redes neurais convolucionais (CNNs) para a classificação de sinais em tempo real, porém, sem explorar explicitamente o uso de marcos corporais na modelagem temporal. Em contraste, técnicas que combinam *MediaPipe* e LSTM têm se mostrado mais robustas frente à variabilidade de usuários e ambientes.

Além disso, Ma et al. (2021) propuseram um método de reconhecimento dinâmico de língua de sinais baseado em uma rede Residual-LSTM melhorada. Os autores incorporaram mecanismos de aprendizagem residual e atenção por canal em uma rede convolucional, aprimorando a extração de características espaciais das imagens, e utilizaram uma rede LSTM para capturar as características temporais das sequências de imagens. Para a detecção das mãos, foi empregada a rede YOLO (*You Only Look Once*) (REDMON et al., 2016), diminuindo ainda mais a carga computacional. O método foi avaliado no *SLR dataset* (ZHANG et al., 2016) e alcançou acurácia de 91,2%, demonstrando eficácia na classificação de sinais dinâmicos.

De forma complementar, Ridwang et al. (2023) propuseram uma abordagem baseada na biblioteca *MediaPipe*, combinada com uma versão modificada de LSTM, para o reconhecimento de sinais a partir de movimentos dinâmicos das mãos, da pose e da face. No estudo, 20 movimentos dinâmicos foram processados e o modelo alcançou acurácia média de 99,4% no treinamento, com 85% de acurácia na detecção de palavras isoladas e 80% em sentenças completas, indicando avanços significativos na tradução de sinais dinâmicos.

Em síntese, a utilização de marcos corporais para descrever poses e gestos, combinada ao emprego de redes LSTM, configura-se como uma das abordagens mais promissoras para o reconhecimento dinâmico de línguas de sinais, pois explora simultaneamente a geometria do movimento e as dependências temporais intrínsecas às sequências de vídeo. Diante disso, o presente trabalho adota uma abordagem semelhante, explorando representações baseadas em marcos corporais em conjunto com modelos LSTM para o reconhecimento de sinais.

4 CONJUNTO DE DADOS

No contexto deste trabalho, não foram encontrados conjuntos de dados que contemplassem sinais da Língua Brasileira de Sinais (Libras) voltados a aplicações em contextos específicos, como o hospitalar. Dessa forma, optou-se pela criação de um conjunto de dados próprio, desenvolvido com o objetivo de atender às necessidades do trabalho e representar um cenário mais realista de uso.

O desenvolvimento desse conjunto de dados teve início na disciplina Trabalho Multidisciplinar, ofertada pelo Departamento de Mecânica Aplicada e Computacional da Universidade Federal de Juiz de Fora (UFJF). A partir do tema proposto na disciplina, a construção do conjunto de dados foi realizada em colaboração com outros dois discentes.

4.1 DESCRIÇÃO GERAL

O primeiro passo para a construção do conjunto de dados consistiu na definição do escopo temático dos sinais a serem incluídos. Assim, o conjunto de dados elaborado contempla sinais comumente utilizados durante o atendimento em ambientes hospitalares, visando facilitar a comunicação entre profissionais de saúde e pacientes surdos. Todos os sinais selecionados são de natureza dinâmica, o que exige maior precisão em sua execução e captura.

Para a gravação, foram convidados profissionais fluentes em Libras, com diversidade de sexo, cor de pele, altura e vestimentas, de modo a aumentar a variabilidade e reduzir possíveis vieses no conjunto de dados. Esses profissionais fazem parte do Grupo de Estudos Linguísticos da Libras (GELLI), vinculado à Faculdade de Letras da UFJF, o que garantiu a qualidade linguística das execuções e a adequação dos sinais selecionados.

As gravações foram realizadas tendo como plano de fundo uma lona de cor sólida verde, o que proporcionou uma padronização do ambiente e facilitou o processo de segmentação do corpo e das mãos dos participantes. Essa escolha também permite, caso necessário, a aplicação da técnica de *chroma key* para substituição do fundo, possibilitando o uso das gravações em diferentes contextos ou etapas de processamento.

Ao todo, o conjunto de dados contém 10 sinais distintos, realizados por 9 indivíduos, cada um repetindo os gestos em até cinco vezes, totalizando 492 vídeos. Os sinais selecionados foram: “Alergia”, “Autismo”, “Bronquite”, “Câimbra”, “Convulsão”, “Diarreia”, “Fratura”, “Frieira”, “Infarto” e “Sarampo”. Cada sinal é caracterizado por uma sequência de movimentos que envolvem mãos, braços e postura corporal, evidenciando a natureza dinâmica dos gestos e a necessidade de um processamento computacional capaz de capturar a complexidade temporal dessas ações.

A escolha dos 10 sinais incluídos no conjunto de dados foi guiada pela relevância

clínica e pela frequência com que essas condições ou sintomas são encontrados em ambientes hospitalares. Dessa forma, priorizou-se sinais que representam doenças ou manifestações de saúde comuns, como “Alergia”, “Bronquite” e “Diarreia”, bem como situações de urgência médica, como “Convulsão” e “Infarto”. Essa seleção visa garantir que o conjunto de dados seja útil para a comunicação prática entre profissionais de saúde e pacientes surdos, atendendo a necessidades reais de atendimento e triagem hospitalar.

A variação na quantidade de amostras dos sinais se deve ao fato de alguns participantes terem gravado todos os sinais mais de uma vez, enquanto outros realizaram apenas parte deles. Um indivíduo em particular foi gravado em diferentes condições — trocando vestimentas, acessórios (como óculos) ou penteado (cabelo preso ou solto) —, o que contribuiu para a diversidade do material coletado. A Tabela 1 apresenta a distribuição do número de vídeos por sinal.

Tabela 1 – Distribuição do número de vídeos por sinal no conjunto de dados de Libras construído neste trabalho.

Sinal	Quantidade de vídeos
Alergia	50
Autismo	50
Bronquite	48
Câimbra	48
Convulsão	50
Diarreia	49
Fratura	50
Frieira	50
Infarto	50
Sarampo	47
Total	492

Fonte: Elaborada pelo autor (2026).

Com os sinais definidos e os participantes selecionados, deu-se início ao processo de gravação dos vídeos. Essa etapa foi conduzida de forma padronizada, com controle sobre as condições de iluminação, posicionamento da câmera e instruções aos participantes, conforme descrito na próxima seção.

4.2 AQUISIÇÃO DOS VÍDEOS

O conjunto de dados foi gravado utilizando uma câmera *Canon EOS Rebel T6i*, configurada para capturar vídeos na resolução de 1280×720 *pixels* (HD), em alguns casos, ou de 640×480 em outros, com taxa de 30 quadros por segundo (*frames per second* – FPS). Para garantir a estabilidade da câmera e evitar vibrações, foi utilizado um tripé fixo.

As gravações foram realizadas em um ambiente interno com controle de iluminação, de forma a eliminar interferências provenientes de fontes de luz externa. Conforme comentado anteriormente, o fundo consistiu em uma lona verde sólida. A câmera foi posicionada a aproximadamente 1,5 metros de distância dos participantes, de modo que o enquadramento mostrasse cerca de um palmo acima da cabeça e um palmo abaixo da região do umbigo, conforme ilustrado na Figura 10. Essa configuração assegurou que todos os movimentos dos braços e mãos ficassem completamente visíveis.

Figura 10 – Exemplo de enquadramento respeitando as configurações definidas.



Fonte: Elaborada pelo autor (2026).

Antes de cada gravação, o participante era instruído a adotar uma posição inicial padronizada, com as mãos unidas à frente do abdômen. Em seguida, era feita a confirmação da correta compreensão do sinal a ser executado.

Cada vídeo contém a execução de um sinal específico, realizado cinco vezes consecutivas (com exceção de alguns casos) sem interrupção da gravação. Após cada repetição, o participante retornava à posição inicial antes de iniciar novamente o gesto. Esse procedimento garantiu a uniformidade entre as amostras, além de permitir que cada vídeo pudesse ser posteriormente segmentado, para isolar cada execução individual.

Dessa maneira, as execuções de cada sinal presentes nos vídeos foram segmentadas. Em outras palavras, cada vídeo contendo múltiplas repetições consecutivas de um mesmo sinal foi dividido em vários vídeos menores, de modo que cada arquivo resultante passasse a conter apenas uma execução individual.

Esse processo foi realizado utilizando a ferramenta *Online Video Cutter*¹, que permitiu realizar o corte dos vídeos de maneira simples e intuitiva.

4.3 CONJUNTO DE DADOS DE TESTE ADICIONAL

Além do conjunto de dados desenvolvido com profissionais de Libras integrantes do GELLI, foram gravados vídeos adicionais com uma pessoa externa a esse grupo, mas que também possui domínio na Língua Brasileira de Sinais. Ao todo, foram gravados 60 vídeos, sendo 6 vídeos correspondentes para cada sinal.

Diferentemente do processo de gravação descrito anteriormente, nesse caso o ambiente não era controlado, tanto no que diz respeito à padronização do fundo quanto às condições de iluminação. O equipamento utilizado também foi distinto sendo ele um *smartphone* Samsung modelo S20 FE, na resolução de 1080×1920 e taxa de 60 FPS.

No entanto, a dinâmica das gravações, bem como o processo de segmentação das execuções individuais, seguiu exatamente os mesmos procedimentos descritos anteriormente.

Assim, esse conjunto de dados será utilizado para avaliar os modelos desenvolvidos a partir do conjunto de dados principal.

¹ <<https://online-video-cutter.com/pt/>>

5 MÉTODO PROPOSTO

Este capítulo descreve as etapas de pré-processamento aplicadas ao conjunto de dados desenvolvido, abordando desde a preparação inicial dos vídeos até os procedimentos de extração e organização dos marcos corporais. Além disso, apresenta a estrutura da rede neural utilizada para a tarefa de classificação dos sinais, detalhando os componentes que compõem o modelo.

5.1 PRÉ-PROCESSAMENTO DOS DADOS

Após a aquisição dos vídeos, foi realizada a etapa de pré-processamento, com o objetivo de preparar o conjunto de dados para utilização nos experimentos. Essa fase é essencial para garantir a preparação das amostras para a entrada de modelos de aprendizado de máquina.

5.1.1 Seleção de quadros

Mesmo com a utilização dos mesmos equipamentos e a padronização das condições de gravação, os participantes não executam os sinais exatamente no mesmo intervalo de tempo. Considerando esse aspecto e o fato de que, em sua maioria, as redes neurais requerem entradas de tamanho fixo, optou-se por selecionar 30 quadros de cada vídeo. Essa quantidade foi definida de modo a preservar a essência e as principais transições do movimento, garantindo representatividade do sinal.

Assim, é necessário um método para realizar a seleção desses quadros. De modo semelhante a Sarmento (2023), no presente trabalho, foram realizados experimentos utilizando a seleção de quadros baseada em uma distribuição uniforme e em uma distribuição normal. Pelos resultados obtidos, percebeu-se que o melhor critério de escolha é uma amostragem com distribuição normal, que possui a seguinte função de densidade de probabilidade:

$$f(x) = \frac{1}{\sigma(\sqrt{2\pi})} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \quad (5.1)$$

onde μ é a média e σ é o desvio padrão.

A distribuição normal funciona melhor pelo fato de que os movimentos principais tendem a ocorrer na região central do vídeo, uma vez que tanto no início quanto no fim os participantes permanecem em posição neutra. Desse modo, utilizando esse critério de seleção, as amostras escolhidas concentram-se aproximadamente na metade da duração de cada vídeo. No entanto, é importante destacar que esse tipo de amostragem pode enfrentar certo tipo de limitação quando não existe uma padronização da execução do movimento.

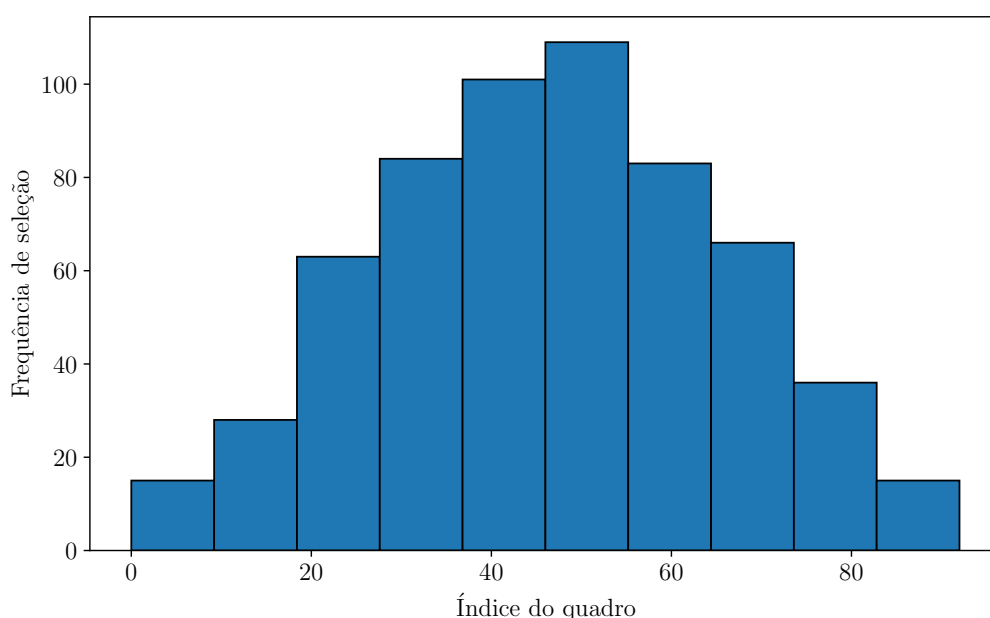
Isso porque, ela considera que a parte mais relevante do movimento se encontra próximo ao quadro central da execução, o que pode não ocorrer em casos não padronizados.

O valor de μ foi definido como a metade do total de quadros, enquanto σ foi definido como $\mu \times 0,4$, o que equivale a 20% do comprimento total do vídeo. Dessa forma, o intervalo $\mu \pm \sigma$ cobre cerca de 40% da região central, onde normalmente ocorre a maior parte do movimento. Como a densidade da distribuição normal é máxima na média e decresce gradualmente com o afastamento, essa configuração aumenta a probabilidade de selecionar quadros centrais, sem excluir completamente quadros mais distantes.

Utilizando um dos vídeos do conjunto de dados como exemplo, foram realizados 20 sorteios considerando um total de 95 quadros. Nesse caso, os índices variam de 0 a 94, o que resulta em $\mu = 47$ e $\sigma = 18,8$. Assim, os quadros cujos índices estão no intervalo entre 28 e 66 (isto é, $\mu \pm \sigma$) possuem maior probabilidade de serem selecionados, enquanto os demais quadros, mais distantes da média, apresentam probabilidade significativamente menor. Esse comportamento é consequência direta da forma da distribuição normal utilizada na etapa de amostragem.

Dessa forma, a Figura 11 apresenta o histograma dos índices dos quadros amostrados, sendo possível observar que a maior frequência de seleções se concentra em torno da média.

Figura 11 – Histograma dos índices dos quadros amostrados utilizando distribuição normal.



Fonte: Elaborada pelo autor (2026).

Porém, como pode-se perceber, esse tipo de amostragem envolve certa aleatoriedade. Portanto, em cada seleção, os índices dos quadros nem sempre serão exatamente os mesmos.

5.1.2 Aumento de Dados

Um dos principais desafios no treinamento de modelos de aprendizado profundo é o sobreajuste (*overfitting*), que ocorre quando o modelo aprende detalhes e ruídos específicos do conjunto de treinamento, ajustando-se excessivamente a ele. Como resultado, o modelo apresenta bom desempenho durante o treinamento, mas falha ao lidar com dados novos, devido à sua baixa capacidade de generalização. Esse fenômeno pode ocorrer quando o conjunto de treinamento é limitado em variabilidade ou quando o modelo é treinado por tempo excessivo ou possui complexidade desnecessariamente alta (IBM, 2024).

Para mitigar esse problema, foi aplicado o aumento de dados (*data augmentation*), técnica que consiste em expandir artificialmente o conjunto de amostras por meio de modificações controladas nos dados originais, de modo a introduzir variações que favorecem a robustez do modelo.

No presente trabalho, foram realizadas combinações das seguintes operações de transformação geométrica, aplicadas a cada quadro dos vídeos, juntamente com o redimensionamento para a resolução de 640×480 *pixels*:

- Rotação: variação aleatória no intervalo de $[-10^\circ, 10^\circ]$;
- Translação horizontal e vertical: deslocamentos aleatórios de até 20 *pixels* nos eixos x e y ;
- Recorte (*crop*): remoção de até 10% da área total da imagem a partir do centro, mantendo o foco na região central.

Um aspecto importante a ser ressaltado é que, para cada vídeo, todas as transformações foram aplicadas de forma consistente a todos os seus quadros, garantindo a coerência temporal da sequência.

Além disso, essas operações foram integradas ao processo de seleção de quadros baseado na distribuição normal apresentada anteriormente. Dessa forma, para cada sorteio realizado, os quadros selecionados foram submetidos simultaneamente às transformações e ao redimensionamento. Esse procedimento foi repetido 20 vezes para cada vídeo, resultando em novas variações do mesmo gesto e contribuindo para um conjunto de dados mais diverso e representativo.

Por fim, é importante destacar que duas operações comuns em aumento de dados não foram aplicadas neste trabalho: o espelhamento horizontal e transformações relacionadas à cor. O espelhamento horizontal, embora amplamente utilizado em tarefas de visão computacional, foi evitado por não garantir a preservação da dinâmica e da validade linguística dos sinais em Libras. Em línguas de sinais, a lateralidade, a mão dominante e a posição espacial relativa dos gestos desempenham papel fundamental na construção do

significado, de modo que a simples inversão espacial pode resultar em sinais semanticamente distintos ou em execuções que não ocorrem naturalmente. Além disso, a ausência de participantes com diferentes mãos dominantes no conjunto de dados inviabiliza a validação adequada do comportamento do modelo frente a sinais espelhados.

Já as transformações relacionadas à cor foram evitadas pelo fato de as etapas posteriores de processamento envolverem a extração de marcos corporais, de modo que alterações cromáticas poderiam comprometer a precisão desse processo e, conseqüentemente, o desempenho do modelo.

Concluídas as etapas de preparação e ampliação do conjunto de dados, o material obtido foi então utilizado como entrada para o processo de extração dos marcos corporais. Essa etapa é fundamental para a construção das representações que serão utilizadas pelo modelo de aprendizado profundo, cuja estrutura e funcionamento são detalhados na próxima seção.

5.2 PROCESSAMENTO E EXTRAÇÃO DOS MARCOS CORPORAIS

O processo de extração dos marcos corporais utilizados neste trabalho foi realizado por meio da solução *Holistic*, da biblioteca *MediaPipe*. Essa abordagem permite detectar simultaneamente múltiplas regiões corporais (mãos, face e pose), garantindo maior consistência espacial e temporal entre os quadros.

Realizar a extração simultânea de todos os 543 marcos é uma tarefa computacionalmente custosa, uma vez que envolve a aplicação de diferentes modelos especializados. Nesse contexto, são empregados modelos distintos para a extração dos marcos da face, das mãos e da pose corporal, os quais podem ser integrados de forma otimizada em uma abordagem unificada, como implementado em MediaPipe GitHub Holistic (2025).

Entretanto, como cada componente é especializado na estimação de um tipo específico de marco (face, mãos ou pose), os modelos são projetados para operar de forma complementar. Para isso, a abordagem adota um *pipeline* integrado, no qual cada região do corpo é processada de maneira adequada à sua natureza, permitindo que os diferentes modelos atuem de forma coordenada e consistente na extração dos marcos.

O processo se inicia com a estimação da pose e, a partir dos marcos obtidos, são definidas três regiões de interesse (*Regions Of Interest* – ROIs): uma para o rosto e duas para as mãos (uma para cada). Logo, a imagem de entrada com resolução total é recortado para cada ROI, permitindo que modelos específicos de face e mãos façam a estimação detalhada de seus marcos. Por fim, todos esses pontos de referência são unificados aos marcos da pose, resultando em um conjunto total de 543 marcos.

De modo a facilitar as detecções contínuas das ROIs da face e mãos em vídeos, pode-se utilizar uma estratégia de rastreamento temporal. Basicamente, a abordagem

parte da suposição de que essas regiões não mudam drasticamente de posição entre quadros consecutivos. Em situações de movimentos rápidos, o rastreador pode perder a região de interesse, o que faz com que uma nova detecção deva ser realizada. Desse modo, o *MediaPipe Holistic* mitiga esse problema aplicando a previsão da pose a cada quadro, de modo a reduzir o tempo de resposta e manter a consistência semântica entre corpo e partes (por exemplo, evitando confundir mãos esquerda e direita ou misturar membros de pessoas diferentes na cena).

Para finalizar, um outro desafio é que, mesmo com a pose estimada, as ROIs extraídas ainda podem ser imprecisas para recortes refinados de face e mãos. Para contornar isso, o *pipeline* utiliza modelos leves de recorte os quais atuam como transformadores espaciais (*spatial transformers*), ajustando as regiões com maior precisão. Esses modelos têm custo computacional reduzido, consumindo apenas cerca de 10% do tempo de inferência dos modelos principais.

As configurações adotadas foram baseadas em Sarmiento (2023) e ajustadas para melhorar a estabilidade e a precisão dos marcos corporais extraídos. A seguir, são descritas as principais definições empregadas nesta etapa:

- **Modo de imagem estática desabilitado:** permite o processamento contínuo dos vídeos, mantendo o rastreamento entre quadros sem a necessidade de realizar uma nova detecção do corpo a cada imagem;
- **Complexidade do modelo igual a 2:** configura o modelo de detecção de marcos de pose em seu nível máximo de complexidade (0, 1 ou 2), aumentando a acurácia da estimativa ao custo de maior tempo de processamento;
- **Suavização temporal dos marcos ativada:** aplica filtragem entre quadros consecutivos, reduzindo ruídos e flutuações indesejadas na detecção;
- **Segmentação desativada:** tanto a geração das máscaras de segmentação quanto sua suavização foram desabilitadas, já que essas informações não eram necessárias para o propósito deste trabalho;
- **Confiança mínima de detecção igual a 0,9:** especifica o limiar mínimo para considerar uma detecção inicial como válida;
- **Confiança mínima de rastreamento igual a 0,9:** define o limiar mínimo exigido para manter o rastreamento confiável dos marcos ao longo da sequência.

Vale destacar que, neste trabalho, os marcos faciais não foram utilizados. Isso se deve ao fato de que, em alguns sinais do conjunto, os movimentos não apresentam expressividade facial significativa ou as expressões são sutis, tornando difícil extrair informações consistentes. A inclusão dos marcos faciais nesses casos poderia introduzir

ruído nos dados e prejudicar o aprendizado do modelo. Dessa forma, optou-se por utilizar apenas os marcos das mãos e da pose corporal, garantindo maior uniformidade e foco nas informações mais relevantes para a tarefa de classificação, sem descartar a possibilidade de explorar expressões faciais em sinais onde elas sejam mais pronunciadas.

Com os marcos extraídos e organizados, a etapa seguinte consiste em estruturar esses dados de forma adequada para o treinamento do modelo de aprendizado profundo. A partir do processamento realizado pelo *MediaPipe Holistic*, cada quadro retorna 75 marcos tridimensionais — 33 referentes à pose e 42 às mãos (21 para cada mão).

A forma mais usual de empregar os marcos como entrada para modelos sequenciais é a concatenação de todas as coordenadas em um único vetor unidimensional, processo chamado de achatamento dos marcos 3D. Cada marco fornecido pelo *MediaPipe Holistic* possui três coordenadas:

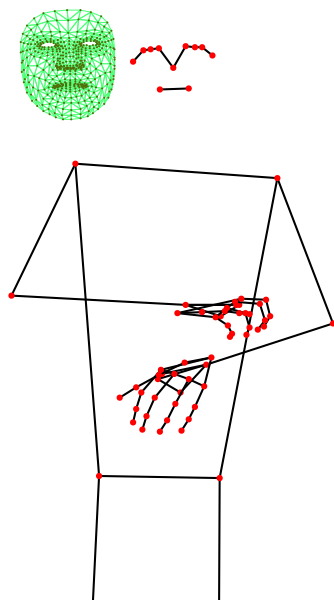
- x : posição horizontal do ponto na imagem;
- y : posição vertical do ponto na imagem;
- z : profundidade relativa do ponto em relação ao plano da câmera.

Dessa forma, cada quadro é convertido em um vetor contendo 225 valores, considerando que cada um dos 75 marcos possui essas três coordenadas (x, y, z) . Para cada vídeo, ao se utilizar sequências de 30 quadros, isso resulta em tensores de dimensão 30×225 , prontos para serem utilizados como entrada em modelos de classificação sequencial.

Após o processo de achatamento, todos esses vetores foram armazenados em arquivos da biblioteca NumPy (HARRIS et al., 2020), no formato `.npy`, o que permite acesso eficiente aos dados e garante consistência no formato das amostras, além de evitar que o processo de extração precise ser reexecutado.

Assim, a Figura 12 apresenta os marcos das mãos, face e pose gerados pelo *MediaPipe Holistic* sobre um vídeo do conjunto de dados desenvolvido neste trabalho. É interessante ressaltar que, nesse caso, somente a parte superior do indivíduo é identificada, isto se deve pelo enquadramento do vídeo. Vale destacar que, embora a figura exiba os marcos apenas em duas dimensões, o *MediaPipe Holistic* fornece as coordenadas em 3D desses pontos de referência.

Figura 12 – Marcos da face, pose e mãos estimados pelo *MediaPipe Holistic*. Para melhor visualização, os marcos da face foram deslocados levemente para a esquerda, evitando que fiquem sobrepostos aos marcos de pose.



Fonte: Elaborada pelo autor (2026).

5.3 MODELO DE REDE NEURAL

No presente trabalho, desenvolveu-se um modelo que consiste em quatro camadas principais: Entrada, Máscara, LSTM e Classificação.

Os vetores de marcos tridimensionais achatados são inicialmente passados pela camada de Entrada, seguidos de uma camada de Máscara (*masking*), responsável por ignorar valores nulos provenientes do processo de extração dos marcos.

Em seguida, a camada LSTM é empregada para capturar as dependências temporais presentes nos gestos, utilizando uma função de ativação ReLU bem como o *dropout* recorrente para técnica de regularização. A Figura 13 ilustra o funcionamento interno de uma célula LSTM com a presença do *dropout* recorrente, representado por conexões com linhas pontilhadas.

biblioteca *TensorFlow* (ABADI et al., 2015). A Tabela 2 apresenta o número de parâmetros por cada camada presente na arquitetura base dos modelos.

Tabela 2 – Relação do número de parâmetros pelo tipo de camada.

Tipo de camada	Dimensão de saída	Número de parâmetros
Entrada	$n_{\text{quadros}} \times n_{\text{características}}$	0
Máscara	$n_{\text{quadros}} \times n_{\text{características}}$	0
LSTM	$n_{\text{neurônios}}$	$4 \cdot ((n_{\text{características}} + 1) \cdot n_{\text{neurônios}} + n_{\text{neurônios}}^2)$
Classificação	n_{classes}	$(n_{\text{neurônios}} + 1) \cdot n_{\text{classes}}$

Fonte: Elaborada pelo autor (2026).

Como mencionado anteriormente, neste trabalho definiu-se $n_{\text{quadros}} = 30$. Cada quadro fornece 225 características, totalizando $n_{\text{características}} = 225$. Para ilustrar o cálculo do número de parâmetros, considere um exemplo em que a camada LSTM é configurada com 64 neurônios, isto é, $n_{\text{neurônios}} = 64$. Nesse cenário, essa camada possui 74.240 parâmetros. De forma análoga, para um problema com $n_{\text{classes}} = 10$, a camada de saída contém 650 parâmetros.

Esses valores são apresentados apenas como referência, uma vez que diferentes configurações de número de neurônios e de classes são analisadas nos experimentos.

6 EXPERIMENTOS E RESULTADOS

Este capítulo descreve o ambiente computacional, os critérios de avaliação, procedimentos adotados e as configurações utilizadas na condução dos experimentos realizados. Além disso, são apresentados e analisados os resultados obtidos com os modelos propostos, incluindo uma discussão crítica sobre seu desempenho, limitações e principais características observadas durante a avaliação.

6.1 AMBIENTE COMPUTACIONAL

Para a realização dos experimentos foi utilizada uma máquina equipada com processador Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz, com 16 GB de memória RAM e com sistema operacional *Windows 11 Home Single Language*.

No entanto, por conveniência e praticidade no desenvolvimento, os códigos foram implementados e executados no ambiente virtual *Windows Subsystem for Linux* ²¹, utilizando o sistema operacional *Ubuntu 22.04.5 LTS*.

Como comentado, os modelos foram desenvolvidos utilizando a linguagem de programação *Python* na versão 3.10.2 com a biblioteca *TensorFlow*. Para gerenciamento de dependências e isolamento do ambiente de execução, utilizou-se a ferramenta *Pipenv*², que permite criar ambientes virtuais independentes e gerenciar versões específicas de bibliotecas. Esse isolamento evita conflitos entre pacotes e permite organizar diferentes configurações do ambiente de forma segura e independente.

6.2 PROTOCOLO DE TREINAMENTO

Em primeiro lugar, o conjunto de dados foi dividido por meio da técnica de validação cruzada *k-fold*. Essa técnica consiste em dividir o conjunto de dados em k partes aproximadamente iguais, de modo que, a cada iteração, $k - 1$ partes são utilizadas para treinamento enquanto a parte restante é destinada ao teste. Esse processo é repetido até que todas as partes tenham sido utilizadas uma vez como conjunto de teste, resultando em k estimativas independentes de desempenho (STONE, 1974).

Neste trabalho, adotou-se $k = 10$, em conformidade com o esquema *leave-one-subject-out*. Nesse procedimento, cada indivíduo é isolado para compor o conjunto de teste, enquanto os demais são utilizados no treinamento. Como discutido em seções anteriores, apesar de haver repetições de um mesmo participante no conjunto de dados, as diferenças decorrentes das condições de gravação levaram à consideração de cada ocorrência como

²¹ <<https://learn.microsoft.com/windows/wsl/>>.

² <<https://pipenv.pypa.io/>>.

um indivíduo distinto. Dessa forma, obteve-se um total de 10 sujeitos, viabilizando o uso da técnica *10-fold* alinhada ao esquema *leave-one-subject-out*.

Todos os experimentos utilizaram como função de perda a entropia cruzada categórica e o otimizador AdamW com decaimento de peso igual a 0,005. Entretanto, parâmetros como a quantidade de neurônios e a taxa do *dropout* recorrente presente na camada LSTM, tamanho do *batch*, taxa de aprendizado e número de épocas variam entre os modelos e os cenários experimentais. Os valores dos parâmetros avaliados estão apresentados na Tabela 3.

Tabela 3 – Parâmetros avaliados no treinamento dos modelos.

Parâmetro	Valores testados
Neurônios na camada LSTM	32, 64, 128
Taxa do <i>dropout</i> recorrente	1×10^{-1} , 2×10^{-1}
Tamanho do <i>batch</i>	32, 64, 128, 256
Taxa de aprendizado	1×10^{-3} , 1×10^{-4}
Épocas	150, 250, 300, 350, 400

Fonte: Elaborada pelo autor (2026).

Além disso, cada experimento conta com 20 avaliações para cada conjunto de teste da validação cruzada. Isso ocorre porque, conforme descrito anteriormente, o processo de seleção de quadros baseado em amostragem pela distribuição normal gera 20 versões distintas de cada vídeo. Assim, para cada *fold* do esquema *leave-one-subject-out*, o conjunto de teste é avaliado 20 vezes, produzindo estimativas mais estáveis e permitindo uma análise mais robusta do desempenho dos modelos. No entanto, é interessante ressaltar que, essas 20 amostras, são as mesmas obtidas durante o processo de aumento de dados descrito na Seção 5.1.2, porém, sem que fossem submetidas às transformações aplicadas aos dados de treinamento.

6.3 CRITÉRIOS DE AVALIAÇÃO

Para avaliar os modelos de classificação, foram utilizados alguns critérios de avaliação que determinam a qualidade dos resultados. Inicialmente, é importante compreender os elementos que compõem esses critérios. Para isso, a Tabela 4 apresenta a estrutura de uma matriz de confusão para um problema de classificação binária, destacando as quantidades de predições corretas e incorretas realizadas pelo modelo.

Tabela 4 – Estrutura de uma matriz de confusão para um problema de classificação binária.

	Classe Predita	
	Positiva	Negativa
Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Elaborada pelo autor (2026).

Os elementos da matriz de confusão possuem interpretações específicas no contexto da avaliação de classificadores. No caso binário, esses termos são definidos da seguinte forma:

- Verdadeiro Positivo (VP): quantidade de amostras que pertencem à classe positiva e foram corretamente classificadas como positivas pelo modelo.
- Falso Positivo (FP): quantidade de amostras que pertencem à classe negativa, mas foram incorretamente classificadas como positivas.
- Falso Negativo (FN): quantidade de amostras que pertencem à classe positiva, mas foram incorretamente classificadas como negativas. Nesse caso, o modelo deixa de identificar uma ocorrência real da classe positiva.
- Verdadeiro Negativo (VN): quantidade de amostras que pertencem à classe negativa e foram corretamente classificadas como negativas.

Em problemas de classificação com N classes, esses mesmos conceitos podem ser interpretados de forma análoga, considerando cada classe individualmente como a “classe de interesse”. A matriz de confusão passa a ser organizada de modo que cada linha representa as classes reais e cada coluna representa as classes preditas. Assim, para uma classe C_i , têm-se:

- Verdadeiro Positivo (VP): quantidade de amostras de C_i corretamente classificadas como C_i (elemento da diagonal principal).
- Falso Positivo (FP): quantidade de amostras que não pertencem a C_i e foram classificadas incorretamente como C_i (demais elementos da coluna de C_i).
- Falso Negativo (FN): quantidade de amostras de C_i que não foram classificadas como C_i (demais elementos da linha de C_i).
- Verdadeiro Negativo (VN): quantidade de amostras que não pertencem a C_i e não foram classificadas como C_i .

Dentre os critérios de interesse, tem-se a acurácia, responsável por informar quantas predições foram classificadas corretamente dentre todas as predições realizadas. Seu cálculo é definido como:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}.$$

Já, a precisão indica a quantidade de predições corretas dentre aquelas que o modelo fez para a classe C_i , podendo ser calculada da seguinte forma:

$$\text{Precisão} = \frac{VP}{VP + FP}.$$

Ou seja, uma alta precisão indica que quando o modelo realiza a predição para a classe C_i , poucos casos são pertencentes à outras classes que não seja essa. No entanto, a avaliação isolada da precisão, pode trazer alguns problemas, pois para casos onde há uma alta precisão, o modelo pode estar ignorando diversos casos onde a classe deveria ser classificada.

Com, isso, o critério de revocação (*recall*) é responsável por determinar a quantidade correta de predições em que determinada classe era esperada. Sua fórmula é dada por:

$$\text{Recall} = \frac{VP}{VP + FN}.$$

Dessa forma, valores elevados de *recall* indicam que o modelo apresenta boa capacidade de identificar as instâncias pertencentes à classe considerada, minimizando a ocorrência de falsos negativos. Já, no caso em que se deseja observar simultaneamente tanto a precisão quanto o *recall*, isso pode ser feito utilizando o critério F_1 score o qual representa uma média harmônica entre esses dois critérios. Ele é definido como:

$$F_1 = 2 \cdot \frac{\text{precisão} \cdot \text{recall}}{\text{precisão} + \text{recall}}.$$

No entanto, é possível também ter uma generalização do F_1 score, chamado de F_β score que realiza uma ponderação, por meio da inserção de um valor β , no valor do *recall* que define o quanto mais importante ele será do que a precisão para o cálculo do critério. Assim tem-se:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precisão} \cdot \text{recall}}{\beta^2 \cdot \text{precisão} + \text{recall}}.$$

Dessa maneira, é possível verificar que, quando o valor de β é maior do que 1, sendo então o critério F_2 score, ocorre uma priorização do *recall* ao invés da precisão. Neste trabalho, os modelos serão avaliados utilizando os critérios de acurácia, precisão, *recall*, F_1 score e F_2 score, calculadas tanto sobre os conjuntos de teste gerados pela técnica de validação cruzada quanto sobre o conjunto de dados de teste adicional.

6.4 RESULTADOS

Nesta seção são apresentados os resultados obtidos nos experimentos realizados para o problema de classificação de sinais de Libras. Os desempenhos dos modelos são analisados com base nas 20 avaliações realizadas para cada conjunto de teste da validação cruzada, bem como nas matrizes de confusão do melhor conjunto e do conjunto de dados de teste adicional.

6.4.1 Problema de classificação de 10 classes

Em primeiro lugar, tem-se o problema de classificação das 10 classes que compõem o conjunto de dados desenvolvido neste trabalho. O experimento que obteve melhor resultado utilizou um tamanho de lote (*batch size*) igual a 128, taxa de aprendizado igual a 0,001, número de neurônios na camada LSTM igual a 64, taxa do *dropout* recorrente igual a 0,2 e um número de épocas igual a 300.

A Tabela 5 apresenta a acurácia geral obtida bem como a acurácia de todas as classes em cada uma das 20 avaliações realizadas para os conjuntos de teste da validação cruzada. Além disso, são exibidos a média das acurácias e o desvio padrão, permitindo uma análise mais estável do desempenho do modelo.

Tabela 5 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de 10 classes.

Avaliação	Alergia	Autismo	Bronquite	Câimbra	Convulsão	Diarreia	Fratura	Frieira	Infarto	Sarampo	Geral
1	74,00	94,00	97,92	56,25	92,00	100,00	98,00	60,00	80,00	91,49	84,35
2	66,00	96,00	100,00	58,33	90,00	95,92	96,00	52,00	86,00	93,62	83,33
3	76,00	92,00	97,92	58,33	90,00	97,96	96,00	56,00	82,00	93,62	83,94
4	68,00	92,00	97,92	58,33	90,00	95,92	94,00	60,00	82,00	95,74	83,33
5	74,00	90,00	97,92	64,58	90,00	100,00	92,00	56,00	86,00	95,74	84,55
6	78,00	94,00	97,92	56,25	90,00	100,00	94,00	56,00	82,00	93,62	84,15
7	68,00	96,00	100,00	60,42	90,00	100,00	96,00	56,00	82,00	93,62	84,15
8	74,00	94,00	100,00	52,08	90,00	97,96	92,00	50,00	82,00	91,49	82,32
9	78,00	96,00	100,00	54,17	90,00	95,92	96,00	50,00	82,00	89,36	83,13
10	72,00	92,00	100,00	52,08	90,00	95,92	98,00	56,00	86,00	95,74	83,74
11	74,00	98,00	100,00	54,17	92,00	95,92	96,00	60,00	82,00	95,74	84,76
12	74,00	90,00	100,00	56,25	90,00	97,96	94,00	58,00	80,00	95,74	83,54
13	76,00	92,00	97,92	62,50	90,00	97,96	94,00	58,00	86,00	93,62	84,76
14	76,00	92,00	97,92	58,33	90,00	100,00	96,00	62,00	84,00	93,62	84,96
15	72,00	96,00	97,92	52,08	90,00	97,96	92,00	58,00	86,00	93,62	83,54
16	72,00	96,00	100,00	54,17	88,00	93,88	92,00	52,00	86,00	93,62	82,72
17	72,00	94,00	100,00	58,33	90,00	97,96	98,00	56,00	84,00	93,62	84,35
18	70,00	88,00	100,00	54,17	90,00	97,96	94,00	56,00	84,00	93,62	82,72
19	78,00	94,00	97,92	58,33	90,00	100,00	98,00	52,00	82,00	91,49	84,15
20	74,00	94,00	97,92	58,33	90,00	97,96	96,00	54,00	86,00	93,62	84,15
Média	73,30	93,50	98,96	56,87	90,10	97,86	95,10	55,90	83,50	93,62	83,83
Desvio padrão	3,30	2,44	1,04	3,30	0,77	1,76	2,05	3,32	2,09	1,65	0,72

Fonte: Elaborada pelo autor (2026).

Com base nos valores obtidos, observa-se que o modelo apresenta uma boa qualidade geral, com média das acurácias gerais entre os diferentes conjuntos de teste igual a 83,83%. Além disso, o modelo demonstra boa estabilidade, uma vez que as acurácias gerais são bastante próximas entre si, o que é evidenciado pelo desvio padrão reduzido de 0,72%.

Outro ponto a ser destacado é que, para a maior parte das classes, os resultados permaneceram consistentes ao longo dos conjuntos, o que pode ser constatado pelos baixos valores de desvio padrão por classe. Isso indica que o modelo não apenas possui bom desempenho médio, mas também é estável, mantendo previsões semelhantes mesmo sob variações no conjunto de dados.

Entretanto, observa-se que algumas classes, em especial “Alergia”, “Câimbra” e “Frieira”, apresentam valores médios de acurácia inferiores aos das demais, além de desvios padrão ligeiramente mais elevados. Esse comportamento sugere maior sensibilidade do modelo à composição do conjunto de teste para essas classes.

Com o objetivo de investigar esse comportamento com maior nível de detalhe, a Tabela 6 apresenta os resultados de acurácia obtidos para cada classe no conjunto que apresentou o melhor valor de acurácia geral, o conjunto 14, cuja acurácia foi de 84,96%. Essa apresentação individualizada permite analisar com maior precisão a variação do desempenho entre os *folds*, especialmente para as classes que apresentaram maior instabilidade na análise global.

Tabela 6 – Resultados da acurácia, geral e por classe, em porcentagem, para cada *fold* da validação cruzada para o conjunto que obteve melhor acurácia geral (conjunto 14) do problema de 10 classes.

<i>Fold</i>	Alergia	Autismo	Bronquite	Câimbra	Convulsão	Diarreia	Fratura	Frieira	Infarto	Sarampo	Geral
1	100,00	100,00	100,00	20,00	100,00	100,00	100,00	80,00	80,00	100,00	88,00
2	100,00	100,00	66,67	100,00	0,00	100,00	80,00	20,00	80,00	100,00	73,91
3	40,00	100,00	100,00	0,00	100,00	100,00	100,00	100,00	100,00	80,00	82,00
4	100,00	100,00	100,00	100,00	100,00	100,00	80,00	0,00	100,00	100,00	88,00
5	60,00	40,00	100,00	50,00	100,00	100,00	100,00	0,00	100,00	0,00	69,57
6	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
7	80,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	98,00
8	0,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	90,00
9	80,00	80,00	100,00	0,00	100,00	100,00	100,00	20,00	0,00	100,00	68,00
10	100,00	100,00	100,00	20,00	100,00	100,00	100,00	100,00	80,00	100,00	90,00
Média	76,00	92,00	96,67	59,00	90,00	100,00	96,00	62,00	84,00	88,00	84,79
Desvio padrão	32,00	18,33	10,00	43,00	30,00	0,00	8,00	43,31	29,39	29,93	10,58

Fonte: Elaborada pelo autor (2026).

A partir da Tabela 6, é possível reforçar as observações já discutidas anteriormente. Nota-se que as classes “Alergia”, “Câimbra” e “Frieira” que apresentaram desempenho mais baixo na análise global exibem uma maior variação entre os *folds*, evidenciada pelos altos valores de desvio padrão. Essa oscilação indica que o desempenho do modelo para essas classes depende fortemente de quais indivíduos compõem o conjunto de teste em cada partição, o que sugere menor capacidade de generalização para esses casos.

Por outro lado, classes como “Bronquite”, “Diarreia” e “Fratura” mantêm desempenhos mais consistentes entre os *folds*, apresentando desvios padrão menores. Essa estabilidade reforça que, nessas classes, o modelo consegue capturar padrões mais estáveis, menos sensíveis à variação entre sujeitos.

A Tabela 7 apresenta os valores de precisão, *recall*, F_1 score e F_2 score obtidos para

cada classe no conjunto 14, permitindo uma avaliação mais detalhada do desempenho do modelo.

Tabela 7 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 14) do problema de 10 classes.

Classe	Precisão (%)	Recall (%)	F_1 score (%)	F_2 score (%)
Alergia	74,51	76,00	75,25	75,70
Autismo	93,88	92,00	92,93	92,37
Bronquite	87,04	97,92	92,16	95,53
Câimbra	87,50	58,33	70,00	62,50
Convulsão	90,00	90,00	90,00	90,00
Diarreia	98,00	100,00	98,99	99,59
Fratura	72,73	96,00	82,76	90,23
Frieira	67,39	62,00	64,58	63,01
Infarto	100,00	84,00	91,30	86,78
Sarampo	84,62	93,62	88,89	91,67

Fonte: Elaborada pelo autor (2026).

A partir da Tabela 7, é possível observar diferenças significativas no comportamento do modelo entre as classes. De modo geral, classes como “Autismo”, “Bronquite”, “Convulsão”, “Diarreia” e “Sarampo” apresentam resultados elevados e relativamente balanceados entre precisão e *recall*, indicando que o modelo consegue tanto identificar corretamente os exemplos positivos quanto evitar falsos positivos. Nesse grupo, destaca-se a classe “Diarreia”, que atinge valores superiores a 98% em todas os critérios, evidenciando um desempenho bastante consistente.

Por outro lado, algumas classes apresentam desequilíbrios importantes. A classe “Câimbra” possui precisão razoável (87,50%), mas um *recall* bastante baixo (58,33%), refletindo uma quantidade significativa de falsos negativos. Esse comportamento faz com que seus valores de F_1 e principalmente F_2 diminuam consideravelmente, já que o F_2 penaliza de forma mais intensa situações em que o modelo deixa de identificar casos positivos. Situação semelhante ocorre na classe “Frieira”, que apresenta desempenho baixo e relativamente equilibrado entre precisão e *recall*, indicando dificuldade geral do modelo em distinguir essa categoria.

Algumas classes ainda revelam padrões específicos. “Fratura”, por exemplo, possui precisão mais baixa, mas um *recall* alto, sugerindo tendência do modelo a classificar mais exemplos como pertencentes a essa classe, aumentando verdadeiros positivos, porém ao custo de mais falsos positivos. Já a classe “Infarto” apresenta *recall* menor que a precisão, o que indica que o modelo é mais conservador e deixa de detectar alguns casos.

Além disso, a Figura 14 apresenta a matriz de confusão geral para o conjunto 14, englobando todos os conjuntos de testes referentes à cada *fold*. A matriz de confusão apresentada na Figura 14 complementa os resultados da Tabela 7, permitindo visualizar

de forma explícita como ocorrem os erros entre as classes. As classes que apresentaram desempenho mais baixo na Tabela 7, como “Alergia”, “Câimbra” e “Frieira”, são também aquelas que exibem os maiores índices de confusão. A classe “Alergia” é frequentemente confundida com “Frieira”, e o inverso também ocorre, o que está de acordo com os valores reduzidos de F_1 score e F_2 score observados para essas categorias. Esse comportamento sugere que ambas as classes compartilham características visuais semelhantes, levando o modelo a dificuldade em diferenciá-las.

Figura 14 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 14) para o problema de 10 classes.

Alergia	76,00% 38	0,00% 0	0,00% 0	0,00% 0	2,00% 1	0,00% 0	0,00% 0	14,00% 7	0,00% 0	8,00% 4
Autismo	0,00% 0	92,00% 46	0,00% 0	4,00% 2	0,00% 0	0,00% 0	4,00% 2	0,00% 0	0,00% 0	0,00% 0
Bronquite	0,00% 0	0,00% 0	97,92% 47	0,00% 0	0,00% 0	2,08% 1	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Câimbra	2,08% 1	6,25% 3	0,00% 0	58,33% 28	0,00% 0	0,00% 0	22,92% 11	10,42% 5	0,00% 0	0,00% 0
Convulsão	10,00% 5	0,00% 0	0,00% 0	0,00% 0	90,00% 45	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Diarreia	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	100,00% 49	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Fratura	0,00% 0	0,00% 0	0,00% 0	4,00% 2	0,00% 0	0,00% 0	96,00% 48	0,00% 0	0,00% 0	0,00% 0
Frieira	14,00% 7	0,00% 0	0,00% 0	0,00% 0	8,00% 4	0,00% 0	10,00% 5	62,00% 31	0,00% 0	6,00% 3
Infarto	0,00% 0	0,00% 0	14,00% 7	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	84,00% 42	2,00% 1
Sarampo	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	6,38% 3	0,00% 0	93,62% 44
	Alergia	Autismo	Bronquite	Câimbra	Convulsão	Diarreia	Fratura	Frieira	Infarto	Sarampo

Fonte: Elaborada pelo autor (2026).

De forma semelhante, a classe “Câimbra” apresenta o menor valor de *recall* entre todas as classes, e a matriz de confusão evidencia essa dificuldade ao mostrar que grande parte dos seus erros ocorre especificamente com a classe “Fratura”. Esse padrão reforça

a interpretação de que o modelo, ao tentar reconhecer instâncias de “Câimbra”, tende a superestimar a classe “Fratura”, o que explica o alto *recall* observado para esta última.

Outro caso interessante é o da classe “Infarto”, que na Tabela 7 apresenta precisão maior que *recall*. A matriz de confusão revela a origem desse comportamento: uma parcela considerável de exemplos de “Infarto” é classificada incorretamente como “Bronquite”. Isso justifica a redução do *recall* e evidencia uma relação de confusão bem definida entre essas duas classes.

Analisando esses fatores, observou-se que os casos de confusão estão relacionados principalmente à proximidade visual e gestual entre alguns sinais. A classe “Frieira” apresenta movimentos muito semelhantes aos da classe “Alergia”, diferenciando-se apenas em etapas iniciais do gesto. Da mesma forma, as classes “Bronquite” e “Infarto” compartilham um padrão gestual próximo à região do tórax, o que contribui para as confusões observadas na matriz de confusão. Já a classe “Câimbra” apresenta movimentos que lembram os de “Fratura”, pois ambas envolvem gestos executados à frente do corpo com indicação de dor ou ruptura.

Diante desses padrões, foi feita a junção desses casos em pares. Cada par representa um caso de confusão: “Alergia” e “Frieira”, “Bronquite” e “Infarto” e, por fim, “Câimbra” e “Fratura”. Em seguida, foi desenvolvido um modelo que realiza a classificação de um total de 7 classes: “Alergia_Frieira”, “Autismo”, “Bronquite_Infarto”, “Câimbra_Fratura”, “Convulsão”, “Diarreia” e “Sarampo”. Posteriormente, esses pares são classificados por meio de modelos especialistas, também desenvolvidos neste trabalho.

6.4.2 Problema de classificação de 7 classes

Para o problema de classificação de 7 classes, o experimento que obteve melhor resultado utilizou um tamanho de lote igual a 128, taxa de aprendizado igual a 0,001, número de neurônios na camada LSTM igual a 32, taxa do *dropout* recorrente igual a 0,2 e um número de épocas igual a 400. Destaca-se ainda que, para manter o balanceamento do conjunto de dados, nos sinais presentes nas classes unificadas foram selecionadas as 10 primeiras amostras de cada classe antes da união, totalizando 20 amostras por classe combinada. Dessa forma, garantiu-se que todas as classes permanecessem com praticamente a mesma quantidade de instâncias. No entanto, os conjuntos de teste foram mantidos os mesmos, sem alteração nesse quesito.

A Tabela 8 apresenta a acurácia geral obtida bem como a acurácia geral de todas as classes em cada uma das 20 avaliações realizadas para os conjuntos de teste da validação cruzada. Além disso, são exibidos a média das acurácias e o desvio padrão, permitindo uma análise mais estável do desempenho do modelo.

Tabela 8 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de 7 classes.

Avaliação	Alergia_Frieira	Autismo	Bronquite_Infarto	Câimbra_Fratura	Convulsão	Diarreia	Sarampo	Geral
1	91,00	100,00	100,00	97,96	88,00	95,92	100,00	96,14
2	89,00	100,00	98,98	98,98	84,00	97,96	100,00	95,53
3	88,00	100,00	98,98	98,98	88,00	97,96	100,00	95,73
4	90,00	100,00	98,98	98,98	88,00	97,96	100,00	96,14
5	93,00	100,00	98,98	95,92	88,00	100,00	100,00	96,34
6	91,00	100,00	100,00	96,94	88,00	97,96	100,00	96,14
7	90,00	100,00	97,96	96,94	90,00	95,92	97,87	95,33
8	91,00	100,00	100,00	98,98	88,00	95,92	100,00	96,34
9	90,00	100,00	98,98	97,96	90,00	93,88	100,00	95,73
10	90,00	100,00	98,98	96,94	88,00	95,92	97,87	95,33
11	91,00	100,00	98,98	97,96	90,00	93,88	100,00	95,93
12	90,00	100,00	100,00	98,98	90,00	97,96	100,00	96,54
13	93,00	100,00	97,96	97,96	88,00	97,96	97,87	96,14
14	92,00	100,00	100,00	97,96	88,00	97,96	100,00	96,54
15	91,00	100,00	98,98	96,94	88,00	97,96	95,74	95,53
16	91,00	100,00	98,98	98,98	88,00	93,88	100,00	95,93
17	93,00	100,00	98,98	96,94	88,00	93,88	100,00	95,93
18	91,00	100,00	98,98	98,98	90,00	97,96	100,00	96,54
19	92,00	100,00	97,96	98,98	88,00	95,92	100,00	96,14
20	93,00	100,00	98,98	100,00	90,00	97,96	100,00	97,15
Média	91,00	100,00	99,10	98,10	88,40	96,80	99,50	96,06
Desvio padrão	1,34	0,00	0,64	1,03	1,36	1,76	1,14	0,45

Fonte: Elaborada pelo autor (2026).

De modo geral, observa-se uma melhora significativa no desempenho do modelo quando comparado ao cenário original de 10 classes. A acurácia média geral aumentou para 96,06%, o que representa um ganho de 12,12 pontos percentuais em relação ao caso anterior. Além disso, o modelo manteve elevada estabilidade entre os diferentes conjuntos, evidenciada pelo baixo desvio padrão de 0,45%.

Esse comportamento sugere que a estratégia de agrupamento das classes que apresentavam maior confusão contribui efetivamente para reduzir ambiguidades e melhorar a separabilidade entre os sinais. O impacto é particularmente perceptível nos pares anteriormente responsáveis pela maior parte dos erros de classificação: o grupo “Alergia_Frieira” obteve acurácia média de 91,00%, enquanto os grupos “Bronquite_Infarto” e “Câimbra_Fratura” alcançaram médias de 99,10% e 98,10%, respectivamente, evidenciando um ganho expressivo de desempenho após a união das classes.

Com base nesses resultados, procede-se à análise detalhada do conjunto com melhor desempenho geral, o conjunto 20, cuja acurácia alcançou 97,15%. A Tabela 9 apresenta as acurácias por classe em cada *fold* para esse conjunto, possibilitando uma investigação mais aprofundada do comportamento do modelo sob a variação de indivíduos.

Tabela 9 – Resultados da acurácia geral e por classe, em porcentagem, para cada *fold* da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) do problema de 7 classes.

<i>Fold</i>	Alergia_Frieira	Autismo	Bronquite_Infarto	Câimbra_Fratura	Convulsão	Diarreia	Sarampo	Geral
<i>Fold</i> 1	90,00	100,00	100,00	100,00	100,00	100,00	100,00	98,00
<i>Fold</i> 2	80,00	100,00	100,00	100,00	100,00	100,00	100,00	95,65
<i>Fold</i> 3	70,00	100,00	100,00	100,00	0,00	100,00	100,00	84,00
<i>Fold</i> 4	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
<i>Fold</i> 5	100,00	100,00	100,00	100,00	100,00	80,00	100,00	97,83
<i>Fold</i> 6	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
<i>Fold</i> 7	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
<i>Fold</i> 8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
<i>Fold</i> 9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
<i>Fold</i> 10	90,00	100,00	90,00	100,00	100,00	100,00	100,00	96,00
Média	93,00	100,00	99,00	100,00	90,00	98,00	100,00	97,25
Desvio padrão	10,05	0,00	3,00	0,00	30,00	6,00	0,00	4,65

Fonte: Elaborada pelo autor (2026).

A partir dos valores apresentados na Tabela 9, observa-se uma melhora em relação aos resultados obtidos anteriormente para o problema de 10 classes (Tabela 6). As médias das acurácias por classe aumentaram de forma consistente, refletindo diretamente no acréscimo da acurácia geral média dos *folds*, que passou de 84,79% para 97,25%. Essa melhora é acompanhada por uma redução substancial no desvio padrão da acurácia geral, de 10,58% para apenas 4,65%, indicando que o modelo tornou-se mais estável frente à variação entre indivíduos.

Outro aspecto relevante é a diminuição dos desvios padrão das acurácias individuais das classes. De forma geral, todas as classes apresentaram maior consistência entre os *folds* quando comparadas ao cenário anterior. A única exceção permanece sendo a classe “Convulsão”, cujo desvio padrão ainda é elevado (30%), característica que já havia sido observada na Tabela 6. Esse comportamento sugere que essa classe, em particular, mantém maior variabilidade entre os participantes, possivelmente devido a diferenças mais marcantes na execução do sinal.

Esses resultados reforçam que a estratégia de fusão das classes previamente confundidas não apenas elevou o desempenho médio do classificador, mas também contribuiu para uma resposta mais homogênea ao longo dos diferentes conjuntos de teste.

A Tabela 10 apresenta os critérios de precisão, *recall*, F_1 score e F_2 score para o conjunto que obteve a maior acurácia geral.

Tabela 10 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 20) do problema de 7 classes.

Classe	Precisão (%)	Recall (%)	F_1 score (%)	F_2 score (%)
Alergia_Frieira	94,90	93,00	93,94	93,37
Autismo	100,00	100,00	100,00	100,00
Bronquite_Infarto	98,98	98,98	98,98	98,98
Câimbra_Fratura	98,99	100,00	99,49	99,80
Convulsão	97,83	90,00	93,75	91,46
Diarreia	97,96	97,96	97,96	97,96
Sarampo	90,38	100,00	94,95	97,92

Fonte: Elaborada pelo autor (2026).

A análise dos critérios apresentados na Tabela 10 confirma o impacto positivo da estratégia de união das classes anteriormente responsáveis pelos maiores índices de confusão. De modo geral, observa-se um desempenho elevado em todas os critérios, com valores de precisão, *recall*, F_1 score e F_2 score superiores a 90% para todas as classes, o que evidencia uma melhora significativa em relação ao cenário de 10 classes.

A classe “Alergia_Frieira”, formada pela união de duas das categorias mais problemáticas na classificação original, apresenta resultados equilibrados e superiores a 93%, indicando que a união contribuiu diretamente para a redução dos erros de confusão observados anteriormente. Resultado semelhante ocorre para o par “Bronquite_Infarto”, que obteve valores próximos de 99% em todos os critérios, demonstrando que o agrupamento eliminou praticamente todas as ambiguidades vistas entre essas duas classes no problema anterior.

O par “Câimbra_Fratura” também apresentou desempenho expressivo, com F_1 score de 99,49% e *recall* de 100%, revelando que a união foi particularmente eficaz para esse caso. Já a classe “Convulsão” mantém resultados elevadas, embora seu *recall* (90%) seja inferior às demais, o que está alinhado com as variações observadas nas análises por *fold*. Ainda assim, o desempenho geral permanece estável.

As classes que não sofreram fusão (“Autismo”, “Diarreia” e “Sarampo”) também apresentam resultados muito altos, reforçando a consistência do modelo na nova configuração de classes. Entre elas, destaca-se a classe “Autismo”, com desempenho perfeito em todas os critérios.

A Figura 15 apresenta a matriz de confusão geral para o conjunto 20, englobando todos os conjuntos de testes referentes à cada *fold*.

Figura 15 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) para o problema de 7 classes.

Alergia_Frieira	93,00% 93	0,00% 0	0,00% 0	1,00% 1	1,00% 1	0,00% 0	5,00% 5
Autismo	0,00% 0	100,00% 50	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Bronquite_Infarto	0,00% 0	0,00% 0	98,98% 97	0,00% 0	0,00% 0	1,02% 1	0,00% 0
Câimbra_Fratura	0,00% 0	0,00% 0	0,00% 0	100,00% 98	0,00% 0	0,00% 0	0,00% 0
Convulsão	8,00% 4	0,00% 0	2,00% 1	0,00% 0	90,00% 45	0,00% 0	0,00% 0
Diarreia	2,04% 1	0,00% 0	0,00% 0	0,00% 0	0,00% 0	97,96% 48	0,00% 0
Sarampo	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	100,00% 47
	Alergia_Frieira	Autismo	Bronquite_Infarto	Câimbra_Fratura	Convulsão	Diarreia	Sarampo

Fonte: Elaborada pelo autor (2026).

Por meio da análise da matriz de confusão apresentada na Figura 15, é possível observar que o modelo obteve um desempenho consistente para o problema de 7 classes, alcançando mais de 90% de acerto em todas elas. Além disso, nota-se que os erros entre classes foram substancialmente reduzidos, o que confirma que as confusões pontuais presentes no problema de 10 classes deixaram de ocorrer nesta configuração.

Observa-se também que as predições incorretas tornam-se mais esparsas e concentradas em poucas ocorrências residuais, indicando que o modelo passou a aprender padrões mais discriminativos após a união das classes originalmente ambíguas. Esse comportamento é especialmente evidente nos pares que apresentavam maior similaridade semântica no cenário anterior: “Alergia_Frieira”, “Bronquite_Infarto” e “Câimbra_Fratura”. Para essas combinações, a matriz revela praticamente ausência de erros, o que está em plena

concordância com os resultados da Tabela 10, nas quais todas essas classes superaram 93% de desempenho nos principais critérios.

Outro ponto relevante é a maior diagonalização da matriz, que evidencia uma capacidade mais clara do modelo em distinguir corretamente cada uma das classes, sem dispersão significativa para categorias adjacentes. A classe “Convulsão” continua sendo a mais sensível a variações, mas mesmo assim apresenta desempenho elevado, com poucos desvios fora da diagonal principal. Já classes como “Autismo”, “Diarreia” e “Sarampo” mantêm acurácia praticamente perfeita, reforçando a estabilidade do modelo nessas categorias.

Dessa forma, as próximas etapas do estudo concentram-se no desenvolvimento dos modelos especialistas responsáveis por discriminar, individualmente, as classes que foram agrupadas em pares.

6.4.3 Problemas de classificação utilizando especialistas

Para este caso, foram desenvolvidos 3 modelos distintos, um para cada par de classes. Em primeiro lugar, foi desenvolvido o modelo responsável por classificar o sinal entre as classes “Bronquite” e “Infarto”, bem como o modelo responsável por classificar o sinal entre as classes “Câimbra” e “Fratura”. Isso porque, eles seguiram a mesma metodologia de desenvolvimento dos processos descritos nas seções anteriores.

Assim, para o modelo que classifica os sinais entre “Bronquite” e “Infarto”, o experimento que obteve melhor resultado utilizou um tamanho de lote igual a 32, taxa de aprendizado igual a 0,001, número de neurônios na camada LSTM igual a 32, taxa do *dropout* recorrente igual a 0,2 e um número de épocas igual a 150.

A Tabela 11 apresenta a acurácia geral obtida bem como a acurácia geral das classes “Bronquite” e “Infarto” em cada uma das 20 avaliações realizadas para os conjuntos de teste da validação cruzada. Além disso, são exibidos a média das acurácias e o desvio padrão, permitindo uma análise mais estável do desempenho do modelo.

Tabela 11 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de classificação dos sinais “Bronquite” e “Infarto”.

Avaliação	Bronquite	Infarto	Geral
1	97,92	88,00	92,86
2	100,00	88,00	93,88
3	97,92	90,00	93,88
4	97,92	90,00	93,88
5	97,92	88,00	92,86
6	95,83	88,00	91,84
7	97,92	90,00	93,88
8	97,92	90,00	93,88
9	97,92	88,00	92,86
10	97,92	88,00	92,86
11	97,92	90,00	93,88
12	97,92	88,00	92,86
13	97,92	90,00	93,88
14	100,00	88,00	93,88
15	97,92	88,00	92,86
16	97,92	90,00	93,88
17	97,92	90,00	93,88
18	97,92	90,00	93,88
19	97,92	88,00	92,86
20	97,92	90,00	93,88
Média	98,02	89,00	93,42
Desvio padrão	0,80	1,00	0,60

Fonte: Elaborada pelo autor (2026).

Com base nos valores apresentados na Tabela 11, observa-se que o modelo destinado à discriminação entre os sinais “Bronquite” e “Infarto” apresenta desempenho consistente ao longo das 20 avaliações realizadas. As acurácias gerais mantêm-se em um intervalo estreito, variando entre 91,84% e 93,88%, o que evidencia a estabilidade do modelo mesmo diante das variações inerentes aos conjuntos de teste da validação cruzada.

Ao analisar individualmente o desempenho por classe, nota-se que a classe “Bronquite” apresenta acurácia sistematicamente elevada, com valores majoritariamente superiores a 97% e atingindo o máximo de 100% em duas avaliações. Essa regularidade é confirmada pela média de 98,02% e por um desvio padrão de apenas 0,80%, indicando que o modelo praticamente não possui dificuldade em reconhecer corretamente os casos de bronquite.

Por outro lado, a classe “Infarto”, embora também apresente desempenho satisfatório, exibe valores ligeiramente mais baixos, variando entre 88% e 90%, com média de 89,00% e desvio padrão de 1,00%. Esses resultados sugerem que, apesar de o modelo apresentar boa capacidade discriminativa, os casos dessa classe são relativamente mais

desafiadores, o que possivelmente se deve à semelhança estrutural de certos padrões de movimento em relação à classe “Bronquite”, origem da confusão observada no problema de 10 classes.

Diversas avaliações atingiram o maior valor de acurácia geral, igual a 93,88%. Assim, para fins de continuidade das análises, selecionou-se de maneira arbitrária o conjunto 20, que obteve esse desempenho máximo. A Tabela 12 apresenta, então, os resultados detalhados de acurácia por *fold* para esse conjunto selecionado.

Tabela 12 – Resultados da acurácia geral e por classe, em porcentagem para cada *fold* da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) do problema de classificação dos sinais “Bronquite” e “Infarto”.

<i>Fold</i>	Bronquite	Infarto	Geral
<i>Fold</i> 1	100,00	100,00	100,00
<i>Fold</i> 2	100,00	0,00	37,50
<i>Fold</i> 3	100,00	100,00	100,00
<i>Fold</i> 4	100,00	100,00	100,00
<i>Fold</i> 5	100,00	100,00	100,00
<i>Fold</i> 6	100,00	100,00	100,00
<i>Fold</i> 7	80,00	100,00	90,00
<i>Fold</i> 8	100,00	100,00	100,00
<i>Fold</i> 9	100,00	100,00	100,00
<i>Fold</i> 10	100,00	100,00	100,00
Média	98,00	90,00	92,75
Desvio padrão	6,00	30,00	18,66

Fonte: Elaborada pelo autor (2026).

Ao analisar os resultados apresentados na Tabela 12, observa-se que o modelo obteve um desempenho consistente na maior parte dos *folds*, alcançando acurácia geral de 100% em oito das dez partições avaliadas. Esse comportamento reforça a capacidade do modelo em distinguir corretamente os sinais “Bronquite” e “Infarto” na maior parte das divisões de dados.

Entretanto, destaca-se o *fold* 2, no qual o modelo classificou incorretamente todas as instâncias da classe “Infarto”, resultando em acurácia de 0% para essa classe e uma queda expressiva na acurácia geral. Esse comportamento sugere que, nessa partição específica, o conjunto de teste possuía características menos representativas ou particularmente desafiadoras para o modelo, o que resultou em um desequilíbrio no desempenho entre as classes.

Apesar desse caso isolado, a média das acurácias permanece elevada para ambas as classes, com 98,00% para “Bronquite” e 90,00% para “Infarto”, valores que estão alinhados com os resultados gerais obtidos nas 20 avaliações anteriores. O desvio padrão mais elevado para a classe “Infarto” (30,00%) reflete justamente essa oscilação pontual, mas

não compromete a evidência de que o modelo, no geral, lida bem com a distinção entre os dois sinais.

A Tabela 13 apresenta os critérios de precisão, *recall*, F_1 score e F_2 score referentes ao conjunto 20, que obteve a maior acurácia geral no problema de classificação entre os sinais “Bronquite” e “Infarto”.

Tabela 13 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 20) do problema de classificação dos sinais “Bronquite” e “Infarto”.

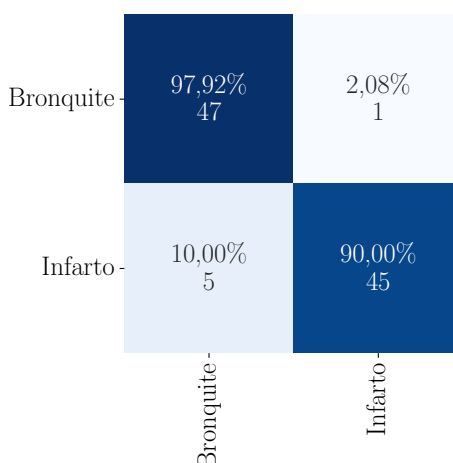
Classe	Precisão (%)	Recall (%)	F_1 score (%)	F_2 score (%)
Bronquite	90,38	97,92	94,00	96,31
Infarto	97,83	90,00	93,75	91,46

Fonte: Elaborada pelo autor (2026).

Por meio dos resultados apresentados na Tabela 13, observa-se que ambas as classes apresentam resultados equilibrados, indicando que o modelo conseguiu discriminar de forma satisfatória os sinais de “Bronquite” e “Infarto”. A classe “Bronquite” obteve maior *recall* (97,92%), sugerindo que o modelo raramente deixou de identificá-la quando presente, enquanto a classe “Infarto” apresentou maior precisão (97,83%), indicando menor ocorrência de falsos positivos. Os valores de F_1 score e F_2 score reforçam essa tendência, evidenciando um desempenho globalmente consistente, ainda que com pequenas assimetrias entre as classes.

Complementarmente, a Figura 16 exibe a matriz de confusão correspondente, permitindo uma visualização direta dos acertos e erros realizados pelo modelo.

Figura 16 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 20) para o problema de classificação dos sinais “Bronquite” e “Infarto”.



Fonte: Elaborada pelo autor (2026).

Tendo em vista esses resultados, observa-se que as classes “Bronquite” e “Infarto” apresentaram redução nos erros de classificação. Em particular, no modelo de 10 classes, a classe “Infarto” apresentava confusão principalmente com a classe “Bronquite”, conforme ilustrado na Figura 14. No entanto, a matriz de confusão apresentada na Figura 16 indica que esses erros foram significativamente reduzidos, evidenciando que o modelo especialista conseguiu separar de forma mais adequada os padrões característicos de cada classe. Assim, confirma-se que a estratégia de isolamento adotada contribuiu diretamente para reduzir as ambiguidades presentes no modelo geral para este caso.

Agora, para o modelo que classifica os sinais entre “Câimbra” e “Fratura”, o experimento que obteve melhor resultado utilizou um tamanho de lote igual a 64, taxa de aprendizado igual a 1×10^{-4} , número de neurônios na camada LSTM igual a 32, taxa do *dropout* recorrente igual a 0,2 e um número de épocas igual a 150.

A Tabela 14 apresenta a acurácia geral e por classe obtidas em cada uma das 20 avaliações realizadas para os conjuntos de teste da validação cruzada.

Tabela 14 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada para o problema de classificação dos sinais “Câimbra” e “Fratura”.

Avaliação	Câimbra	Fratura	Geral
1	89,58	94,00	91,84
2	83,33	98,00	90,82
3	83,33	94,00	88,78
4	85,42	94,00	89,80
5	81,25	96,00	88,78
6	85,42	94,00	89,80
7	81,25	96,00	88,78
8	87,50	94,00	90,82
9	93,75	94,00	93,88
10	75,00	94,00	84,69
11	89,58	96,00	92,86
12	89,58	98,00	93,88
13	89,58	98,00	93,88
14	89,58	94,00	91,84
15	83,33	94,00	89,80
16	89,58	94,00	91,84
17	85,42	96,00	90,82
18	83,33	96,00	89,80
19	91,67	94,00	92,86
20	87,50	96,00	91,84
Média	86,25	95,30	90,87
Desvio padrão	4,29	1,45	2,18

Fonte: Elaborada pelo autor (2026).

Tendo em vista esses valores, é possível ver que o modelo conseguiu apresentar

boa performance, com um valor médio de acurácia de 90,87% e um desvio padrão de 2,18%, o que indica um comportamento estável ao longo das avaliações. Observa-se ainda que a classe “Fratura” manteve consistentemente acurácias mais altas, enquanto a classe “Câimbra” apresentou maior variabilidade. Apesar disso, ambas as classes alcançaram resultados consideráveis, reforçando a capacidade do modelo em distingui-las de forma confiável.

Novamente, vários conjuntos atingiram o maior valor de acurácia geral, sendo este de 93,88%. Com isso, o conjunto escolhido para construir a Tabela 15, que apresenta os resultados de acurácia obtidos para cada classe em cada *fold*, foi o conjunto 13.

Tabela 15 – Resultados da acurácia geral e por classe, em porcentagem, para cada *fold* da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 13) do problema de classificação dos sinais “Câimbra” e “Fratura”.

<i>Fold</i>	Câimbra	Fratura	Geral
<i>Fold 1</i>	100,00	100,00	100,00
<i>Fold 2</i>	100,00	100,00	100,00
<i>Fold 3</i>	80,00	100,00	100,00
<i>Fold 4</i>	100,00	80,00	90,00
<i>Fold 5</i>	75,00	100,00	88,89
<i>Fold 6</i>	100,00	100,00	100,00
<i>Fold 7</i>	100,00	100,00	100,00
<i>Fold 8</i>	100,00	100,00	100,00
<i>Fold 9</i>	40,00	100,00	70,00
<i>Fold 10</i>	100,00	100,00	100,00
Média	89,50	98,00	94,89
Desvio padrão	18,77	6,00	9,28

Fonte: Elaborada pelo autor (2026).

Diante desses resultados, é notável que foi possível proporcionar uma boa diminuição nas confusões entre as classes, principalmente para a classe “Câimbra”. Observa-se que, apesar de alguns *folds* apresentarem acurácia mais baixa para essa classe — como é o caso do *fold 9*, com 40,00% — o desempenho geral permanece elevado, refletido pela média de 89,50%. Já a classe “Fratura” manteve um comportamento ainda mais estável, com acurácia média de 98,00% e variabilidade reduzida ao longo dos *folds*.

Para verificar de maneira mais detalhada, a Tabela 16 apresenta os critérios de precisão, *recall*, F_1 score e F_2 score gerais, referentes ao conjunto que apresentou maior acurácia geral — o conjunto 9.

Tabela 16 – Avaliação de desempenho por classe para o conjunto que obteve melhor acurácia geral (conjunto 13) do problema de classificação dos sinais “Câimbra” e “Fratura”.

Classe	Precisão (%)	Recall (%)	F_1 score (%)	F_2 score (%)
Câimbra	97,73	89,58	93,48	91,10
Fratura	90,74	98,00	94,23	96,46

Fonte: Elaborada pelo autor (2026).

A partir dos valores apresentados na Tabela 16, observa-se que ambas as classes apresentam desempenho elevado e relativamente equilibrado na discriminação entre “Câimbra” e “Fratura”. A classe “Câimbra” apresenta alta precisão (97,73%), indicando baixa taxa de falsos positivos, embora com um valor de *recall* ligeiramente inferior (89,58%). Por outro lado, a classe “Fratura” apresenta *recall* elevado (98,00%), evidenciando boa capacidade de identificação das instâncias verdadeiras dessa classe, ainda que com uma precisão um pouco menor (90,74%).

Os valores de F_1 score e F_2 score, igualmente elevados e praticamente idênticos para ambas as categorias, reforçam essa consistência, demonstrando que o modelo manteve bom equilíbrio entre sensibilidade e precisão mesmo considerando diferentes pesos atribuídos ao *recall*. Esses resultados confirmam que o processo de separação posterior das classes originalmente agrupadas foi eficaz, minimizando as confusões que se manifestavam no cenário de 10 classes e garantindo uma classificação mais confiável para esse par específico.

Além disso, a Figura 17 apresenta a matriz de confusão correspondente, de modo a complementar a análise dos resultados e permitir uma visualização direta dos acertos e erros realizados pelo modelo.

Figura 17 – Matriz de confusão geral da validação cruzada do conjunto que obteve melhor acurácia geral (conjunto 13) para o problema de classificação dos sinais “Câimbra” e “Fratura”.

Câimbra	89,58% 43	10,42% 5
Fratura	2,00% 1	98,00% 49
	Câimbra	Fratura

Fonte: Elaborada pelo autor (2026).

Tendo em vista os resultados apresentados pela matriz de confusão da Figura 17, observa-se uma melhoria substancial na capacidade de distinção entre as classes quando comparada ao desempenho obtido no modelo com 10 classes. O índice de erro, que anteriormente era de 22,92%, reduz-se para apenas 10,42% neste novo modelo, indicando que a especialização permitiu ao classificador separar com maior clareza os padrões característicos de cada sinal. Esses resultados reforçam, mais uma vez, a eficácia da abordagem baseada em modelos especialistas, que se mostra especialmente vantajosa em cenários onde determinadas classes apresentam similaridades significativas.

Com esses dois problemas de classes unificadas resolvidos, foi desenvolvido o último modelo que classifica os sinais entre as classes “Alergia” e “Frieira”. Para este caso, percebeu-se que a diferença entre os sinais se dá, justamente, no início do movimento. Desse modo, foi desenvolvida uma estratégia que realiza a identificação do início do movimento em uma execução de sinal.

A detecção do início do movimento foi realizada por meio de uma análise temporal dos marcos das mãos. Inicialmente, para cada quadro do vídeo, foram carregados os marcos produzidos pelo modelo de detecção, considerando-se apenas os quadros válidos, isto é, aqueles em que os pontos das mãos não estão totalmente zerados. Para cada quadro válido, foram extraídos os marcos correspondentes às regiões das mãos esquerda e direita. Em seguida, calculou-se o centro de cada mão a partir da média das coordenadas (x, y) dos seus respectivos marcos. Essa etapa permite representar o movimento das mãos ao longo do tempo de forma compacta e menos sensível a variações estruturais individuais.

Após essa etapa, as sequências referentes à componente vertical (eixo y) dos centros

das mãos foram suavizados utilizando o filtro de Savitzky–Golay, com o mesmo princípio de Konstantinovskiy (2024), com janela de tamanho 5 e polinômio de ordem 3. Esse procedimento reduz ruídos quadro a quadro, facilitando a identificação de tendências reais de movimento. A partir das sequências suavizadas, foram calculadas as velocidades aproximadas das mãos por meio do operador de gradiente, e a energia do movimento em cada quadro foi definida como a soma dos módulos dessas velocidades, refletindo aumentos simultâneos de atividade das duas mãos.

Para determinar o início do gesto, adotou-se um limiar adaptativo baseado no comportamento inicial da energia. Primeiro, consideram-se os 5 primeiros quadros como fase de repouso, e calcula-se um limiar dado pela soma entre a média e o desvio padrão da energia nessa janela. O início do movimento é então identificado como o primeiro quadro em que a energia permanece acima desse limiar durante 5 quadros consecutivos, garantindo que oscilações isoladas não sejam interpretadas como movimento real.

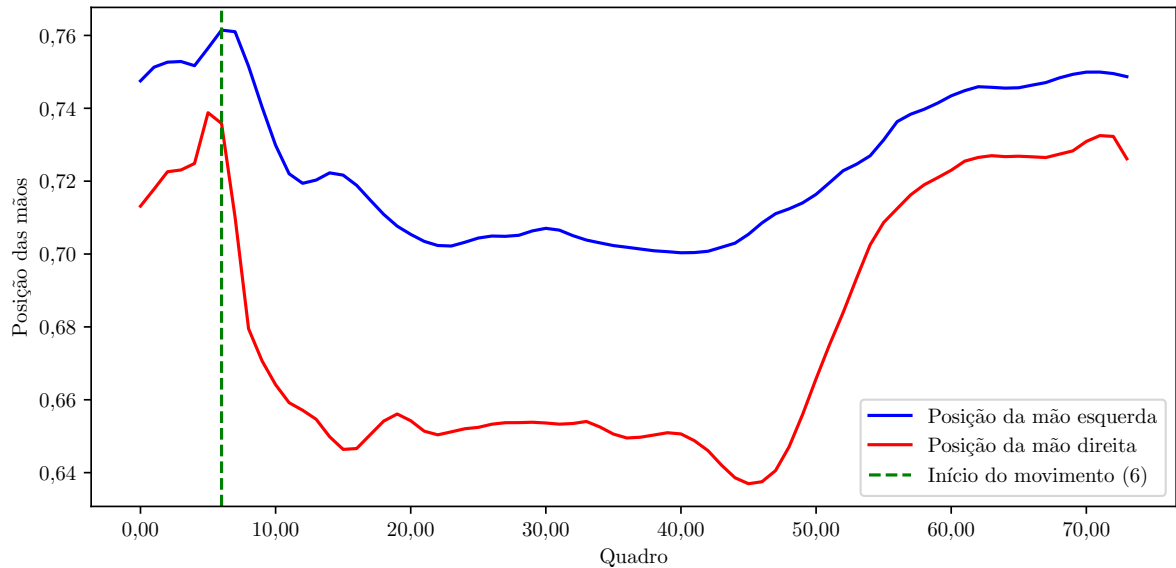
Caso o índice detectado ocorra muito cedo, ou seja, dentro da própria janela inicial, uma segunda estimativa do limiar é realizada. Nessa etapa, utiliza-se uma janela de referência compreendida entre os quadros 10 e 20 para recalcular o limiar, agora definido como a média acrescida de duas vezes o desvio padrão da energia. Esse critério mais restritivo não pressupõe necessariamente a ausência de movimento nessa região, mas visa elevar o limiar de decisão, tornando a detecção mais robusta a ruídos e a pequenas variações iniciais. O algoritmo então busca novamente a ocorrência de 5 quadros consecutivos acima desse novo limiar. Esse mecanismo de correção permite lidar com vídeos em que a fase inicial é ruidosa ou atípica, assegurando que o início do movimento seja marcado apenas quando ocorre uma elevação consistente e sustentada da energia do movimento.

A partir do instante detectado, são selecionados 30 quadros que compõem a entrada utilizada na etapa de classificação. Esses quadros são amostrados de forma uniformemente espaçada, com espaçamento aproximado de $s = \lceil f/30 \rceil$ quadros, onde f é a taxa de aquisição do vídeo em FPS. Assim, em um vídeo de 60 FPS, por exemplo, o espaçamento resultante é $s = 2$ e, portanto, os quadros são amostrados em intervalos regulares (de 2 em 2), garantindo que o segmento analisado mantenha duração temporal compatível com o comportamento do sinal observado. Desse modo, pode-se perceber que não seria possível fazer o processo de aumento de dados utilizando o sorteio de quadros por meio da distribuição normal. Dessa maneira, cada execução foi replicada 20 vezes, porém apenas com as transformações descritas na Seção 5.1.2.

Assim, como para cada vídeo tem-se 20 execuções, o processo de identificação do início do movimento é aplicado apenas para o primeiro e a mesma seleção temporal é replicada para suas variações, assegurando consistência entre todas as amostras. As Figuras 18 e 19 apresentam, respectivamente, o gráfico da posição das mãos durante a execução do movimento e o gráfico da energia calculada pelo movimento das mãos ao

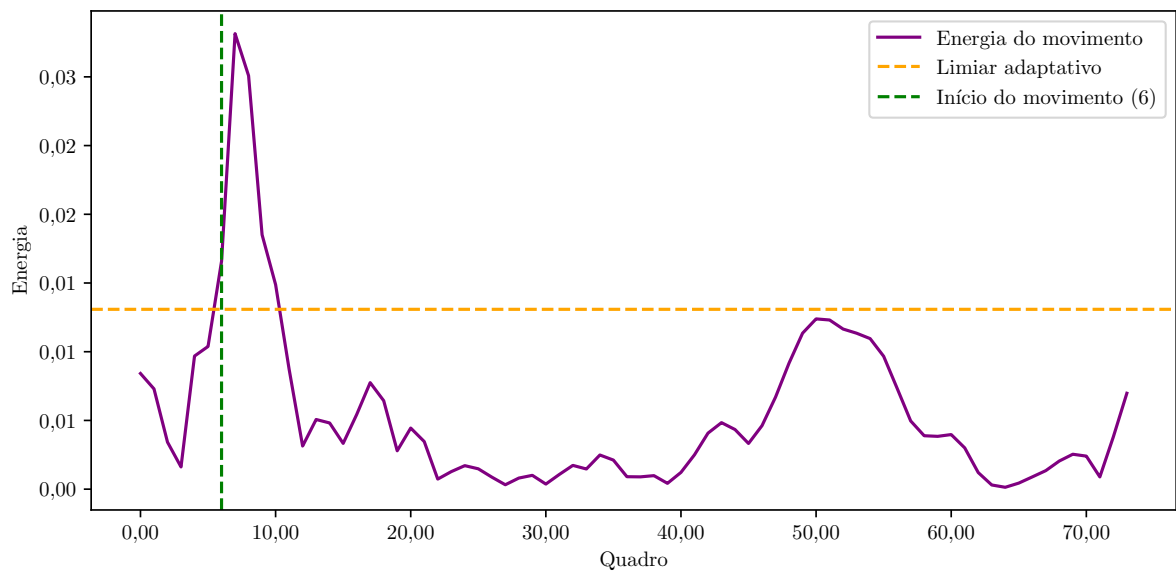
longo da execução para uma das amostras da classe “Alergia”.

Figura 18 – Gráfico das posições das mãos ao longo da execução do sinal com identificação do início do movimento.



Fonte: Elaborada pelo autor (2026).

Figura 19 – Gráfico da energia do movimento calculada ao longo da execução do movimento com início do movimento e limiar identificados.



Fonte: Elaborada pelo autor (2026).

Com o conjunto de dados para este caso definido, experimentos foram realizados e

o que obteve melhor resultado utilizou um tamanho de lote igual a 32, taxa de aprendizado igual a 0,001, número de neurônios na camada LSTM igual a 128, taxa do *dropout* recorrente igual a 0,1 e um número de épocas igual a 250.

Considerando o desempenho global do modelo a partir da validação cruzada, a classe “Alergia” apresentou uma acurácia de 98,00% , enquanto a classe “Frieira” obteve acurácia de 96,00% . Para o desempenho geral do modelo, a acurácia foi de 97,00%.

A Tabela 17 apresenta os resultados obtidos de acurácia para cada classe em cada *fold* para o conjunto de teste proveniente da validação cruzada. Diferente dos casos anteriores, aqui, não há mais de um conjunto, pois não foi realizado o sorteio dos quadros com distribuição normal, conforme explicado anteriormente.

Tabela 17 – Resultados da acurácia, em porcentagem, e estatísticas por classe para cada *fold* da validação cruzada para o problema de classificação dos sinais “Alergia” e “Frieira”.

<i>Fold</i>	Alergia	Frieira	Geral
<i>Fold 1</i>	100,00	80,00	90,00
<i>Fold 2</i>	100,00	80,00	90,00
<i>Fold 3</i>	100,00	100,00	100,00
<i>Fold 4</i>	100,00	100,00	100,00
<i>Fold 5</i>	100,00	100,00	100,00
<i>Fold 6</i>	100,00	100,00	100,00
<i>Fold 7</i>	100,00	100,00	100,00
<i>Fold 8</i>	80,00	100,00	90,00
<i>Fold 9</i>	100,00	100,00	100,00
<i>Fold 10</i>	100,00	100,00	100,00
Média	98,00	96,00	97,00
Desvio padrão	6,00	8,00	4,58

Fonte: Elaborada pelo autor (2026).

A Tabela 17 apresenta um desempenho bastante consistente ao longo dos 10 *folds*, com acurácias gerais variando entre 90% e 100%. Observa-se que, diferentemente do comportamento verificado no modelo de 10 classes, ambas as classes passaram a ser distinguidas de maneira clara e estável. A classe “Alergia” apresentou média de acerto de 98%, enquanto “Frieira” atingiu 96%, ambas com desvios padrão 6% e 8%, respectivamente, indicando que o modelo mantém sua capacidade discriminativa mesmo sob diferentes particionamentos dos dados.

A Tabela 18 apresenta os critérios gerais de desempenho do modelo para o problema de classificação binária entre os sinais “Alergia” e “Frieira”, incluindo precisão, *recall*, F_1 score e F_2 score.

Tabela 18 – Avaliação de desempenho por classe para o problema de classificação dos sinais “Alergia” e “Frieira”.

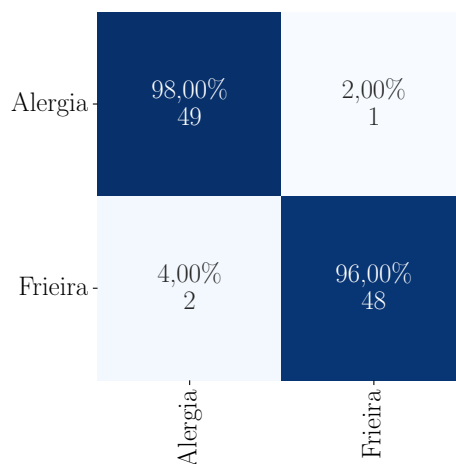
Classe	Precisão (%)	Recall (%)	F_1 score (%)	F_2 score (%)
Alergia	96,08	98,00	97,03	97,61
Frieira	97,96	96,00	96,97	96,39

Fonte: Elaborada pelo autor (2026).

Os resultados evidenciam um comportamento bastante equilibrado entre as duas categorias, com valores superiores a 96% em todos os critérios. A classe “Alergia” apresenta um *recall* ligeiramente superior (98%), indicando que o modelo raramente deixa de detectar ocorrências dessa classe. Por outro lado, “Frieira” mantém maior precisão (97,96%), sugerindo menor taxa de falsos positivos. A proximidade entre os valores de F_1 score e F_2 score em ambas as classes reforça a estabilidade do modelo e indica que não há assimetria relevante entre sensibilidade e precisão. Dessa forma, a capacidade do modelo especialista em separar adequadamente as duas classes é confirmada, superando as dificuldades observadas no cenário de classificação com 10 classes.

A Figura 20 apresenta a matriz de confusão geral obtida a partir dos conjuntos de teste da validação cruzada para o problema de classificação entre os sinais “Alergia” e “Frieira”.

Figura 20 – Matriz de confusão geral da validação cruzada para o problema de classificação dos sinais “Alergia” e “Frieira”.



Fonte: Elaborada pelo autor (2026).

Observa-se que o modelo apresenta poucos erros de classificação, mantendo elevado acerto para ambas as classes. A maioria absoluta das amostras é corretamente atribuída às suas respectivas categorias, o que confirma a boa separabilidade entre os padrões gestuais e reforça os resultados já evidenciados pela Tabela 18. Os poucos erros restantes

distribuem-se de forma equilibrada entre as classes, indicando que não há tendência de viés ou favorecimento de uma categoria específica.

Diante desses resultados, observa-se o impacto significativo da estratégia de classificação por pares de sinais, a qual reduziu de forma expressiva o número de confusões entre as classes. Com essa melhora evidenciada tanto pelos critérios quanto pela matriz de confusão, o próximo passo consiste em integrar esses modelos especialistas ao modelo geral de 7 classes, a fim de avaliar de maneira conjunta as melhorias alcançadas e verificar o quanto essa abordagem contribui para o desempenho final do modelo.

6.4.4 Integração do modelo de 7 classes com os modelos especialistas

Para avaliar o desempenho da abordagem hierárquica com modelos especialistas, foram realizadas as 20 execuções de validação cruzada sobre os conjuntos de teste. A Tabela 19 apresenta os resultados detalhados, incluindo a acurácia por classe em cada avaliação, bem como a média e o desvio padrão das 20 execuções, fornecendo uma visão completa da estabilidade e do desempenho do modelo.

Tabela 19 – Resultados da acurácia geral e por classe, em porcentagem, para os conjuntos de teste da validação cruzada utilizando estratégia de modelos especialistas.

Avaliação	Alergia	Autismo	Bronquite	Câimbra	Convulsão	Diarreia	Fratura	Frieira	Infarto	Sarampo	Geral
1	88,00	100,00	97,92	85,42	88,00	95,92	94,00	88,00	88,00	100,00	92,48
2	86,00	100,00	97,92	81,25	84,00	97,96	98,00	86,00	88,00	100,00	91,87
3	88,00	100,00	95,83	81,25	88,00	97,96	94,00	84,00	90,00	100,00	91,87
4	90,00	100,00	95,83	83,33	88,00	97,96	94,00	84,00	90,00	100,00	92,28
5	90,00	100,00	95,83	79,17	88,00	100,00	92,00	90,00	88,00	100,00	92,28
6	88,00	100,00	95,83	79,17	88,00	97,96	94,00	90,00	88,00	100,00	92,07
7	86,00	100,00	93,75	75,00	90,00	95,92	96,00	88,00	90,00	97,87	91,26
8	88,00	100,00	97,92	85,42	88,00	95,92	94,00	90,00	90,00	100,00	92,89
9	88,00	100,00	95,83	89,58	90,00	93,88	94,00	86,00	88,00	100,00	92,48
10	88,00	100,00	95,83	68,75	88,00	95,92	94,00	88,00	88,00	97,87	90,45
11	86,00	100,00	95,83	85,42	90,00	93,88	96,00	92,00	90,00	100,00	92,89
12	92,00	100,00	97,92	87,50	90,00	97,96	98,00	84,00	88,00	100,00	93,50
13	90,00	100,00	93,75	85,42	88,00	97,96	98,00	90,00	90,00	97,87	93,09
14	88,00	100,00	100,00	85,42	88,00	97,96	94,00	94,00	88,00	100,00	93,50
15	86,00	100,00	95,83	77,08	88,00	97,96	96,00	92,00	88,00	95,74	91,67
16	90,00	100,00	95,83	87,50	88,00	93,88	94,00	88,00	90,00	100,00	92,68
17	90,00	100,00	95,83	79,17	88,00	93,88	96,00	90,00	90,00	100,00	92,28
18	88,00	100,00	97,92	81,25	90,00	97,96	96,00	88,00	88,00	100,00	92,68
19	88,00	100,00	95,83	89,58	88,00	95,92	94,00	90,00	86,00	100,00	92,68
20	92,00	100,00	95,83	87,50	90,00	97,96	96,00	90,00	90,00	100,00	93,90
Média	88,50	100,00	96,35	82,71	88,40	96,73	95,10	88,60	88,80	99,47	92,44
Desvio padrão	1,77	0,00	1,45	5,15	1,36	1,76	1,61	2,69	1,17	1,14	0,78

Fonte: Elaborada pelo autor (2026).

Ao comparar os resultados obtidos na Tabela 19 com os resultados obtidos na Tabela 5, observa-se uma melhora significativa tanto na acurácia média geral quanto na consistência das predições.

No caso do primeiro modelo de 10 classes, a média de acurácia geral foi de 83,83%, enquanto o modelo hierárquico alcançou 92,44% de média, indicando maior estabilidade entre as 20 execuções. Além disso, a acurácia por classe apresenta ganhos notáveis, especialmente para sinais que eram mais confusos no problema original, como “Alergia”,

que tinha uma média de acurácia de 73,30% e agora apresenta uma média de acurácia de 88,50%, e “Frieira” que tinha uma média de acurácia de 55,90% e agora apresenta uma média de acurácia de 88,60%.

Esses resultados evidenciam que a decomposição hierárquica do problema e o uso de classificadores especialistas para pares de sinais conflitantes permitem reduzir erros específicos, melhorar o desempenho global e garantir maior consistência nas predições, confirmando a efetividade da abordagem proposta.

Utilizando o mesmo conjunto de dados empregado na avaliação do problema de 10 classes (conjunto 14), de modo a fazer uma comparação mais justa, a Tabela 20 apresenta os resultados de acurácia por classe em cada *fold*, assim como a média e o desvio padrão correspondentes. Dessa forma, é possível avaliar não apenas o desempenho médio, mas também a consistência do sistema ao longo da validação cruzada.

Tabela 20 – Resultados da acurácia geral e por classe, em porcentagem para cada *fold* da validação cruzada utilizando estratégia de modelos especialistas.

<i>Fold</i>	Alergia	Autismo	Bronquite	Câimbra	Convulsão	Diarreia	Fratura	Frieira	Infarto	Sarampo	Geral
<i>Fold 1</i>	100,00	100,00	100,00	100,00	100,00	100,00	80,00	80,00	100,00	100,00	96,00
<i>Fold 2</i>	80,00	100,00	100,00	100,00	80,00	100,00	100,00	80,00	0,00	100,00	82,61
<i>Fold 3</i>	40,00	100,00	100,00	40,00	0,00	100,00	100,00	100,00	100,00	100,00	78,00
<i>Fold 4</i>	100,00	100,00	100,00	100,00	100,00	100,00	80,00	100,00	100,00	100,00	98,00
<i>Fold 5</i>	100,00	100,00	100,00	100,00	100,00	80,00	100,00	100,00	80,00	100,00	95,65
<i>Fold 6</i>	100,00	100,00	100,00	40,00	100,00	100,00	100,00	100,00	100,00	100,00	94,00
<i>Fold 7</i>	80,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	98,00
<i>Fold 8</i>	80,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	98,00
<i>Fold 9</i>	100,00	100,00	100,00	80,00	100,00	100,00	100,00	100,00	100,00	100,00	98,00
<i>Fold 10</i>	100,00	100,00	100,00	100,00	100,00	100,00	80,00	80,00	100,00	100,00	96,00
Média	88,00	100,00	100,00	86,00	88,00	98,00	94,00	94,00	88,00	100,00	93,43
Desvio padrão	18,33	0,00	0,00	23,75	29,93	6,00	9,17	9,17	29,93	0,00	6,76

Fonte: Elaborada pelo autor (2026).

Por meio da Tabela 20, observa-se uma melhora consistente nos resultados em relação à Tabela 6. A acurácia média geral aumenta de 84,79% para 93,43%, enquanto o desvio padrão cai de 10,58% para 6,76%, indicando um sistema não apenas mais preciso, mas também mais estável entre os diferentes *folds*.

Ao se comparar numericamente o desempenho classe a classe, os ganhos tornam-se ainda mais expressivos. A classe “Alergia”, por exemplo, aumenta sua média de acurácia de 76,00% para 88,00%, enquanto “Câimbra” evolui de 59,00% para 86,00%, um dos saltos mais significativos. A classe “Frieira”, que no modelo de 10 classes alcançava apenas 62,00% em média, passa a atingir 94,00% com o uso dos especialistas. De forma semelhante, “Infarto” apresenta um avanço de 84,00% para 88,00%, corrigindo incertezas observadas anteriormente. Mesmo classes que já possuíam desempenho satisfatório, como “Sarampo” e “Autismo”, apresentam melhorias (de 88,00% para 100,00% e de 92,00% para 100,00%, respectivamente), reforçando a qualidade do processo.

Esses ganhos mostram que os modelos especialistas atuam precisamente sobre as classes mais problemáticas, reduzindo os erros que anteriormente ocorriam de forma

recorrente. Além disso, nota-se que classes que já exibiam alta acurácia no modelo de 10 classes mantêm seus valores ou os ampliam, evidenciando que o refinamento não introduziu efeitos adversos sobre padrões bem estabelecidos.

Ademais, a Tabela 21 apresenta os critérios de desempenho obtidos após a aplicação da estratégia baseada em modelos especialistas. Essa tabela permite avaliar de forma quantitativa o impacto da abordagem proposta sobre cada classe individualmente, possibilitando uma comparação direta com os resultados do modelo original de 10 classes.

Tabela 21 – Avaliação de desempenho por classe utilizando estratégia de modelos especialistas.

Classe	Precisão (%)	Recall (%)	F_1 score (%)	F_2 score (%)
Alergia	84,62	88,00	86,27	87,30
Autismo	100,00	100,00	100,00	100,00
Bronquite	87,27	100,00	93,20	97,17
Câimbra	89,13	85,42	87,23	86,13
Convulsão	100,00	88,00	93,62	90,16
Diarreia	100,00	97,96	98,97	98,36
Fratura	90,38	94,00	92,16	93,25
Frieira	97,92	94,00	95,92	94,76
Infarto	100,00	88,00	93,62	90,16
Sarampo	88,68	100,00	94,00	97,51

Fonte: Elaborada pelo autor (2026).

Ao comparar a Tabela 21 com a Tabela 7, observa-se que a adoção dos modelos especialistas promoveu melhorias claras e quantificáveis nos critérios por classe. No caso de “Alergia”, por exemplo, a precisão aumentou de 74,51% para 84,62%, enquanto o *recall* passou de 76,00% para 88,00%. A classe “Câimbra”, anteriormente uma das mais problemáticas, apresentou um salto expressivo no *recall*, que evoluiu de 58,33% para 85,42%, e no F_1 score, que aumentou de 70,00% para 87,23%. Melhorias semelhantes são observadas em Frieira”, cujo F_1 score passou de 64,58% para 95,92%, e em “Fratura”, cujo desempenho elevou-se de 82,76% para 92,16%, reduzindo significativamente a taxa de erros.

Além disso, classes que já apresentavam desempenho elevado no modelo original, como “Diarreia” e “Autismo”, mantiveram ou ultrapassaram seus valores anteriores, indicando que a abordagem não prejudicou categorias bem definidas. De modo geral, os critérios de F_1 score e F_2 score tornaram-se mais homogêneos entre as classes, sugerindo um sistema mais equilibrado e com menor dependência de casos específicos.

A Figura 21 complementa a análise ao apresentar a matriz de confusão correspondente ao modelo após a utilização dos especialistas. Por meio dela, torna-se possível visualizar como a estratégia influenciou a distribuição dos acertos e erros entre as classes, oferecendo uma perspectiva mais clara sobre as confusões que foram reduzidas ou

eliminadas.

Figura 21 – Matriz de confusão geral da validação cruzada utilizando estratégia de modelos especialistas.

Alergia	88,00% 44	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	2,00% 1	0,00% 0	10,00% 5
Autismo	0,00% 0	100,00% 50	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Bronquite	0,00% 0	0,00% 0	100,00% 48	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Câimbra	4,17% 2	0,00% 0	0,00% 0	85,42% 41	0,00% 0	0,00% 0	10,42% 5	0,00% 0	0,00% 0	0,00% 0
Convulsão	12,00% 6	0,00% 0	0,00% 0	0,00% 0	88,00% 44	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Diarreia	0,00% 0	0,00% 0	0,00% 0	2,04% 1	0,00% 0	97,96% 48	0,00% 0	0,00% 0	0,00% 0	0,00% 0
Fratura	0,00% 0	0,00% 0	0,00% 0	6,00% 3	0,00% 0	0,00% 0	94,00% 47	0,00% 0	0,00% 0	0,00% 0
Frieira	0,00% 0	0,00% 0	2,00% 1	2,00% 1	0,00% 0	0,00% 0	0,00% 0	94,00% 47	0,00% 0	2,00% 1
Infarto	0,00% 0	0,00% 0	12,00% 6	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	88,00% 44	0,00% 0
Sarampo	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	0,00% 0	100,00% 47
	Alergia	Autismo	Bronquite	Câimbra	Convulsão	Diarreia	Fratura	Frieira	Infarto	Sarampo

Fonte: Elaborada pelo autor (2026).

A partir da matriz de confusão apresentada na Figura 21, é possível aprofundar a avaliação do impacto dos modelos especialistas ao compará-la diretamente com a matriz original da Figura 14. Enquanto o modelo de 10 classes apresentava confusões concentradas especialmente nos pares “Alergia”/“Frieira”, “Bronquite”/“Infarto” e “Câimbra”/“Fratura”, a matriz resultante após a aplicação dos especialistas mostra uma redução substancial desses erros. As células fora da diagonal principal, antes mais expressivas, tornam-se menos frequentes e com valores menores, evidenciando que os especialistas atuaram precisamente nas regiões de maior ambiguidade.

Em particular, as confusões entre “Alergia” e “Frieira”, que anteriormente al-

cançavam 28% nos conjuntos de teste, reduzem-se para apenas 2% após o tratamento especializado. De forma semelhante, os erros envolvendo “Câimbra” e “Fratura” tornam-se consideravelmente menos recorrentes, o que se reflete diretamente no aumento observado no *recall* e no F_1 score dessas classes na Tabela 21. Além disso, o erro total entre esse par de classes cai de 26,92% para 16,42%, evidenciando a vantagem da utilização dos modelos especialistas para resolver ambiguidades específicas. Por fim, as confusões entre “Bronquite” e “Infarto” também sofrem redução, contribuindo para a melhora nos acertos da classe “Infarto” e reforçando o impacto positivo da abordagem proposta.

Desse modo, percebe-se que a estratégia baseada na utilização de modelos especialistas promoveu uma melhoria significativa no desempenho do sistema de classificação. A redução da confusão entre classes ambíguas, aliada ao aumento das acurácias médias e à diminuição da variabilidade entre os *folds*, demonstra que o tratamento individualizado das classes com maior similaridade foi fundamental para aprimorar o desempenho do modelo. Esses resultados reforçam não apenas a eficácia da abordagem proposta, mas também sua capacidade de mitigar ambiguidades inerentes aos sinais que compõem o problema original de 10 classes.

6.4.5 Avaliação dos modelos no conjunto de dados de teste adicional

Tendo verificado o impacto positivo da integração entre o modelo de 7 classes e os modelos especialistas, torna-se relevante avaliar se esse ganho de desempenho se mantém quando o sistema é exposto a dados não utilizados em nenhuma etapa da validação cruzada. Dessa forma, nesta subseção são apresentados os resultados obtidos no conjunto de dados de teste adicional, permitindo analisar a capacidade de generalização do modelo em condições mais realistas.

Para verificar a questão da aleatoriedade no processo de seleção dos quadros, foram feitos 20 sorteios para cada instância, assim como para os testes de validação cruzada para o conjunto de dados original e cada sorteio foi avaliado utilizando a estratégia da integração do modelo de 7 classes com os modelos especialistas. Com isso, a Tabela 22 apresenta as acurácias gerais do conjunto e acurácias gerais de cada classe em cada sorteio, bem como a média e desvio padrão de todas essas acurácias.

Tabela 22 – Acurácia geral obtida em cada um dos 20 sorteios do conjunto de teste adicional.

Avaliação	Alergia	Autismo	Bronquite	Câmbra	Convulsão	Diarreia	Fraturo	Frieira	Infarto	Sarampo	Geral
1	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
2	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	83,33	100,00	96,67
3	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	83,33	100,00	96,67
4	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
5	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
6	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	100,00	98,33
7	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
10	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
11	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	100,00	100,00	98,33
12	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
13	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	100,00	98,33
14	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	100,00	98,33
15	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	100,00	98,33
16	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	100,00	98,33
17	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	83,33	100,00	96,67
18	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
19	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
20	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	83,33	100,00	98,33
Média	100,00	100,00	100,00	100,00	100,00	100,00	100,00	96,67	92,50	100,00	98,92
Desvio padrão	0,00	0,00	0,00	0,00	0,00	0,00	0,00	6,67	8,29	0,00	1,21

Fonte: Elaborada pelo autor (2026).

A Tabela 22 permite avaliar o desempenho do modelo frente a dados completamente externos ao processo de validação cruzada. Observa-se que, ao longo dos 20 sorteios realizados, a maior parte das classes manteve acurácia de 100,00%, evidenciando uma forte capacidade de generalização. As únicas variações ocorreram nas classes “Frieira” e “Infarto”, cujas médias foram de 96,67% e 92,50%, com desvios padrão de 6,67% e 8,29%, respectivamente. Ainda assim, a acurácia geral apresentou média de 98,92% e um desvio padrão reduzido de apenas 1,21%, indicando um comportamento altamente estável mesmo diante de diferentes composições do conjunto de dados de teste adicional.

Esses resultados reforçam o desempenho do modelo, especialmente ao considerar que diversos sorteios mantiveram 100,00% de acurácia geral. Como consequência direta desse desempenho, no melhor cenário, todos os critérios de precisão, *recall*, F_1 score e F_2 score alcançam seus valores máximos para todas as classes. Assim, diferentemente dos casos anteriores, tais critérios não são apresentados em tabelas, tampouco é exibida a matriz de confusão.

7 CONCLUSÃO

Neste trabalho, apresentou-se um método para classificar automaticamente dez sinais da Língua Brasileira de Sinais (Libras) relacionados à área da saúde. O estudo envolveu a construção de um conjunto de dados em parceria com alunos da disciplina Trabalho Multidisciplinar do Departamento de Mecânica Aplicada e Computacional da Universidade Federal de Juiz de Fora e o desenvolvimento de modelos de aprendizado de máquina para realizar essa tarefa. A criação desse conjunto de dados, composta por sequências padronizadas de movimentos representando sinais comuns em contextos clínicos, auxiliou na viabilização da modelagem de um problema real e relevante.

A partir do conjunto de dados, foram investigadas diferentes abordagens de classificação, iniciando pelo problema original de 10 classes. A análise das matrizes de confusão evidenciou confusões pontuais entre pares específicos de sinais, o que motivou a adoção de uma estratégia hierárquica baseada em modelos especialistas. Essa reorganização, que reduziu o problema para 7 classes e empregou classificadores dedicados aos pares conflituosos, resultou em melhorias nas acurácias e na estabilidade dos resultados, alcançando uma média de acurácias de 92,44% nos experimentos, reforçando a efetividade do método proposto.

Outro aspecto relevante refere-se às escolhas arquiteturais adotadas. A utilização de marcos tridimensionais extraídos pelo *MediaPipe Holistic* mostrou-se eficaz para representar os movimentos manuais e corporais, reduzindo a influência de fatores externos que afetam modelos baseados diretamente em imagens. Ao abstrair a cena para pontos articulares, o modelo passa a focar na dinâmica espacial dos sinais. Além disso, a inclusão de camadas LSTM permitiu capturar dependências temporais das sequências, contribuindo para o desempenho geral dos modelos.

Ademais, a avaliação em um conjunto de dados de teste adicional confirmou a boa capacidade de generalização dos modelos, com uma média de acurácias de 98,92% e baixo desvio padrão para os experimentos. Esses resultados mostram que o sistema mantém desempenho consistente mesmo diante de novas instâncias, reforçando seu potencial prático para aplicações de apoio à comunicação em ambientes de saúde.

Como trabalhos futuros, destacam-se a expansão do conjunto de sinais, a inclusão de usuários com diferentes mãos dominantes, a avaliação em cenários mais variados e a exploração de novas arquiteturas, como *Vision Transformers* (ViT) e modelos baseados em atenção.

REFERÊNCIAS

- ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <<https://www.tensorflow.org/>>.
- AL-MOHIMEED, B. A. et al. Dynamic sign language recognition based on real-time videos. *IJOE*, v. 18, n. 1, 2021. College of Computer, Qassim University, Buraydah, Saudi Arabia.
- BRASIL. *Lei nº 13.146, de 6 de julho de 2015*. 2015. Diário Oficial da União: seção 1, Brasília, DF, 7 jul. 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). Acesso em: 05 de Setembro 2025. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm>.
- CERNA, L. R. et al. A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. *Expert Systems with Applications*, v. 167, p. 114179, 2021.
- COSTA, C. F. F. et al. A fully automatic method for recognizing hand configurations of brazilian sign language. *Research on Biomedical Engineering*, SciELO Brasil, v. 33, p. 78–89, 2017.
- FENG, W. et al. Audio visual speech recognition with multimodal recurrent neural networks. In: . [S.l.: s.n.], 2017. p. 681–688.
- GALA, A. S. *Profissionais de tecnologia e acessibilidade: quais são eles?* 2023. Acesso em: 26 de Dezembro de 2025. Disponível em: <<https://www.handtalk.me/br/blog/profissionais-de-tecnologia-e-acessibilidade/>>.
- GANDHI, D.; SHAH, K.; CHANDANE, M. Dynamic sign language recognition and emotion detection using mediapipe and deep learning. In: IEEE. *13th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. Virtual Conference: IEEE, 2022. p. 1–6.
- GERS, F. A.; SCHMIDHUBER, J.; CUMMINS, F. Learning to forget: Continual prediction with lstm. *Neural Computation*, MIT Press, v. 12, n. 10, p. 2451–2471, 1999.
- GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. S.l.: Pearson, 2018.
- HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <<https://doi.org/10.1038/s41586-020-2649-2>>.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, MIT Press, 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*, MDPI, v. 10, n. 6, p. 149, 2024. Published: 20 June 2024.
- IBM. *O que é overfitting?* 2024. Acesso em: 26 de Novembro de 2025. Disponível em: <<https://www.ibm.com/br-pt/think/topics/overfitting>>.

KONSTANTINOVSKY, T. *Introduction to the Savitzky–Golay Filter: A Comprehensive Guide (Using Python)*. 2024. Acesso em: 22 de Outubro de 2025. Disponível em: <<https://medium.com/pythoneers/introduction-to-the-savitzky-golay-filter-a-comprehensive-guide-using-python-b2dd07a8e2ce>>.

MA, D. et al. Dynamic sign language recognition based on improved residual-lstm network. In: *The 7th International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2021)*. Beijing, China: [s.n.], 2021.

MediaPipe Face. *Guia de detecção de pontos de referência do rosto*. 2025. Acesso em: 08 de Agosto de 2025. Disponível em: <https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker?hl=pt-br>.

MediaPipe GitHub Holistic. *MediaPipe Holistic*. 2025. <<https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/holistic.md>>. Disponível em: <<https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/holistic.md>>. Acesso em: 08 de Agosto de 2025.

MediaPipe Hands. *Guia de detecção de pontos de referência manuais*. 2025. Acesso em: 08 de Agosto de 2025. Disponível em: <https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker?hl=pt-br>.

MediaPipe Pose. *Guia de detecção de pontos de referência de poses*. 2025. Acesso em: 08 de Agosto de 2025. Disponível em: <https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker?hl=pt-br#:~:text=The%20MediaPipe%20Pose%20Landmarker%20task,with%20single%20images%20or%20video.>>

MIENYE, I. D.; SWART, T. G.; OBAIDO, G. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information, MDPI*, v. 15, n. 9, p. 517, 2024. Disponível em: <<https://doi.org/10.3390/info15090517>>.

NEIDLE, C.; THANGALI, A.; SCLAROFF, S. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In: *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*. [S.l.]: Citeseer, 2012.

OLAH, C. *Understanding LSTM Networks*. 2015. Acesso em: 08 de Outubro de 2025. Disponível em: <<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>>.

OLIVEIRA, A. S. R. et al. Ensino da língua brasileira de sinais durante a graduação em medicina: a percepção dos futuros médicos. *Audiology - Communication Research*, v. 27, 2022. FapUNIFESP (SciELO).

QUADROS, R. M. d.; KARNOPP, L. B. *Língua de sinais brasileira: estudos linguísticos*. Porto Alegre: Artmed, 2007. ISBN 978-85-363-0308-6. Disponível em: <<https://bds.unb.br/handle/123456789/948>>.

REDMON, J. et al. You only look once: Unified, real-time object detection. In: . Las Vegas, Nevada, USA: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. p. 779–788.

REZENDE, T. M.; ALMEIDA, S. G. M.; GUIMARÃES, F. G. Development and validation of a brazilian sign language database for human gesture recognition. *Neural Computing and Applications*, v. 33, p. 10449–10467, 2021.

RIDWANG et al. Dynamic sign language recognition using mediapipe library and modified lstm method. *International Journal on Advanced Science Engineering Information Technology (IJASEIT)*, v. 13, n. 6, Dec 2023. ISSN 2088-5334. Corresponding author: ridwang@unismuh.ac.id.

ROCHA, J. et al. Towards a tool to translate brazilian sign language (libras) to brazilian portuguese and improve communication with deaf. In: *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. [S.l.]: IEEE, 2020. p. 1–4.

RONCHETTI, F. et al. *LSA64: An Argentinian Sign Language Dataset*. 2023. Instituto de Investigación en Informática LIDI, Facultad de Informática, Universidad Nacional de La Plata; Instituto Superior Politécnico José Antonio Echeverría.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, American Psychological Association, v. 65, n. 6, p. 386–408, 1958.

SAMAAN, G. H. et al. Mediapipe's landmarks with rnn for dynamic sign language recognition. *Electronics*, MDPI, v. 11, n. 19, p. 3228, 2022. Disponível em: <<https://www.mdpi.com/2079-9292/11/19/3228>>.

SARMENTO, A. H. A. *Classificação automática de vídeos com sinais da Libras utilizando deep learning*. Dissertação (Dissertação de Mestrado) — Centro Universitário SENAI CIMATEC, Salvador, BA, 2023. Instituto Senai de Inovação em Sistemas de Manufatura e Processamento a Laser.

SEMIENIUTA, S.; SEVERYN, A.; BARTH, E. *Recurrent Dropout without Memory Loss*. 2016. Disponível em: <<https://arxiv.org/abs/1603.05118>>.

STONE, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, [Royal Statistical Society, Oxford University Press], v. 36, n. 2, p. 111–147, 1974. ISSN 00359246. Disponível em: <<http://www.jstor.org/stable/2984809>>.

ZHANG, J. et al. Chinese sign language recognition with adaptive hmm. In: *2016 IEEE International Conference on Multimedia and Expo (ICME)*. [S.l.: s.n.], 2016. p. 1–6.