Universidade Estadual de Londrina

JOÃO PAULO MANASSÉS BERNARDI MONTEIRO

AVALIAÇÃO COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO DA MARCHA NA PARALISIA CEREBRAL:

UMA ANÁLISE EXPLORATÓRIA DO SISTEMA GCS-CP

Londrina
2026

JOÃO PAULO MANASSÉS BERNARDI MONTEIRO

AVALIAÇÃO COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO DA MARCHA NA PARALISIA CEREBRAL: UMA ANÁLISE EXPLORATÓRIA DO SISTEMA GCS-CP

Dissertação apresentada ao Programa de Pós-Graduação em Ciências da Reabilitação (Programa Associado entre Universidade Estadual de Londrina [UEL] e Universidade Pitágoras Unopar [UNOPAR]), como requisito parcial à obtenção do título de Mestre em Ciências da Reabilitação.

Orientador: Prof. Dr. Felipe Arruda Moura
Co-orientador: Prof. Dr. Alessandro Giurizatto Melanda

Londrina
2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

M775a Monteiro, João Paulo Manassés Bernardi.
Avaliação comparativa de modelos de aprendizado de máquina para a classificação da marcha na paralisia cerebral: uma análise exploratória do sistema GCS-CP / João Paulo Manassés Bernardi Monteiro. - Londrina, 2026.
157 f. : il.

Orientador: Felipe Arruda Moura.
Coorientador: Alessandro Giurizatto Melanda.
Dissertação (Mestrado em Ciências da Reabilitação) - Universidade Estadual de Londrina, Centro de Ciências da Saúde, Programa de Pós-Graduação em Ciências da Reabilitação, 2026.
Inclui bibliografia.

1. Paralisia cerebral - Tese. 2. Marcha - Tese. 3. Aprendizado de máquina - Tese. 4. Análise da marcha - Tese. I. Moura, Felipe Arruda. II. Melanda, Alessandro Giurizatto. III. Universidade Estadual de Londrina. Centro de Ciências da Saúde. Programa de Pós-Graduação em Ciências da Reabilitação. IV. Título.

CDU 61

JOÃO PAULO MANASSÉS BERNARDI MONTEIRO

AVALIAÇÃO COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO DA MARCHA NA PARALISIA CEREBRAL: UMA ANÁLISE EXPLORATÓRIA DO SISTEMA GCS-CP

Dissertação apresentada ao Programa de Pós-Graduação em Ciências da Reabilitação (Programa Associado entre Universidade Estadual de Londrina [UEL] e Universidade Pitágoras Unopar [UNOPAR]), como requisito parcial à obtenção do título de Mestre em Ciências da Reabilitação.

BANCA EXAMINADORA

Prof. Dr. Felipe Arruda Moura
Universidade Estadual de Londrina
(Orientador)

Prof. Dr. Alessandro Giurizatto Melanda
Universidade Estadual de Londrina
(Coorientador)

Profa. Dra. Adriane Mara de Souza Muniz
Escola de Educação Física do Exército

Prof. Dr. Paulo Roberto Garcia Lucareli
Universidade Nove de Julho – UNINOVE

Londrina, 10 de agosto de 2026.

Dedico este trabalho à minha esposa Janette: guerreira, forte e mãe exemplar. A base sobre a qual tudo se sustenta: nossa casa, nossa família, e também esta conquista.

AGRADECIMENTOS

A realização deste trabalho foi uma jornada longa e, muitas vezes, árdua, mais desafiadora do que eu imaginava quando a iniciei. Enfrentei dificuldades, momentos de dúvida e a tentação de desistir, mas escolhi persistir a cada dia. Bebi de conhecimentos que jamais imaginei explorar e me deparei com versões de mim mesmo que não conhecia. Concluo este percurso mais forte, mais crítico e mais maduro, saio profundamente diferente de quem entrou, e isso só foi possível graças ao apoio inestimável de muitas pessoas. Valeu cada passo.

Em primeiro lugar, expresso minha profunda gratidão ao Professor orientador Felipe Arruda Moura, não apenas pela constante orientação acadêmica e intelectual, mas sobretudo pela amizade construída nos encontros e conversas na sala de aula e no laboratório, pela paciência e por ter sido um grande apoio em momentos difíceis. Guardo grande respeito e admiração pelo profissional e pela pessoa que conheci.

Agradeço também ao Professor co-orientador Alessandro Giurizatto Melanda, que acreditou em mim antes mesmo de eu acreditar que seria capaz de realizar este trabalho. Suas valiosas sugestões e perspicazes questionamentos enriqueceram significativamente este estudo.

Agradeço minha mãe Elisabete, que sempre lutou pelas minhas conquistas, tornando-as nossas.

MONTEIRO, JOÃO. **AVALIAÇÃO COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA PARA A CLASSIFICAÇÃO DA MARCHA NA PARALISIA CEREBRAL: UMA ANÁLISE EXPLORATÓRIA DO SISTEMA GCS-CP.** 2026. Dissertação (Mestrado em Ciências da Reabilitação) – Universidade Estadual de Londrina, Londrina, 2026.

RESUMO

Introdução: A paralisia cerebral (PC) é a deficiência física mais comum com início na infância, caracterizada por distúrbios permanentes do movimento e da postura. As alterações da marcha são frequentes e funcionalmente relevantes, porém sua classificação clínica permanece desafiadora devido à subjetividade das avaliações visuais. **Objetivo:** Avaliar técnicas de aprendizado de máquina para a classificação automática de padrões da marcha em crianças com PC, com base no Sistema de Classificação da Marcha para Crianças com Paralisia Cerebral (GCS-CP), enfatizando a análise do desempenho por classe como critério de aplicabilidade clínica. **Métodos:** Estudo transversal retrospectivo com 115 crianças e adolescentes com PC espástica (5 a 19 anos; GMFCS I–III), utilizando registros de análise cinemática tridimensional da marcha com divisão rigorosa por paciente entre conjuntos de treino e teste. O GCS-CP classifica padrões em três componentes: Apoio e Balanço (plano sagital) e Transverso (plano transverso), definidos com base em ângulos articulares específicos do quadril, joelho e tornozelo durante o ciclo da marcha. Foram avaliados nove algoritmos de aprendizado de máquina (Regressão Logística, Máquina de Vetores de Suporte, Árvores de Decisão, K-NN, *Naive Bayes*, Floresta Aleatória, LightGBM, XGBoost e Redes Neurais Artificiais) e três modelos de fusão (*Voting Hard*, *Voting Soft* e *Stacking*), comparando dois *pipelines* de variáveis (Cinemática Angular e Cinemática Geral, este último adicionando parâmetros espaço-temporais), duas estratégias de redução de dimensionalidade (com e sem seleção de variáveis por consenso, neste caso substituída por análise de componentes principais, PCA) e duas estratégias de balanceamento (com e sem SMOTE). Adicionalmente, foram avaliadas abordagens hierárquicas em cascata e três estratégias de otimização clínica orientadas à minimização de falsos negativos. **Resultados:** Os melhores modelos alcançaram Acurácia Balanceada de 0,791 para o Transverso (Regressão Logística, *Pipeline* Cinemática Geral, PCA), 0,717 para o Apoio (*Naive Bayes, Pipeline* Cinemática Angular, com consenso) e 0,702 para o Balanço (Árvore de Decisão, abordagem Hierárquica), atingindo o limiar de 0,70 adotado neste estudo. A análise comparativa entre estratégias metodológicas revelou padrão alvo-dependente sistemático em cinco eixos: seleção por consenso × PCA, classificação *Flat* × Hierárquica, com × sem SMOTE, *pipelines* de variáveis, e algoritmos individuais × modelos de fusão, sem superioridade universal de nenhuma escolha em nenhum eixo. Como exemplo, PCA foi superior para Transverso e para Balanço em reexecuções com sementes aleatórias distintas (+5,5 pontos percentuais), enquanto a seleção por consenso foi claramente superior para Apoio (+8,9 pontos percentuais). A análise por classe revelou heterogeneidade clinicamente relevante, com sensibilidade de 100% para Transverso Externo e apenas 56,6% para marcha em salto e 50,6% para marcha rígida nos melhores modelos antes da otimização clínica. As estratégias de otimização clínica demonstraram que sensibilidades próximas ou superiores a 90% são alcançáveis para quatro das cinco classes patológicas avaliadas, com eliminação completa dos falsos negativos para os padrões Sagital Agachado e Transverso Externo. **Conclusão:** Técnicas de aprendizado de máquina demonstram potencial para auxiliar a classificação objetiva de padrões da marcha na PC, com desempenho compatível com o teto

de confiabilidade interavaliadores humanos. Métricas agregadas mostraram-se insuficientes para avaliação clínica completa, sendo a análise por classe e as estratégias de otimização orientadas à sensibilidade patológica pré-requisitos para implementação clínica responsável.

Palavras-chave: Paralisia cerebral; Marcha; Aprendizado de máquina; Análise da marcha; Classificação automática.

MONTEIRO, JOÃO. **Comparative evaluation of machine learning models for gait classification in cerebral palsy: an exploratory analysis of the GCS-CP system**. 2026. Dissertation (Master's in Rehabilitation Sciences) – State University of Londrina, Londrina, 2026.

ABSTRACT

Introduction: Cerebral palsy (CP) is the most common physical disability with childhood onset, characterized by permanent disorders of movement and posture. Gait abnormalities are frequent and functionally relevant; however, their clinical classification remains challenging due to the subjectivity of visual assessments. **Objective:** To evaluate machine learning techniques for the automatic classification of gait patterns in children with CP, based on the Gait Classification System for Children with Cerebral Palsy (GCS-CP), emphasizing per-class performance analysis as a criterion for clinical applicability. **Methods:** Retrospective cross-sectional study with 115 children and adolescents with spastic CP (aged 5 to 19 years; GMFCS I–III), using three-dimensional kinematic gait analysis records with rigorous patient-level split between training and test sets. The GCS-CP classifies patterns into three components: Stance and Swing (sagittal plane) and Transverse (transverse plane), defined based on specific joint angles of the hip, knee and ankle during the gait cycle. Nine machine learning algorithms (Logistic Regression, SVM, Decision Tree, K-NN, Naive Bayes, Random Forest, LightGBM, XGBoost and Artificial Neural Networks) and three ensemble models (Voting Hard, Voting Soft and Stacking) were evaluated, comparing two variable pipelines (Angular Kinematics and General Kinematics, the latter adding spatiotemporal parameters), two dimensionality reduction strategies (with and without consensus-based feature selection, the latter replaced by principal component analysis, PCA) and two balancing strategies (with and without SMOTE). Hierarchical cascade approaches and three clinical optimization strategies aimed at minimizing false negatives were additionally evaluated. **Results:** The best models achieved Balanced Accuracy of 0.791 for the Transverse pattern (Logistic Regression, General Kinematics Pipeline, PCA), 0.717 for the Stance pattern (Naive Bayes, Angular Kinematics Pipeline, consensus), and 0.702 for the Swing pattern (Decision Tree, Hierarchical approach), reaching the 0.70 threshold adopted in this study. Comparative analysis across methodological strategies revealed a systematic target-dependent pattern in five dimensions: consensus-based selection × PCA, *Flat* × Hierarchical classification, with × without SMOTE, variable pipelines, and individual algorithms × *ensemble* models, with no universal superiority of any choice on any axis. As an example, PCA was superior for Transverse and for Swing in re-executions with distinct random seeds (+5.5 percentage points), while consensus-based selection was clearly superior for Stance (+8.9 percentage points). Per-class analysis revealed clinically relevant heterogeneity, with 100% sensitivity for external rotation but only 56.6% for jump gait and 50.6% for stiff gait in the best models before clinical optimization. Clinical optimization strategies demonstrated that sensitivities near or above 90% are achievable for four of the five pathological classes evaluated, with complete elimination of false negatives for the Sagittal Crouch and External Rotation patterns. **Conclusion:** Machine learning techniques demonstrate potential to support objective gait pattern classification in CP, with performance compatible with the ceiling of human inter-rater reliability. Aggregate metrics proved insufficient for complete clinical evaluation; per-class analysis and sensitivity-oriented optimization strategies are prerequisites for responsible clinical implementation.

Keywords: Cerebral palsy; Gait; Machine learning; Gait analysis; Automatic classification.

LISTA DE ILUSTRAÇÕES

Figura 1 – Regiões cerebrais afetadas na PC espástica, discinética e atáxica..... 26
Figura 2 – O ciclo da marcha e suas fases em relação à PC espástica..... 30
Figura 3 – Árvore abrangente de parâmetros na análise da marcha..... 32
Figura 4 – Classificação hierárquica do plano sagital..... 37
Figura 5 – Classificação hierárquica do plano transverso..... 37
Figura 6 – Dados cinemáticos do plano sagital..... 39
Figura 7 – Taxonomia das aplicações atuais e futuras de inteligência..... 40
Figura 8 – Avaliação da qualidade metodológica 41
Figura 9 – Avaliação da qualidade em estudos de machine learning..... 42
Figura 10 – Características cinemáticas extraídas do ciclo da marcha..... 44
Figura 11 – Arquitetura básica de uma Rede Neural Artificial..... 47
Figura 12 – Arquitetura dos modelos de fusão..... 50
Figura 13 – Fluxograma de seleção dos participantes e divisão dos dados..... 54
Figura 14 – Fluxograma do *pipeline* de aprendizado de máquina..... 58
Figura 15 – Seleção por Consenso: *Pipeline* Cinemática Angular | Apoio Geral..... 82
Figura 16 – Seleção por Consenso: *Pipeline* Cinemática Angular | Balanço Geral..... 82
Figura 17 – Seleção por Consenso: *Pipeline* Cinemática Angular | Transverso Geral..... 83
Figura 18 – Seleção por Consenso: *Pipeline* Cinemática Geral | Apoio Geral..... 83
Figura 19 – Seleção por Consenso: *Pipeline* Cinemática Geral | Balanço Geral..... 84
Figura 20 – Seleção por Consenso: *Pipeline* Cinemática Geral | Transverso Geral..... 84
Figura 21 – Matriz de confusão do modelo vencedor para o alvo Transverso..... 88
Figura 22 – Intervalos de confiança *Bootstrap* 95%, dez melhores modelos: Transverso.. 89
Figura 23– AUC ROC macro – Transverso Geral 90
Figura 24 – Matriz de confusão do modelo vencedor para o alvo Apoio..... 93
Figura 25 – Intervalos de confiança *Bootstrap* 95%, dez melhores modelos: Apoio..... 94
Figura 26 – AUC ROC macro – Apoio..... 95
Figura 27 – Matriz de confusão do melhor modelo *Flat* para o alvo Balanço..... 98
Figura 28 – Matriz de confusão do melhor modelo Hierárquico para o alvo Balanço..... 98
Figura 29 – Intervalos de confiança *Bootstrap* 95%, dez melhores modelos: Balanço..... 99
Figura 30 – AUC ROC macro – Balanço..... 101
Figura 31 – Acurácia Balanceada: *Flat* vs. Hierárquica por alvo..... 103
Figura 32 – Acurácia Balanceada: Cinemática Angular vs. Geral por alvo..... 106

Figura 33 – Otimização de ponto de corte: classe TE (Transverso Externo)..... 109
Figura 34 – Otimização de ponto de corte: classe SS (Sagital Salto)..... 110
Figura 35 – Otimização de ponto de corte: classe SA (Sagital Agachado)..... 110
Figura 36 – Otimização de ponto de corte: TI (Transverso Interno)..... 111
Figura 37 – Otimização de ponto de corte: classe BR (Balanço Rígido)..... 111
Figura 38 – Matriz de confusão da Estratégia E3 (Retreinamento), alvo Apoio..... 113

LISTA DE QUADROS

Quadro 1 – Sistema de Classificação Hierárquica da Marcha de Davids e Bagley (GCS-CP)..... 36

LISTA DE TABELAS

Tabela 1 – Classificação Topográfica e Motora da Paralisia Cerebral..... 24
Tabela 2 – Fatores de Risco para Paralisia Cerebral por Período..... 25
Tabela 3 – Padrões de Lesão Cerebral por Idade Gestacional 27
Tabela 4 – Déficits Neuromusculares na PC Espástica..... 28
Tabela 5 – Variáveis da Análise Instrumentada da Marcha 31
Tabela 6 – Classificação de Winters para Hemiplegia Espástica..... 33
Tabela 7 – Estudos de Aprendizado de Máquina na Classificação da Marcha em PC 43
Tabela 8 – Comparativo de Algoritmos de aprendizado de máquina..... 48
Tabela 9 – Comparação entre a base de dados original e empilhada 56
Tabela 10 – Divisão dos dados entre conjuntos de treino e teste 61
Tabela 11 – Distribuição das classes por variável-alvo nos conjuntos de treino e teste 62
Tabela 12 – Configuração da seleção de variáveis por consenso.....63
Tabela 13 – Modelos e espaço de busca de hiperparâmetros.....67
Tabela 14 – Estratégias de regularização e prevenção de *overfitting*..... 71
Tabela 15 – Métricas de avaliação utilizadas no estudo..... 76
Tabela 16 – Principais variáveis selecionadas por alvo e fundamentação clínica..... 85
Tabela 17 – Melhor modelo por variável-alvo no conjunto de teste..... 86
Tabela 18 – 5 Melhores: Transverso Geral por Acurácia Balanceada no conjunto de teste ... 87
Tabela 19 – Análise Comparativa de Desempenho por Classe: Transverso.....91
Tabela 20 – 5 Melhores: Apoio por Acurácia Balanceada no conjunto de teste..... 92
Tabela 21 – Análise Comparativa de Desempenho por Classe: Apoio..... 96
Tabela 22 – 5 Melhores: Balanço por Acurácia Balanceada no conjunto de teste..... 97
Tabela 23 – Análise Comparativa de Desempenho por Classe: Balanço.....102
Tabela 24 – Resumo comparativo das abordagens por alvo.....102
Tabela 25 – Comparação entre abordagens *Flat* e Hierárquica por alvo.....103
Tabela 26 – Comparação Cinemática Angular vs. Geral por alvo..... 105
Tabela 27 – Comparação entre variantes "Com SMOTE" e "Sem SMOTE" por alvo 107

Tabela 28 – Comparação de Falsos Negativos por estratégia de otimização clínica. 108
Tabela 29 – Compromisso sensibilidade × especificidade (Estratégia E2) por classe.....112
Tabela 30 – Viabilidade clínica das estratégias de otimização por classe patológica.....137

LISTA DE ABREVIATURAS E SIGLAS

3DGA	Análise Tridimensional da Marcha (*Three-Dimensional Gait Analysis*)
Abord.	Abordagem (PCA ou Seleção de variáveis)
Acc	Acurácia (*Accuracy*)
ADM	Amplitude de Movimento
ANN	Rede Neural Artificial (*Artificial Neural Network*)
ANOVA	Análise de Variância (*Analysis of Variance*)
AUC	Área Sob a Curva ROC (*Area Under the Curve*)
AVC	Acidente Vascular Cerebral
Bal.Acc	Acurácia Balanceada (*Balanced Accuracy*)
BN	Balanço Normal (classe de classificação)
Bootstrap	Método de reamostragem com reposição para estimativa de intervalos de confiança
BR	Balanço Rígido (Stiff Knee Gait)
CAAE	Certificado de Apresentação de Apreciação Ética
CASIA	*Chinese Academy of Sciences Institute of Automation* (*dataset*)
CI	Contato Inicial
CP	*Cerebral Palsy* (Paralisia Cerebral em inglês)
CV	Validação Cruzada (*Cross-Validation*)
DTree	Árvore de Decisão (*Decision Tree*)
E1	Estratégia 1 de otimização clínica – seleção do melhor modelo por sensibilidade mínima patológica
E2	Estratégia 2 de otimização clínica – ajuste de ponto de corte de decisão para sensibilidade ≥ 90%
E3	Estratégia 3 de otimização clínica – retreinamento com função objetivo orientada à sensibilidade patológica mínima
ECE	Erro de Calibração Esperado
F1-Score	Média Harmônica de Precisão e *Recall*
FN	Falso Negativo
FP	Falso Positivo
G-Mean	Média Geométrica (*Geometric Mean)*
GCS-CP	Sistema de Classificação da Marcha para Paralisia Cerebral (*Gait Classification System for Cerebral Palsy*)

GMFCS	Sistema de Classificação da Função Motora Grossa (*Gross Motor Function Classification System*)
GPS	*Gait Profile Score* (Escore de Perfil da Marcha)
GRF	Força de Reação do Solo (*Ground Reaction Force*)
IA	Inteligência Artificial
IC	Intervalo de Confiança
K-NN	*K-Nearest Neighbors* (K Vizinhos Mais Próximos)
Kappa (κ)	Coeficiente Kappa de Cohen
LightGBM	*Light Gradient Boosting Machine*
L2	Regularização L2 (penalidade quadrática nos pesos)
LogReg	Regressão Logística (*Logistic Regression*)
MCC	*Matthews Correlation Coefficient* (Coeficiente de Correlação de
Matthews)	
Max_ANK_Flex	Flexão máxima do tornozelo *(Maximum Ankle Flexion)*
Max_HIP_Rot	Rotação máxima do quadril *(Maximum Hip Rotation)*
Min_ANK_Flex	Flexão mínima do tornozelo *(Minimum Ankle Flexion)*
Min_KNEE_Flex	Flexão mínima do joelho *(Minimum Knee Flexion)*
Min_KNEE_Rot	Rotação mínima do joelho *(Minimum Knee Rotation)*
ML	Aprendizado de Máquina *(Machine Learning)*
MLP	Perceptron Multicamadas *(Multilayer Perceptron)*
MV SFLDA	*Multivariate Sparse Functional Linear Discriminant Analysis*
Naive Bayes	Classificador Probabilístico Gaussiano baseado no Teorema de *Bayes*
NaN	Not a Number (Valor Ausente)
Optuna	Framework de otimização bayesiana de hiperparâmetros
PC	Paralisia Cerebral
PCA	Análise de Componentes Principais (*Principal Component Analysis*)
pp	Pontos Percentuais
RBF	Função de Base Radial (*Radial Basis Function*) – *kernel* do SVM
R^2	Coeficiente de Determinação
RF	*Random Forest* (Floresta Aleatória)
RN	Recém-nascido
ROC	Receiver Operating Characteristic (Curva ROC)
SA	Sagital Agachado (*Crouch Gait*)
SCALE	*Selective Control Assessment of the Lower Extremity*

Scikit-learn Biblioteca de *Machine Learning* em Python

Sens Sensibilidade (*Sensitivity*)

SMC Controle Motor Seletivo (*Selective Motor Control*)

SMOTE *Synthetic Minority Over-sampling Technique*

SN Sagital Normal

Spec Especificidade (*Specificity*)

SS Sagital Salto (*Jump Gait*)

SVD Decomposição em Valores Singulares (*Singular Value Decomposition*)

SVM Máquina de Vetores de Suporte (*Support Vector Machine*)

TCE Trato Corticoespinhal

TE Transverso Externo

TI Transverso interno

TN Transverso Normal (classe de classificação)

USF *University of South Florida (dataset)*

VN Verdadeiro Negativo

VP Verdadeiro Positivo

Voting Hard *Ensemble* por votação majoritária (predição da classe mais votada entre os classificadores-base)

Voting Soft *Ensemble* por média das probabilidades preditas pelos classificadores-base

XGBoost *eXtreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO..... 19
1.1 Objetivo Geral..... 22
1.2 Objetivos Específicos..... 22
2 REVISÃO DE LITERATURA..... 23
2.1 Paralisia Cerebral: Importância, Definição e Epidemiologia..... 23
2.2 Fatores de Risco e Etiologia..... 24
2.3 Lesões Cerebrais e Fisiopatologia..... 25
2.4 Déficits Neuromusculares na PC Espástica..... 27
2.4.1 Fraqueza Muscular..... 28
2.4.2 Espasticidade e Hipertonia..... 28
2.4.3 Comprometimento do Controle Motor Seletivo..... 29
2.5 Marcha na Paralisia Cerebral..... 29
2.5.1 Análise Cinemática..... 30
2.5.2 Análise Cinética / Dinâmica..... 31
2.6 Classificação dos Padrões de Marcha..... 32
2.6.1 Classificação de Winters para Hemiplegia Espástica..... 33
2.6.2 Sistema de Classificação para Diplegia Espástica..... 33
2.6.3 Sistema de Classificação da Marcha para Crianças com Paralisia Cerebral (GCS-CP)..... 34
2.7 Dificuldade na Classificação e Confiabilidade..... 38
2.8 Inteligência Artificial na Classificação da Marcha..... 39
2.8.1 Vantagens da IA na Prática Clínica..... 40
2.8.2 Aplicações em Paralisia Cerebral..... 42
2.9 Algoritmos de Machine Learning para Classificação da Marcha..... 45
2.9.1 "*Support Vector Machine"* (SVM)..... 46
2.9.2 Redes Neurais Artificiais (ANN)..... 46
2.9.3 Floresta Aleatória..... 47
2.9.4 "*Gradient Boosting*" (LightGBM e XGBoost)..... 48
2.9.5 Modelos de Fusão..... 49
3 MATERIAIS E MÉTODOS..... 52
3.1 Desenho do Estudo e Aspectos Éticos..... 52

3.2 Participantes.....52
3.3 Base de Dados e Estratégia de Empilhamento..... 55
3.3.1 Pré-processamento dos Dados Cinemáticos..... 59
3.4 Divisão dos Dados e Prevenção de Vazamento..... 60
3.5 Variáveis Alvo..... 62
3.6 Seleção de Variáveis.....62
3.7 Modelos de Aprendizado de Máquina..... 65
3.8 Tratamento de Desbalanceamento e Estratégias de Regularização.....69
3.8.1 SMOTE: Balanceamento da Classe Minoritária e Prevenção de Vazamento de Dados.....69
3.8.2 Regularização nas Redes Neurais..... 69
3.8.3 *Overfitting* e seu Monitoramento..... 70
3.9 Análise Complementar: uma abordagem hierárquica.....72
3.9.1 Estrutura da Classificação Hierárquica.....73
3.9.2 Seleção de Características por Etapa..... 73
3.9.3 Cálculo da Sensibilidade Efetiva (Propagação de Erros)..... 73
3.9.4 Análise de Estabilidade Estocástica.....74
3.10 Métricas de Avaliação..... 75
3.11 Estratégias de Otimização Clínica.....79
3.12 Uso de Inteligência Artificial Generativa..... 79
4 RESULTADOS.....81
4.1 Seleção de Atributos e Configuração dos Modelos.....81
4.1.1 Variáveis Mais Relevantes por Alvo..... 85
4.2 Resultados do Conjunto de Teste..... 86
4.3 Ranqueamento Global dos Modelos.....86
4.4 Ranqueamento por alvo: Transverso Geral.....87
4.5 Ranqueamento por alvo: Apoio Geral.....91
4.6 Ranqueamento por alvo: Balanço Geral..... 96
4.7 Comparação *Flat* vs Hierárquico..... 102
4.8 Comparação Angular vs Geral.....104
4.9 Comparação "Com SMOTE" vs "Sem SMOTE"..... 107
4.10 Análise Clínica: Minimização de Falsos Negativos (FN).....107
5 DISCUSSÃO..... 114
5.1 Síntese dos Principais Achados.....114

5.2 Abordagem hierárquica versus abordagem multiclasse direta.....115
5.3 Seleção de Atributos, PCA e Pipelines de Variáveis.....118
5.4 Análise Comparativa de Desempenho por Classe..... 120
5.5 Limites Biomecânicos e Estratégias de Mitigação Clínica.....122
5.5.1 Marcha em joelho rígido (BR): limite biomecânico e teto da Estratégia E2.....122
5.5.2 Marcha em salto (SS): similaridade com padrões adjacentes e ganhos da Estratégia E3.....124
5.6 Modelos de Fusão e Análise de Desempenho..... 125
5.7 Adequação dos Algoritmos ao Problema.....126
5.8 Análise do *Overfitting*..... 128
5.9 Comparação com a Literatura..... 130
5.10 Otimização Clínica e Minimização de Falsos Negativos..... 133
5.11 Implicações Clínicas..... 138
5.12 Limitações.....140
5.13 Contribuições Metodológicas.....142
5.14 Direções para Trabalhos Futuros.....145
6 CONCLUSÃO.....148
REFERÊNCIAS BIBLIOGRÁFICAS..... 150
APÊNDICE A – VARIÁVEIS REGRA IPSILATERAL..... 156

1 INTRODUÇÃO

A paralisia cerebral (PC) representa a deficiência física mais comum com início na infância, com prevalência estimada de 17 milhões de pessoas em todo o mundo e taxas que variam entre os países (1,2). A condição afeta aproximadamente 1 em cada 500 recém-nascidos, com taxas variando de 1,5 a 2,5 por 1.000 nascidos vivos em países desenvolvidos (1,3–5). A descrição atualizada da PC, elaborada por meio de processo multidisciplinar de consenso global, define-a como uma condição neurodesenvolvimental de início precoce e curso ao longo da vida, caracterizada por limitações na atividade decorrentes de desenvolvimento comprometido do movimento e da postura, atribuída a malformação ou lesão do cérebro fetal ou infantil de caráter não degenerativo, embora as manifestações clínicas e funcionais possam mudar com a idade (6).

As manifestações clínicas da PC variam amplamente em termos do tipo de distúrbio do movimento, grau de capacidade funcional, limitação e partes do corpo afetadas (1,7,8). A PC espástica é o tipo mais comum, representando aproximadamente 87% dos casos, sendo seguida pela PC discinética, que afeta aproximadamente 7,5%, e pela PC atáxica, que afeta aproximadamente 4% (7). Os distúrbios motores são frequentemente acompanhados por alterações da sensação, percepção, cognição, comunicação e comportamento, epilepsia e por problemas musculoesqueléticos secundários (1,7). Embora a lesão cerebral inicial seja não progressiva, as deficiências musculoesqueléticas e as limitações funcionais associadas à PC são de fato progressivas ao longo do tempo (1,7).

As alterações na marcha constituem uma das manifestações mais prevalentes e funcionalmente relevantes da PC, resultando da combinação de déficits neuromusculares primários e adaptações secundárias do sistema musculoesquelético (7–9). A espasticidade, a fraqueza muscular, o controle motor seletivo prejudicado e o encurtamento da unidade músculo-tendínea interagem para produzir padrões da marcha característicos descritos na Seção 2.6.2 (9,10). A análise tridimensional da marcha (3DGA) tornou-se o padrão-ouro para a avaliação quantitativa dessas anormalidades, fornecendo dados cinemáticos, cinéticos e eletromiográficos essenciais para o planejamento terapêutico (11–13).

Entretanto, a classificação dos padrões da marcha na PC permanece um desafio clínico significativo. Melanda et al. (14) demonstraram que a confiabilidade intra-avaliador para o Sistema de Classificação da Marcha para Crianças com Paralisia Cerebral

(GCS-CP) varia de moderada a substancial, enquanto a confiabilidade interavaliadores é apenas moderada, indicando considerável variabilidade entre profissionais, variabilidade essa que decorre da natureza multidimensional dos dados da marcha e da sobreposição entre categorias diagnósticas (10,11,14). A necessidade de métodos objetivos e reprodutíveis para a classificação da marcha tornou-se evidente, impulsionando a investigação de abordagens baseadas em inteligência artificial (IA) e aprendizado de máquina (do inglês, *machine learning* - ML) (15–18).

A aplicação de técnicas de ML na análise da marcha tem demonstrado resultados promissores. Em PC, estudos reportam acurácias superiores a 90% com diferentes modelos, aplicados contudo a tarefas distintas, como diferenciação entre crianças com PC e controles saudáveis ou classificação de padrões sagitais segundo o sistema de Rodda e Graham, e não necessariamente ao GCS-CP. Kamruzzaman e Begg (19) obtiveram esse desempenho com máquinas de vetores de suporte (SVM), Slijepcevic et al. (15) com floresta aleatória (RF), e Zhang e Ma (17) com redes neurais artificiais (ANN). Algoritmos como ANN, RF, SVM e métodos de "gradient boosting" como LightGBM e XGBoost têm sido investigados para a classificação automatizada de padrões da marcha patológicos (15,17,20–22). Estas abordagens computacionais oferecem potencial para processamento consistente de dados de alta dimensionalidade, identificação de padrões não lineares e fornecimento de classificações reprodutíveis que podem auxiliar na tomada de decisão clínica (16,18,23).

Apesar dos avanços reportados na literatura (15,17,20,24), persistem seis lacunas relevantes para a aplicação clínica das técnicas de aprendizado de máquina na classificação da marcha em paralisia cerebral. Primeiro, embora algoritmos individuais venham reportando acurácias superiores a 90% para outros sistemas de classificação da marcha (15,17,19), não está estabelecido se esse desempenho se sustenta de forma sistemática para o GCS-CP em níveis clinicamente relevantes (>80%) capazes de superar a variabilidade interavaliadores reportada por Melanda et al. (14). Segundo, as comparações entre estratégias de pré-processamento (seleção supervisionada de variáveis por consenso de múltiplos métodos versus a redução não supervisionada por análise de componentes principais, PCA) permanecem pouco sistematizadas (25,26), apesar do potencial da abordagem por consenso para preservar variáveis biomecânicas interpretáveis. Terceiro, modelos de fusão (*ensembles*) vêm sendo apontados como capazes de superar algoritmos individuais por combinar capacidades complementares (27,28), mas faltam estudos que avaliem sistematicamente esse ganho na classificação da marcha em PC. Quarto, a maior parte dos estudos reporta apenas métricas agregadas (acurácia global, *F1-score* médio), que podem mascarar a heterogeneidade do

desempenho entre classes e falhas sistemáticas na detecção de padrões minoritários clinicamente críticos (29–31), situação particularmente preocupante diante do desbalanceamento natural das classes patológicas. Quinto, a ausência de estratégias de otimização orientadas à minimização de falsos negativos para os subtipos patológicos limita a transposição desses modelos para uso clínico como ferramenta de triagem (32,33). Sexto, embora o sistema GCS-CP proposto por Davids e Bagley (9) e validado por Melanda et al. (14) ofereça uma estrutura de classificação clínica consolidada com três alvos (Apoio, Balanço e Transverso), não foram exploradas de forma abrangente abordagens computacionais que decomponham essa estrutura em etapas sequenciais de triagem (Normal vs. Alterado) e subtipagem, apesar dos benefícios reportados pela classificação hierárquica em outros domínios (34).

Neste contexto, o presente estudo avalia técnicas de ML para classificar os padrões da marcha do GCS-CP em pacientes com PC espástica. Diferentemente de estudos anteriores que se limitam a reportar métricas agregadas (como acurácia global *ou F1-score* médio), este trabalho enfatiza a análise do desempenho por classe como critério essencial de aplicabilidade clínica, reconhecendo que métricas agregadas podem mascarar falhas sistemáticas na detecção de padrões clinicamente relevantes.

Para responder a essas lacunas, foram formuladas seis hipóteses iniciais para este estudo: (H1) algoritmos de ML seriam capazes de classificar padrões da marcha do GCS-CP com acurácia clinicamente relevante (>80%), superando a variabilidade moderada reportada em avaliações interavaliadores humanas; (H2) a comparação entre seleção de atributos por consenso e redução de dimensionalidade por PCA revelaria diferenças sistemáticas de desempenho entre os alvos do GCS-CP, sem superioridade universal de uma estratégia sobre a outra; (H3) modelos de fusão apresentariam desempenho superior aos algoritmos individuais, ao combinar suas capacidades complementares; (H4) a análise por classe revelaria heterogeneidade de desempenho entre diferentes padrões de marcha, com possível superioridade na classificação de padrões mais frequentes ou biomecanicamente mais distintos, justificando a necessidade de métricas individualizadas para avaliação de aplicabilidade clínica; (H5) estratégias de otimização clínica orientadas à sensibilidade patológica mínima permitiriam reduzir substancialmente os falsos negativos para a maioria das classes avaliadas, aproximando os modelos de limiares clinicamente aceitáveis para uso como ferramentas de triagem; e (H6) uma abordagem de classificação hierárquica em cascata, com etapa de triagem (Normal vs. Alterado) seguida de subtipagem, apresentaria desempenho clínico por classe diferente da classificação multiclasse direta (*Flat*), especialmente nas classes patológicas minoritárias.

1.1 Objetivo Geral

Avaliar comparativamente modelos de aprendizado de máquina para a classificação automática dos padrões da marcha do Sistema de Classificação da Marcha para Crianças com Paralisia Cerebral (GCS-CP) em crianças com paralisia cerebral espástica, enfatizando a análise do desempenho por classe como critério de aplicabilidade clínica.

1.2 Objetivos Específicos

a) Comparar o desempenho de nove algoritmos de aprendizado de máquina (Regressão Logística, SVM, Árvores de Decisão, K vizinhos mais próximos (K-NN), Naive Bayes, Floresta Aleatória, LightGBM, XGBoost e Redes Neurais Artificiais) e três modelos de fusão (*Voting Hard*, *Voting Soft* e *Stacking*) na classificação dos três componentes do GCS-CP (Apoio, Balanço e Transverso) em pacientes com paralisia cerebral espástica;

b) Comparar o desempenho dos modelos de classificação utilizando três estratégias de pré-processamento: (I) sem filtragem adicional de variáveis (Sem Consenso, linha de base), (II) seleção supervisionada de variáveis por consenso de múltiplos métodos (Com Consenso) e (III) redução dimensional não supervisionada por Análise de Componentes Principais (PCA), identificando empiricamente a estratégia de maior desempenho por alvo, definida pela Acurácia Balanceada no conjunto de teste.

c) Analisar o desempenho individualizado por classe, por meio de sensibilidade, especificidade e valores preditivos, para cada padrão de marcha classificado, identificando classes com risco de subdiagnóstico clínico;

d) Avaliar uma abordagem de classificação hierárquica em cascata, com etapa de triagem (Normal vs. Alterado) seguida de subtipagem, como alternativa à classificação multiclasse direta (doravante denominada *Flat*), investigando se a decomposição do problema em etapas sequenciais impacta o desempenho clínico por classe.

e) Avaliar estratégias de otimização clínica orientadas à minimização de falsos negativos para as cinco classes patológicas avaliadas, comparando: seleção do melhor modelo por sensibilidade mínima patológica (E1), ajuste de ponto de corte de decisão para sensibilidade $\geq$ 90% (E2) e retreinamento com função objetivo orientada à sensibilidade

patológica mínima via *Optuna* (E3), quantificando a redução de casos patológicos não detectados em relação ao modelo padrão selecionado por acurácia balanceada.

2 REVISÃO DE LITERATURA

Esta revisão segue uma progressão dos fundamentos clínicos da paralisia cerebral aos algoritmos de aprendizado de máquina aplicados à classificação da marcha. Priorizaram-se revisões sistemáticas e estudos clínicos publicados nos últimos dez anos, complementados por trabalhos seminais que estabeleceram os fundamentos teóricos da área.

2.1 Paralisia Cerebral: Importância, Definição e Epidemiologia

A paralisia cerebral é a causa mais comum de incapacidade física com início na infância, representa um dos maiores desafios para os sistemas de saúde em todo o mundo (1,35). Em sua descrição mais atual, a paralisia cerebral é caracterizada como uma condição neurodesenvolvimental de início precoce e curso permanente ao longo da vida, marcada por limitações de atividade decorrentes do desenvolvimento comprometido do movimento e da postura, comumente associadas a alterações do tônus muscular, clinicamente categorizadas como espasticidade, distonia, coreoatetose e/ou ataxia. É atribuída a malformação ou lesão não degenerativa do cérebro fetal ou infantil, ainda que suas manifestações clínicas e funcionais possam mudar com a idade. O fenótipo é heterogêneo e potencialmente complexo, com apresentação única em cada indivíduo, sendo frequente a coexistência de comprometimentos primários e secundários em diferentes domínios do desenvolvimento e da funcionalidade, com impacto relevante sobre a participação na vida diária (6). As nuances dessa descrição são fundamentais: o caráter permanente e de curso prolongado indica que a condição acompanha o indivíduo ao longo da vida; o comprometimento do movimento e da postura refere-se às manifestações motoras primárias; e o termo "não degenerativa" significa que a lesão cerebral original não progride, ainda que suas consequências funcionais possam evoluir com a idade (1,6).

Os registros populacionais de paralisia cerebral, principalmente na Austrália e na Europa, historicamente encontraram prevalência variando de 1,5 a 2,5 por 1.000 nascidos vivos (1,3–5). Globalmente, estima-se que a PC afete aproximadamente 17 milhões de pessoas (1,2). A PC pode ser classificada de acordo com o tipo de distúrbio motor predominante e topografia, conforme apresentado na Tabela 1. Além dessas dimensões, a gravidade do

comprometimento funcional é descrita pelo Sistema de Classificação da Função Motora Grossa (GMFCS), que estratifica a função motora grossa, sobretudo a mobilidade autoiniciada, como sentar, transferir-se e caminhar, em cinco níveis, do mais funcional (I) ao mais dependente (V). Os níveis I a III correspondem, em geral, a indivíduos deambuladores, com ou sem dispositivo auxiliar, enquanto os níveis IV e V indicam mobilidade predominantemente assistida ou por cadeira de rodas (36,37).

Tabela 1 – Classificação Topográfica e Motora da Paralisia Cerebral

Classificação	Tipo	Prevalência	Características Principais
Por distúrbio motor	Espástica	~87%	Hipertonia velocidade-dependente; aumento do reflexo de estiramento
	Discinética	~7,5%	Movimentos involuntários; tônus muscular flutuante
	Atáxica	~4%	Incoordenação; instabilidade postural; tremor intencional
	Mista	~1,5%	Combinação de padrões espástico e discinético
Por topografia	Hemiplegia	-	Um hemicorpo afetado (membro superior e inferior ipsilaterais)
	Diplegia	-	Membros inferiores predominantemente afetados
	Tetraplegia	-	Quatro membros afetados; frequentemente com envolvimento axial

Fonte: *Adaptado de Graham et al.* (1) *e Zhou et al.* (7).

2.2 Fatores de Risco e Etiologia

A etiologia da PC é multifatorial, envolvendo fatores pré-natais, perinatais e pós-natais que podem resultar em lesão do cérebro fetal ou infantil em desenvolvimento, em geral nos primeiros anos de vida, convencionalmente até cerca de dois anos de idade, segundo os critérios dos registros europeu e australiano de paralisia cerebral, e não ao longo de toda a infância (1,7,35). O nascimento prematuro é o fator de risco mais importante, com o risco aumentando de forma constante com o declínio da idade gestacional ao nascimento (1). O risco em bebês nascidos antes de 28 semanas de gestação é aproximadamente 50 vezes maior do que em nascidos a termo (1). Entre os prematuros, evidência de lesão da substância branca na ultrassonografia craniana ou ressonância magnética confere um risco de aproximadamente 50% de desenvolver PC (1). A Tabela 2 resume os principais fatores de risco por período.

Tabela 2 – Fatores de Risco para Paralisia Cerebral por Período

Período	Fator de Risco	Observações
Pré-natal	Prematuridade (<28 semanas)	Risco 50 vezes maior que nascidos a termo
	Lesão substância branca	~50% risco de desenvolver PC
	Inflamação/infecção fetal	Contribuição independente para prematuridade e PC
	Malformações cerebrais	10-15% dos casos detectáveis por neuroimagem
	Fatores genéticos	Podem contribuir para até 30% dos casos
Perinatal	Asfixia ao nascimento	<10% dos casos claramente atribuíveis
	Baixo índice de Apgar	Correlação com risco aumentado em RN termo
	Corioamnionite	Especialmente relevante em prematuros
	AVC isquêmico perinatal	<5% dos casos em RN termo
Fatores Protetores	Sulfato de magnésio	Reduz risco de PC em ~30% no parto prematuro
	Hipotermia terapêutica	Reduz morte e neurodeficiência em asfixia perinatal

Fonte: *Adaptado de Graham et al.* (1) *e McIntyre et al.* (35).

Historicamente, a asfixia ao nascimento foi considerada uma causa primária de PC, mas evidências atuais indicam que apenas menos de 10% dos casos são claramente atribuíveis a asfixia maior ao nascimento (1,7). Os bebês expostos a processos inflamatórios / infecciosos intrauterinos (como corioamnionite) têm maior probabilidade de nascer prematuramente e desenvolver PC; essa inflamação fetal provavelmente contribui de forma independente para ambos os desfechos (1,35).

2.3 Lesões Cerebrais e Fisiopatologia

Em aproximadamente 90% dos casos, a PC resulta de processos destrutivos que lesam o tecido cerebral saudável, e não de anormalidades primárias no desenvolvimento cerebral (1,7). A hipóxia e a isquemia têm sido tradicionalmente propostas como mecanismos principais de lesão, embora a inflamação seja cada vez mais reconhecida como uma via final comum na patogênese (1). Estudos patológicos e de imagem demonstraram combinações variadas de lesões no córtex cerebral, substância branca hemisférica, gânglios da base e cerebelo (1,7,8). A Figura 1 ilustra as regiões cerebrais e os tratos motores afetados nos diferentes tipos de PC.

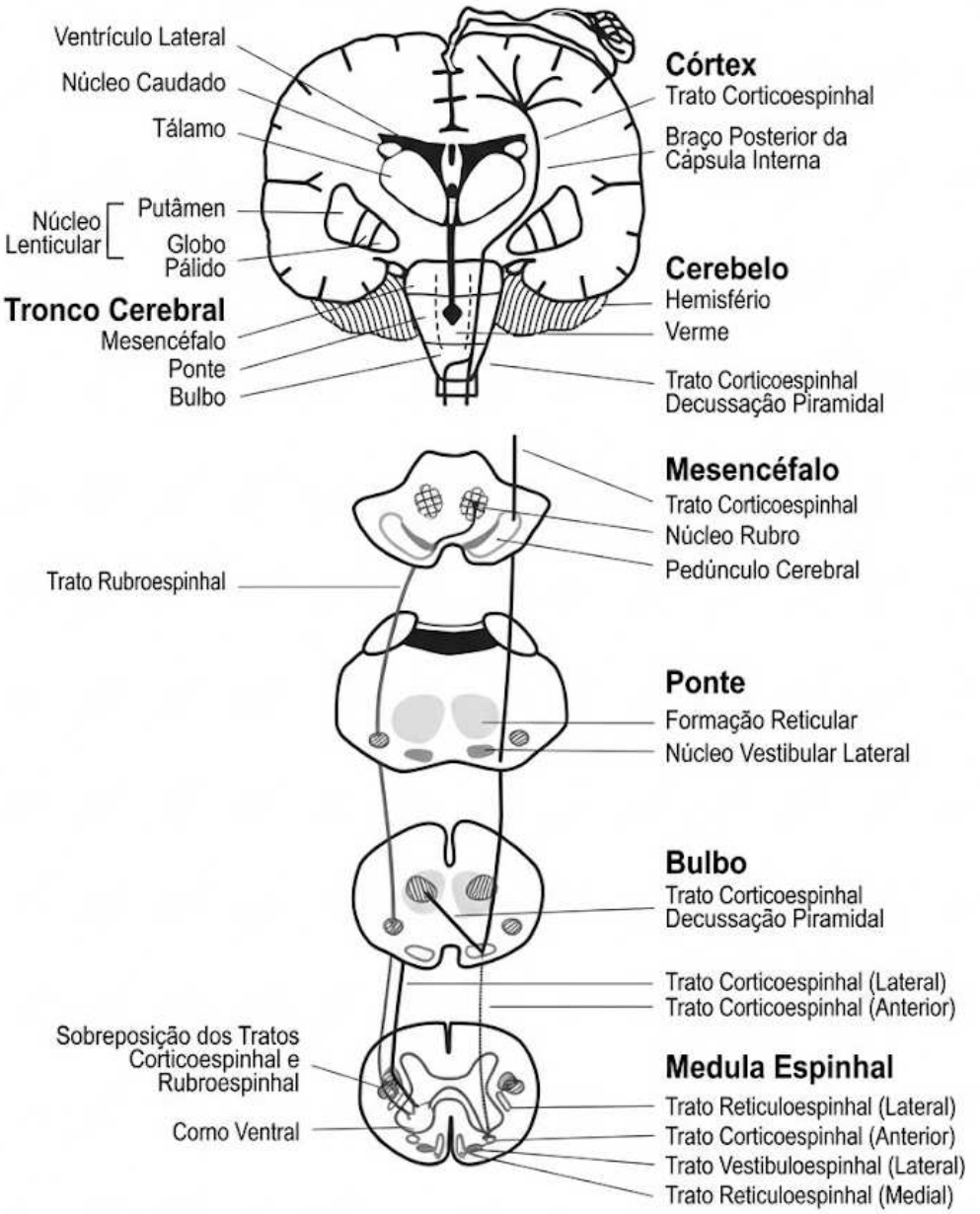

Figura 1 – Regiões cerebrais afetadas na PC espástica, discinética e atáxica.
Nota: Uma porção do homúnculo motor está sobreposta ao córtex. Representações dos ventrículos laterais, núcleos subcorticais, cerebelo, tronco encefálico e tratos corticoespinhal (TCE) e rubroespinhal estão delineados. O vermis do cerebelo (posterior ao tronco encefálico) é representado por linhas pontilhadas. As regiões onde os tratos reticuloespinhal e vestibuloespinhal descendem através de cada camada do tronco encefálico e da medula espinhal estão sombreadas.
Fonte: *Adaptado e traduzido de Zhou J, Butler EE, Rose J. Neurologic correlates of gait abnormalities in cerebral palsy: implications for treatment. Front Hum Neurosci. 2017;11:103. Licença CC-BY 4.0.*

O estágio de maturação cerebral durante o qual ocorrem os eventos patogênicos define o tipo e a localização das lesões, bem como a resposta específica à lesão (1,8). A Tabela 3 apresenta os padrões de lesão cerebral de acordo com a idade gestacional.

Tabela 3 – Padrões de Lesão Cerebral por Idade Gestacional

Idade Gestacional	Estrutura Vulnerável	Padrão de Lesão	Tipo Clínico Resultante
Prematuro (<32 sem)	Substância branca periventricular profunda	Leucomalácia periventricular (LPV) - difusa, focal ou cística	PC espástica diplégica
Termo (>37 sem)	Córtex cerebral, gânglios da base, tálamo	Lesão cortical e subcortical; necrose seletiva	PC espástica, discinética ou mista

Fonte: *Adaptado de Graham et al.* (1).

2.4 Déficits Neuromusculares na PC Espástica

Na PC espástica, a lesão do trato corticoespinhal (TCE) resulta em quatro déficits neuromusculares inter-relacionados que impactam significativamente a função motora: espasticidade, controle motor seletivo (SMC) prejudicado, fraqueza muscular e unidade músculo-tendão encurtada (7). A Tabela 4 resume esses déficits e suas implicações para a marcha.

Tabela 4 – Déficits Neuromusculares na PC Espástica

Déficit	Definição	Mecanismo	Impacto na Marcha
Fraqueza muscular	Redução da força voluntária máxima	Perda de sinais excitatórios do TCE; redução de unidades motoras	Incapacidade de gerar força propulsiva adequada
Espasticidade	Hipertonia velocidade-dependente	Perda de inibição descendente do reflexo de estiramento	Resistência ao movimento; posturas anormais
SMC prejudicado	Incapacidade de isolar ativação muscular seletiva	Sinergias flexoras/extensoras interferindo em movimentos isolados	Co-ativação antagonista; movimentos em bloco
Encurtamento músculo-tendíneo	Unidade músculo-tendão com comprimento reduzido	Crescimento muscular interrompido; estímulo mecânico insuficiente	Contraturas; limitação de amplitude de movimento

Fonte: *Adaptado de Zhou et al.* (7) *e Fowler & Goldberg.* (38).

2.4.1 Fraqueza Muscular

A fraqueza muscular é um déficit frequentemente subestimado na PC espástica, mas contribui significativamente para posturas e movimentos anormais (7,8). É postulada como resultado da perda de sinais motores excitatórios descendentes no TCE, levando à redução da ativação neuromuscular e ao tamanho muscular diminuído (7,8). Relações positivas importantes entre força muscular, função motora grossa e resultados funcionais foram estabelecidas, indicando que a fraqueza é responsável por mais incapacidade do que a espasticidade (1).

2.4.2 Espasticidade e Hipertonia

A definição fisiológica de espasticidade atualmente aceita destaca a importância do reflexo de estiramento dependente da velocidade (1,7,10). A resposta do reflexo de estiramento aumenta aproximadamente linearmente com o aumento da velocidade de estiramento, e o componente reflexo do tônus aumentado pode ser medido em termos da velocidade limiar necessária para evocar a atividade reflexa (1,8). Entretanto, a espasticidade explica apenas parcialmente a má função motora grossa na PC, sendo necessárias abordagens terapêuticas multimodais (1,10).

2.4.3 Comprometimento do Controle Motor Seletivo

O controle motor seletivo (SMC) é a capacidade de isolar a ativação muscular em um padrão selecionado conforme as demandas de uma postura ou movimento voluntário; seu déficit caracteriza-se pela redução dessa capacidade (7). Em vez de ativar os músculos de forma isolada e fracionada, o indivíduo passa a recrutá-los em padrões estereotipados e globais, as chamadas sinergias de flexão ou de extensão, nos quais diversos músculos se contraem em conjunto. Como a marcha exige o controle seletivo e independente de cada articulação ao longo do ciclo, essas sinergias restringem os movimentos articulares isolados e comprometem funções motoras como a locomoção (7,8). Chruscikowski et al. (39), em estudo com 194 crianças com PC bilateral classificadas nos níveis I–III do GMFCS, avaliaram o controle motor seletivo por meio da *Selective Control Assessment of the Lower Extremity* (SCALE), instrumento clínico que gradua, articulação por articulação (quadril, joelho, tornozelo, subtalar e dedos), a capacidade de executar movimentos voluntários isolados no membro inferior. Os autores demonstraram correlação negativa significativa ($r = -0{,}603$; $p < 0{,}001$) entre o escore da SCALE e o *Gait Profile Score* (GPS), indicando que quanto maior o comprometimento do SMC, maior a anormalidade global da marcha. Os autores sugerem que o dano ao trato corticoespinhal pode afetar não apenas o SMC, mas também o desenvolvimento das redes neurais espinhais responsáveis pelo controle automático da marcha.

2.5 Marcha na Paralisia Cerebral

A marcha é uma atividade motora complexa que envolve o movimento cíclico e alternado dos membros inferiores para propulsionar o corpo no espaço, sendo fundamental para a independência funcional e participação social (7–9). O ciclo da marcha é dividido em duas fases principais: a fase de apoio, quando o pé está em contato com o solo (representando aproximadamente 60% do ciclo), e a fase de balanço, quando o membro está avançando no ar (aproximadamente 40% do ciclo) (7,9), conforme ilustrado na Figura 2. As anormalidades da marcha na PC resultam da interação complexa entre déficits neuromusculares primários, alterações musculoesqueléticas secundárias e compensações terciárias (7,9,10). A utilidade da análise instrumentada da marcha no planejamento terapêutico de crianças com PC foi recentemente demonstrada por Schwartz et al., que mostraram que modelos de ML incluindo dados de análise instrumentada predizem desfechos cinemáticos de tratamento, como a

progressão do pé na fase de apoio, com maior acurácia ($R^2 = 0,45$) do que modelos baseados apenas em dados clínicos convencionais ($R^2 = 0,28$) (13).

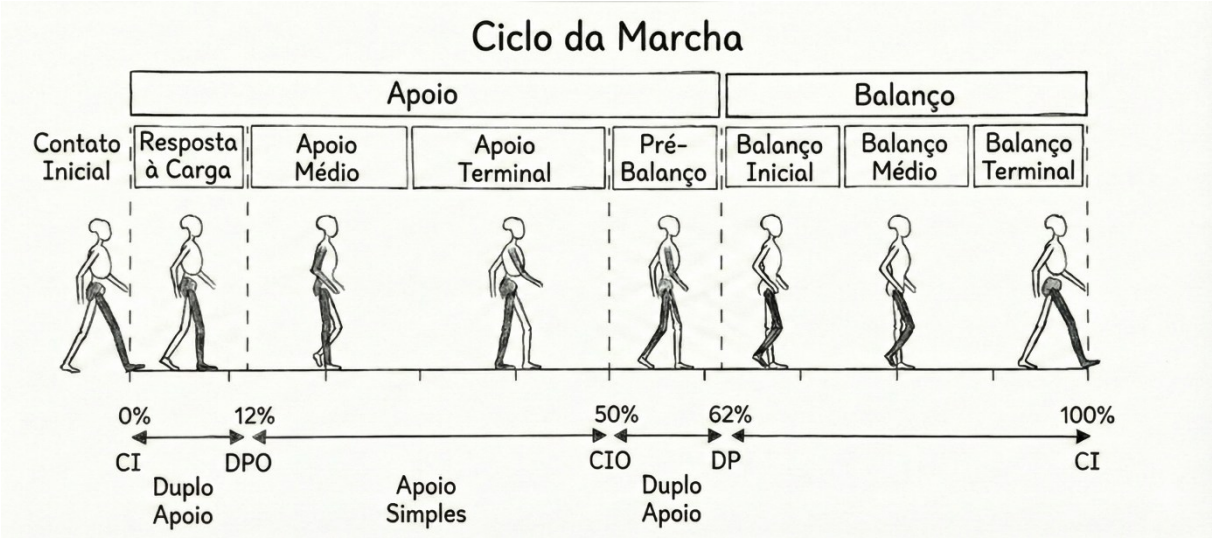

Figura 2 – O ciclo da marcha e suas fases em relação à PC espástica. A fase de apoio representa aproximadamente 60% do ciclo e inclui: contato inicial (CI), resposta à carga, apoio médio, apoio terminal e pré-balanço. A fase de balanço representa aproximadamente 40% do ciclo e inclui: balanço inicial, balanço médio e balanço terminal. CI = contato inicial; DPO = desprendimento do pé oposto; CIO = contato inicial do pé oposto; DP = desprendimento do pé.
Fonte: *Adaptado e traduzido de Zhou J, Butler EE, Rose J. Neurologic correlates of gait abnormalities in cerebral palsy: implications for treatment. Front Hum Neurosci. 2017;11:103. Licença CC-BY 4.0.*

2.5.1 Análise Cinemática

A análise cinemática refere-se ao estudo do movimento dos segmentos corporais e articulações sem levar em consideração as suas causas, produzindo medidas de ângulos articulares, amplitude de movimento, deslocamento, velocidade e aceleração ao longo de vários eixos (12,23). Os sistemas de captura de movimento optoeletrônicos com marcadores retrorrefletivos constituem o padrão-ouro para a análise cinemática tridimensional (3D), permitindo o rastreamento das coordenadas dos marcadores fixados em pontos anatômicos de interesse (9,20). A maioria dos estudos de classificação da marcha na PC utiliza predominantemente dados no plano sagital, embora informações dos três planos possam fornecer uma caracterização mais completa do padrão da marcha (11,15,17).

2.5.2 Análise Cinética / Dinâmica

As variáveis dinâmicas incluem medidas de força, como as forças de reação do solo (GRF) e momentos articulares, que são fundamentais para a compreensão da dinâmica da marcha (9,20). Os momentos articulares internos fornecem informações sobre as demandas impostas aos músculos durante a marcha (23). Estudos demonstram que dados cinemáticos frequentemente superam os dados de GRF isolados para a classificação de padrões da marcha relacionados à PC (15). A Tabela 5 e a Figura 3 resumem as principais variáveis e parâmetros da análise instrumentada da marcha.

Tabela 5 – Variáveis da Análise Instrumentada da Marcha

Tipo de Análise	Variáveis Principais	Equipamento	Aplicação Clínica
Cinemática	Ângulos articulares, ADM, velocidade angular, deslocamento	Câmeras optoeletrônicas + marcadores retrorrefletivos	Identificação de padrões de movimento articular
Cinética	GRF (3 componentes), momentos articulares, potência	Plataformas de força	Análise das forças que geram ou resistem ao movimento
Eletromiografia	Timing e intensidade de ativação muscular	Eletrodos de superfície ou agulha	Identificação de co-contrações e ativação anormal

Nota: ADM = Amplitude de movimento.
Fonte: *Adaptado de Chang et al.* (12) *e Prakash et al.* (23).

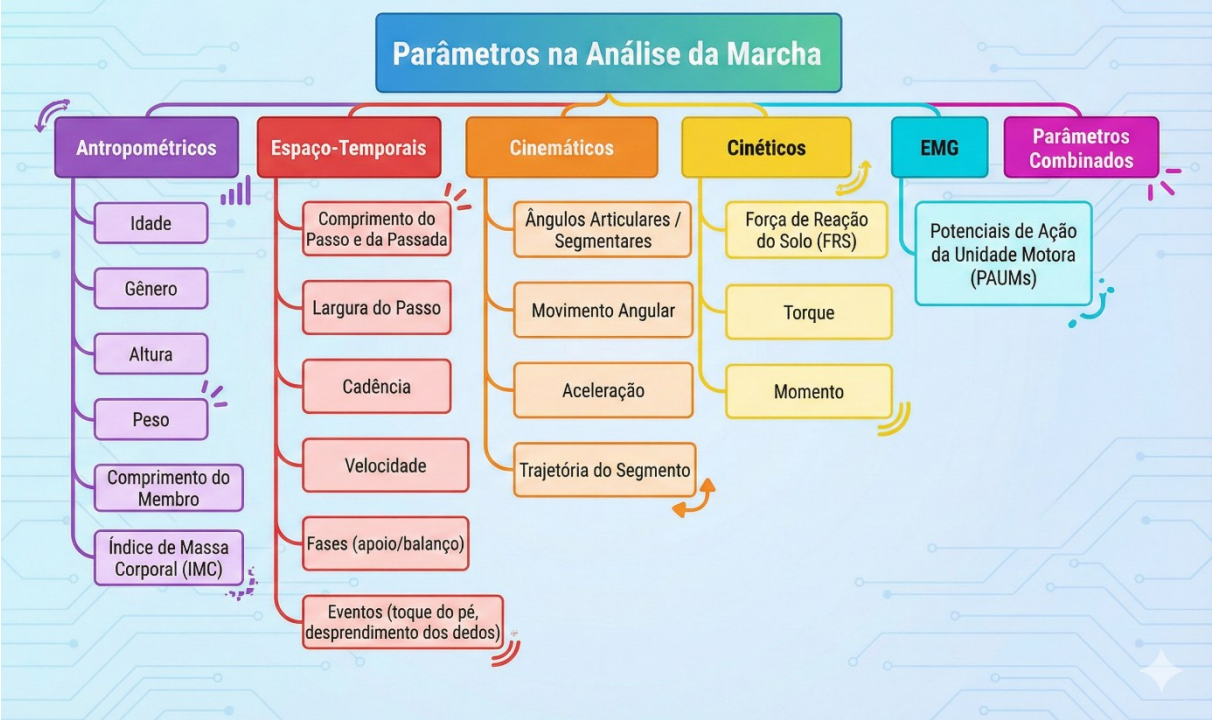

Figura 3 – Árvore abrangente de parâmetros na análise da marcha. Os parâmetros são organizados em seis categorias principais: Antropométricos, Espaço-temporais, Cinemáticos, Cinéticos, EMG e Parâmetros Combinados.
Nota: No presente estudo, foram utilizadas variáveis das categorias espaço-temporal (comprimento do passo, comprimento da passada, cadência e velocidade), cinemática (ângulos articulares da pelve, quadril, joelho, tornozelo e ângulo de progressão do pé nos planos sagital, coronal e transverso).
***Fonte*:** *Elaborada pelo autor, adaptada e traduzida com base em Prakash C, Kumar R, Mittal N. Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. Artif Intell Rev. 2018;49(1):1-40.*

2.6 Classificação dos Padrões de Marcha

A classificação de padrões da marcha na PC espástica permite identificar os desvios predominantes e orientar o planejamento terapêutico. Dois sistemas são amplamente utilizados: a classificação de Winters para hemiplegia e a classificação de Rodda e Graham para diplegia, que foi posteriormente expandida por Davids e Bagley (9,10). Cabe ressaltar que esses sistemas são qualitativos: foram construídos a partir da observação visual dos padrões de marcha e da experiência clínica acumulada pelos próprios autores, e não de critérios quantitativos objetivos. Por dependerem do julgamento do avaliador, estão sujeitos à variabilidade inter e intraexaminador, o que motiva o desenvolvimento de abordagens automatizadas e quantitativas de classificação.

2.6.1 Classificação de Winters para Hemiplegia Espástica

Winters et al. subdividiram a hemiplegia espástica em quatro padrões da marcha com base na cinemática do plano sagital e na eletromiografia dinâmica (EMG) (40). Rodda e Graham refinaram posteriormente o Tipo 2 em dois subtipos (2A e 2B), conforme o comportamento do joelho durante o apoio (10). Esta classificação tem relevância direta para a compreensão do padrão da marcha e manejo terapêutico, conforme apresentado na Tabela 6.

Tabela 6 – Classificação de Winters para Hemiplegia Espástica

Tipo	Tornozelo	Joelho	Quadril	Músculos Envolvidos
1	Pé caído no balanço; apoio normal	Normal	Normal	Dorsiflexores (fraqueza)
2A	Equino verdadeiro	Neutro/Extensão	Normal	Gastrocnêmio-sóleo
2B	Equino verdadeiro	Recurvato	Normal	Gastrocnêmio-sóleo
3	Equino	Flexão + rigidez	Normal	Gastrocnêmio + Isquiotibiais
4	Equino	Flexão + rigidez	Flexão + Adução + RI	Multinível

Fonte: *Adaptado Winters et al.* (40) *e Rodda & Graham* (10).

2.6.2 Sistema de Classificação para Diplegia Espástica

Rodda e Graham (10) propuseram uma classificação dos padrões da marcha na diplegia espástica baseada nos trabalhos de Sutherland e Davids, considerando o plano sagital de forma global: pelve, quadril, joelho e tornozelo. Diferentemente da abordagem de Sutherland e Davids, que focava apenas no movimento do joelho, Rodda e Graham ampliaram a análise para incluir todas as articulações do membro inferior, identificando quatro padrões progressivos: (I) equino verdadeiro, com dominância do músculo gastrocnêmio e tornozelo em flexão plantar durante o apoio com quadril e joelho estendidos; (II) marcha em salto (*jump gait*), com espasticidade dos isquiotibiais e flexores do quadril associada ao equino, resultando em flexão simultânea de quadril e joelho; (III) equino aparente, no qual o tornozelo apresenta dorsiflexão normal, mas a flexão excessiva de quadril e joelho cria a impressão visual de equino; e (IV) marcha em agachamento, caracterizada por dorsiflexão excessiva do tornozelo combinada com flexão acentuada de joelho e quadril, frequentemente associada ao

alongamento inadequado do tendão do calcâneo em crianças mais jovens. Posteriormente, Rodda et al. (41) expandiram essa classificação para cinco grupos, acrescentando um quinto padrão: marcha assimétrica, na qual cada membro inferior se enquadra em um grupo distinto entre os quatro anteriores (por exemplo, um membro em equino aparente e o outro em marcha em salto). Esta classificação estabeleceu a base conceitual e biomecânica sobre a qual o GCS-CP foi desenvolvido.

A validade dos quatro padrões sagitais propostos por Rodda e Graham foi recentemente investigada de forma sistemática por Papageorgiou et al. (42), em estudo conduzido com 270 crianças com PC espástica do Hospital Universitário de Leuven. Utilizando mapeamento estatístico não paramétrico (*statistical non-parametric mapping*) sobre formas de onda cinemáticas contínuas, os autores confirmaram empiricamente a maior parte das regras de classificação originais do equino verdadeiro, marcha em salto, equino aparente e marcha agachada, refinando sua descrição cinemática no plano sagital e complementando-a com assinaturas cinéticas e com desvios nos planos coronal e transverso. Os autores reconheceram, contudo, que padrões cinematicamente vizinhos, particularmente marcha em salto, equino aparente e marcha agachada, compartilham o aumento da flexão do joelho e só podem ser distinguidos com precisão pela posição do tornozelo, limitação que motiva diretamente o uso de algoritmos de aprendizado de máquina capazes de integrar múltiplas variáveis articulares simultaneamente.

2.6.3 Sistema de Classificação da Marcha para Crianças com Paralisia Cerebral (GCS-CP)

O sistema de classificação desenvolvido por Davids e Bagley (GCS-CP), fundamentado nos trabalhos de Sutherland e Davids e de Rodda e Graham, organiza os desvios da marcha em crianças com PC de forma hierárquica, baseada na etiologia fisiopatológica ou biomecânica dos padrões observados (9,10). Este sistema parte do conceito de desvios primários versus compensatórios para identificar padrões comuns e suas causas. Embora a abordagem de Davids e Bagley reconheça desvios nos planos sagital, transverso e coronal, a classificação propriamente dita opera em dois planos: o sagital e o transverso, justamente os de maior relevância e reprodutibilidade na avaliação clínica da marcha, enquanto o plano coronal é observado, mas não compõe a classificação formal. No plano sagital, a classificação é organizada de forma hierárquica em dois níveis: o primeiro distingue a fase de apoio da fase de balanço, e o segundo, o padrão específico dentro de cada fase. Desse arranjo resultam três

componentes de classificação: sagital apoio, sagital balanço e transverso, definidos por critérios cinemáticos objetivos de quadril, joelho e tornozelo. O Quadro 1 mostra o sistema hierárquico completo; a Figura 4 apresenta o componente sagital (fases de apoio e balanço); e a Figura 5, o componente transverso. No presente estudo, os três componentes são abordados como variáveis-alvo independentes de classificação.

A limitação das classificações baseadas exclusivamente no plano sagital foi recentemente evidenciada por Morais Filho et al. (43), que avaliaram 532 pacientes com PC hemiplégica e demonstraram que 42,1% não puderam ser classificados pelo sistema de Winters, Gage e Hicks, identificando quatro padrões adicionais não contemplados pela classificação original. Este achado reforça que classificações uniplanares mostram-se insuficientes para capturar a heterogeneidade fenotípica da marcha na PC. Nesse contexto, o GCS-CP de Davids e Bagley representa um avanço ao hierarquizar e expandir as classificações existentes, incorporando componentes multiplanares: sagital apoio, sagital balanço e transverso, com critérios cinemáticos objetivos, tornando-o potencialmente mais robusto para a classificação automatizada. Essa utilidade clínica foi confirmada por Melanda et al. (14), que, ao avaliarem a aplicabilidade do GCS-CP em crianças com PC, demonstraram que o GCS-CP foi capaz de classificar 96,95% dos pacientes no componente sagital de apoio e 100% nos componentes transverso e sagital de balanço, demonstrando ainda validade e confiabilidade interavaliadores moderada, esta última quantificada pelo coeficiente Kappa, que corrige a concordância entre avaliadores pelo acaso (44,45).

Quadro 1 – Sistema de Classificação Hierárquica da Marcha de Davids e Bagley (GCS-CP)

PLANO / FASE	1º NÍVEL	2º NÍVEL	CARACTERÍSTICAS
SAGITAL Apoio	**Normal**	—	Padrão normal da marcha na fase de apoio
	Salto	Equino Verdadeiro	Flexão plantar do tornozelo; joelho/quadril normal, hiperestendido ou fletido
		Equino Aparente	Tornozelo neutro; flexão excessiva de joelho e/ou quadril
	Agachado	Compensado	Flexão do joelho com inclinação anterior do tronco; momento extensor diminuído
		Não Compensado	Flexão do joelho sem compensação do tronco; momento extensor aumentado
SAGITAL Balanço	**Normal**	—	Flexão adequada do joelho na fase de balanço (pico >45°)
	Rígido	Origem no Joelho	Perfil do quadril normal; espasticidade do reto femoral
		Origem no Quadril	Perfil do quadril anormal; fraqueza ou controle motor fraca
TRANSVERSO Apoio	**Normal**	—	Ângulo de progressão do pé alinhado com linha de progressão
	Transverso Interno	Nível Único	Desvio rotacional em um nível (pé, tíbia ou fêmur)
		Multinível	Desvios rotacionais em mais de um nível
	Transverso Externo	Nível Único	Desvio rotacional em um nível
		Multinível / Off-setting	Multinível ou compensação (anteversão femoral + torção tibial externa)

Fonte: *Adaptado de Davids JR, Bagley AM. J Am Acad Orthop Surg. 2014;22(12):782-90.*

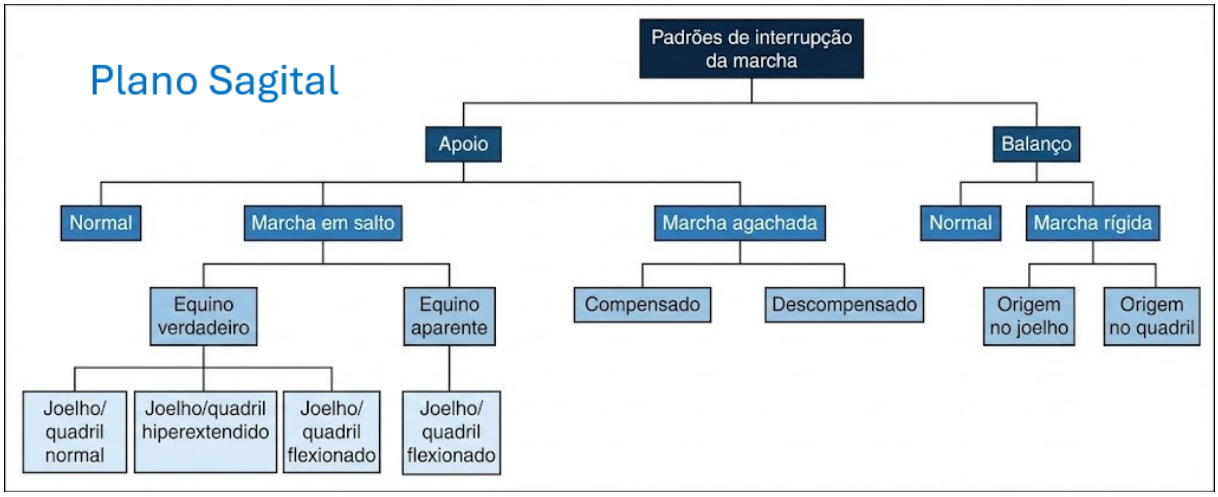

Figura 4 – Classificação hierárquica dos padrões comuns de perturbação da marcha no plano sagital, segundo o sistema GCS-CP.
Nota: Esta classificação hierárquica fundamentou a definição das variáveis alvo do presente estudo: Apoio Geral (normal, salto e agachado) e Balanço Geral (normal e rígido). Os subtipos do nível inferior da hierarquia não foram utilizados como classes de classificação devido ao tamanho amostral. O plano transverso (Transverso Geral: normal, transverso interno e transverso externo), também classificado pelo GCS-CP, não está representado nesta figura.
Fonte: *Elaborada pelo autor, adaptada e traduzida com base em Davids JR, Bagley AM. Identification of common gait disruption patterns in children with cerebral palsy. J Am Acad Orthop Surg. 2014;22(12):782-90.*

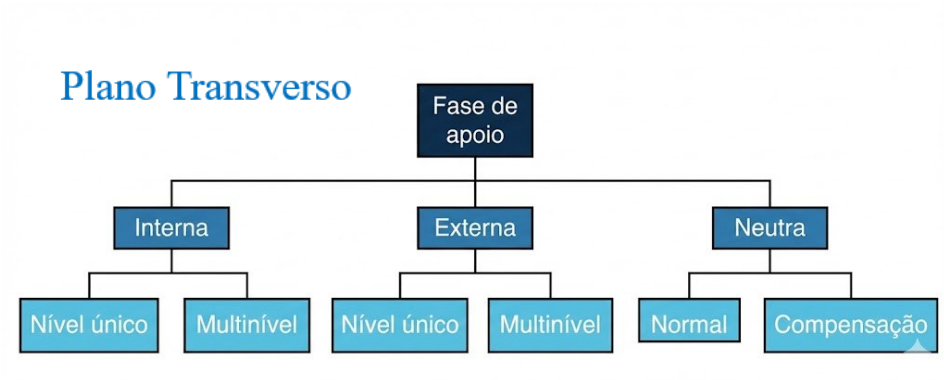

Figura 5 - Classificação hierárquica do componente Transverso do GCS-CP na fase de apoio.
Nota: O componente Transverso classifica o ângulo de progressão do pé na fase de apoio em três categorias: transverso interno, externa e neutra.
Fonte: *Elaborada pelo autor, adaptada e traduzida com base em Davids JR, Bagley AM. Identification of common gait disruption patterns in children with cerebral palsy. J Am Acad Orthop Surg. 2014;22(12):782-90.*

A compreensão dos padrões Salto e Agachado, os dois principais desvios da fase de apoio no GCS-CP, requer o entendimento do conceito de acoplamento flexão plantar-extensão do joelho, que descreve as inter-relações biomecânicas entre o complexo pé-tornozelo e o joelho (10). Durante a fase de apoio, um músculo gastrocnêmio-sóleo ativo, atuando sobre um pé estável e alinhado na linha de progressão, controla a progressão anterior da tíbia sobre o

pé plantígrado. Esse mecanismo direciona o vetor de reação do solo anteriormente ao eixo articular do joelho, gerando um momento extensor que reduz a demanda sobre o quadríceps (10). Quando o gastrocnêmio-sóleo está fraco ou excessivamente alongado, a tíbia avança sem controle adequado, deslocando o vetor de reação do solo posteriormente ao joelho e gerando um momento flexor, resultando em marcha agachada. Em contrapartida, quando o gastrocnêmio está espástico ou encurtado, ocorre flexão plantar excessiva do tornozelo (equino), com o vetor deslocado anteriormente, resultando em marcha em salto ou hiperextensão do joelho (10).

2.7 Dificuldade na Classificação e Confiabilidade

A classificação dos padrões da marcha na PC constitui um desafio clínico significativo devido à complexidade e multidimensionalidade dos dados envolvidos (11,14,46). A natureza contínua das anormalidades da marcha, com frequente sobreposição entre categorias diagnósticas, dificulta a categorização discreta em padrões mutuamente exclusivos (14,46). Além disso, muitos pacientes apresentam padrões mistos que combinam características de diferentes categorias, aumentando a complexidade da classificação (14,46).

Estudos de confiabilidade demonstram variabilidade significativa nas avaliações entre diferentes profissionais. Melanda et al. (14), avaliaram crianças com PC espástica nos níveis I-III do GMFCS e encontraram confiabilidade intra-avaliador alta, substancial a quase perfeita no primeiro nível (κ = 0,791-0,888) e substancial no segundo (κ = 0,714-0,776). Já entre avaliadores diferentes, a concordância caiu para níveis moderados (κ total = 0,596 no primeiro nível; κ total = 0,485 no segundo). A desagregação por plano mostrou que esse desempenho varia conforme o componente: o plano sagital em apoio foi o mais difícil (κ = 0,409), seguido pelo plano transverso (κ = 0,516) e pelo plano sagital em balanço (κ = 0,570). Esses valores são retomados como referência na Seção 5.8. A sobreposição entre os padrões de marcha, visível nas curvas cinemáticas médias dos grupos (Figura 6), ajuda a explicar essa variabilidade, pois não há fronteiras nítidas entre eles (14,46).). A necessidade de métodos mais objetivos e reprodutíveis justifica a investigação de abordagens baseadas em inteligência artificial (14,15,17).

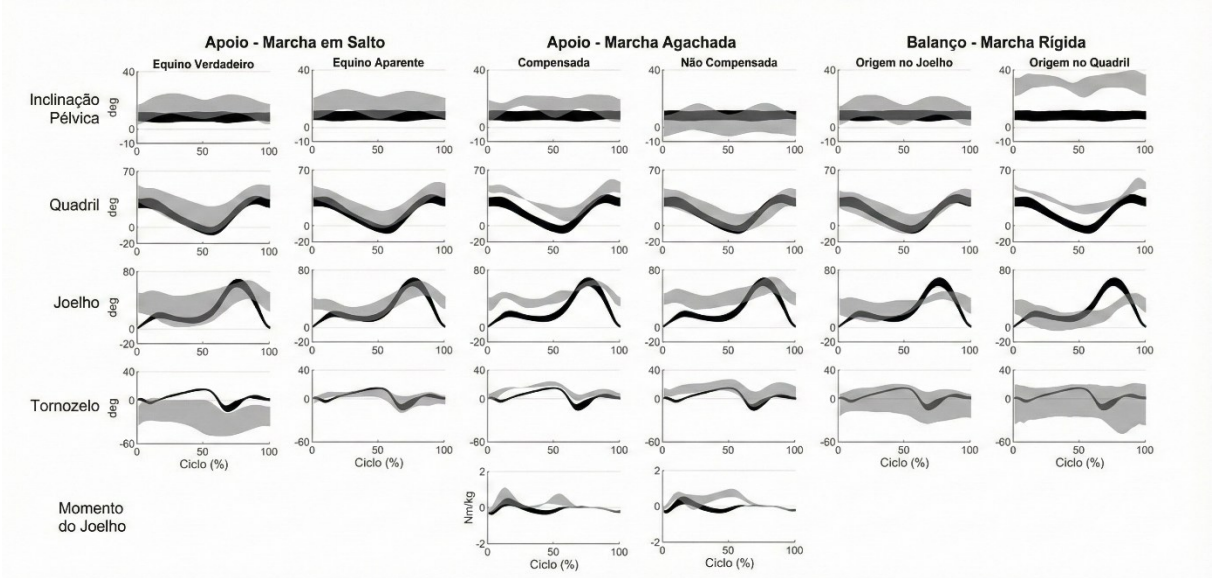

Figura 6 – Dados cinemáticos do plano sagital para os grupos da fase de apoio e balanço. Curvas resumidas para cada padrão da marcha da população do estudo. A faixa preta representa a média ± 1 desvio padrão do banco de dados normal, e a faixa cinza, a média ± 1 desvio padrão de cada padrão. O eixo vertical representa o deslocamento angular articular e o eixo horizontal, 100% do ciclo da marcha. Observa-se a progressão entre os padrões, com aumento gradual da dorsiflexão do tornozelo e da flexão do joelho do equino verdadeiro à marcha agachada.
Nota: Os dados cinemáticos apresentados nesta figura são provenientes do mesmo banco de dados utilizado no presente estudo. A sobreposição visual entre as curvas dos padrões normal e equino verdadeiro (marcha em salto) na fase de apoio ilustra a similaridade biomecânica entre essas classes, o que pode explicar a maior dificuldade de classificação observada para o alvo Apoio Geral em comparação aos demais alvos. Na fase de balanço, a distinção entre marcha normal e rígida é mais evidente no pico de flexão do joelho, consistente com o melhor desempenho dos modelos para o alvo Balanço Geral.
Fonte: *Adaptado e traduzido de Melanda AG, Davids JR, Pauleto AC, Pelegrinelli ARM, Ferreira AEK, Knaut LA, et al. Reliability and validity of the gait classification system in children with cerebral palsy (GCS-CP). Gait Posture. 2022;98:355-61. Reproduzida com permissão do autor.*

2.8 Inteligência Artificial na Classificação da Marcha

A aplicação de técnicas de inteligência artificial (IA) e ML na análise da marcha representa uma área de pesquisa em rápida expansão, com potencial para transformar a avaliação clínica de distúrbios do movimento (13,15–18,23). O interesse crescente no uso de ML no campo da análise clínica da marcha deve-se à sua capacidade de analisar grandes quantidades de dados de forma econômica, rápida e objetiva (15,17). A Figura 7 apresenta uma taxonomia das principais aplicações de inteligência computacional em estudos da marcha. Métodos de ML têm sido aplicados com sucesso para analisar padrões da marcha de pacientes com diversas condições, incluindo PC, esclerose múltipla e recuperação após acidente vascular cerebral (AVC) (15–18,23).

Figura 7 – Taxonomia das aplicações atuais e futuras de inteligência computacional em estudos da marcha. Em geral, a dificuldade de aplicação aumenta da esquerda para a direita, enquanto o design do sistema, análise de dados e complexidade computacional em cada subcategoria aumentam de cima para baixo.
Nota: O presente estudo insere-se na categoria "Detecção de Distúrbios da Marcha", especificamente na aplicação de algoritmos de aprendizado de máquina para classificação de padrões da marcha em crianças com paralisia cerebral.
Fonte: *Elaborada pelo autor, adaptada e traduzida com base em Lai DTH, Begg RK, Palaniswami M. Computational intelligence in gait research: a perspective on current applications and future challenges. IEEE Trans Inf Technol Biomed. 2009;13(5):687-702; e Prakash C, Kumar R, Mittal N. Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. Artif Intell Rev. 2018;49(1):1-40.*

2.8.1 Vantagens da IA na Prática Clínica

As principais vantagens da aplicação de IA na classificação da marcha incluem a objetividade, reprodutibilidade e consistência das classificações (15,17,23). Diferentemente da avaliação visual por especialistas humanos, algoritmos de ML produzem resultados idênticos quando aplicados aos mesmos dados de entrada, eliminando a variabilidade interavaliadores que compromete os métodos tradicionais (14,15). Além disso, sistemas de ML podem processar dados de alta dimensionalidade que seriam impossíveis de analisar manualmente, identificando padrões sutis e relações não lineares entre variáveis (16,17,23).

A capacidade de fornecer medidas de importância de características e, em alguns casos, explicações para as decisões de classificação, representa outra vantagem significativa (15,17), permitindo compreender quais variáveis contribuem para cada predição e potencialmente revelando achados clínicos novos sobre os padrões da marcha (15,17).

Apesar dessas vantagens, uma revisão sistemática recente de Samadi Kohnehshahri et al. (24) analisou 63 estudos sobre aplicação de ML na análise da marcha em pacientes com PC e AVC, identificando importantes desafios para a tradução clínica desses métodos. Entre os algoritmos supervisionados, redes neurais artificiais (18 estudos), máquinas de vetores de suporte (12 estudos) e Floresta Aleatória (8 estudos) foram os mais utilizados, enquanto *k-means* (19 estudos) e *fuzzy clustering* (6 estudos) predominaram entre os métodos não supervisionados (24). A Figura 8 apresenta a avaliação de qualidade metodológica dos estudos incluídos na revisão, demonstrando que aproximadamente 80% foram classificados como de qualidade média ou alta. No entanto, conforme ilustrado na Figura 9, a análise detalhada dos critérios de adequabilidade, viabilidade e confiabilidade mostrou que a relevância clínica permanece comprometida, principalmente devido à falta de justificativa para seleção de características e algoritmos, uso de *datasets* não representativos e ausência de interpretação clínica dos resultados (24).

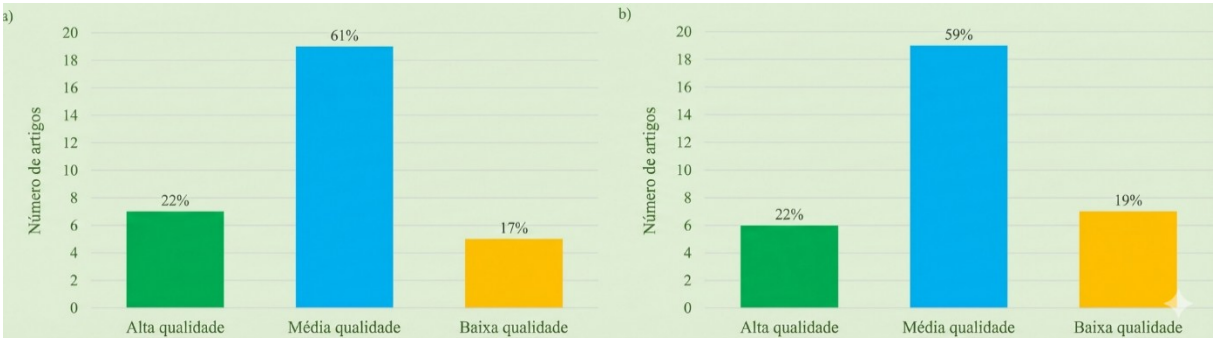

Figura 8 – Avaliação da qualidade metodológica. Estudos de aprendizado de máquina aplicados à análise da marcha em paralisia cerebral e AVC. (a) Estudos com métodos supervisionados (n=31): 22% alta qualidade, 61% qualidade média, 17% baixa qualidade. (b) Estudos com métodos não-supervisionados (n=32): 22% alta qualidade, 59% qualidade média, 19% baixa qualidade.
Nota: A maioria dos estudos analisados apresentou qualidade metodológica média, indicando lacunas importantes em aspectos como validação externa, reprodutibilidade e transparência na descrição dos métodos. O presente estudo buscou endereçar essas limitações por meio de estratégias como prevenção de vazamento de dados, validação cruzada estratificada por paciente e reporte detalhado de hiperparâmetros e métricas.
Fonte: *Elaborada pelo autor com base em Samadi Kohnehshahri F, Merlo A, Mazzoli D, Bò MC, Stagni R. Machine learning applied to gait analysis data in cerebral palsy and stroke: a systematic review. Gait Posture. 2024;111:105-121.*

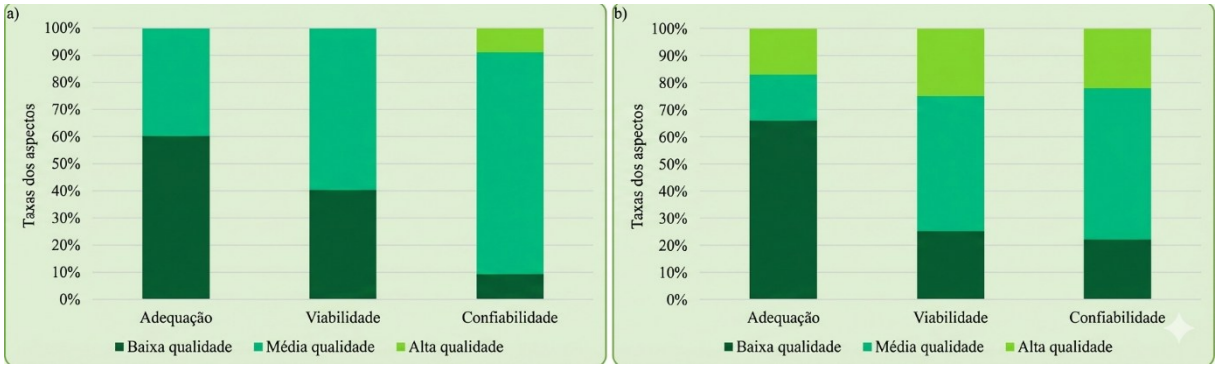

Figura 9 – Avaliação da qualidade em estudos de machine learning. Adequabilidade, viabilidade e confiabilidade. (a) Algoritmos supervisionados. (b) Algoritmos não supervisionados. A adequabilidade refere-se à adequação do algoritmo para atender aos requisitos da aplicação clínica; a viabilidade refere-se à praticidade dos métodos em ambientes clínicos reais; e a confiabilidade refere-se à consistência e robustez dos algoritmos quando aplicados aos dados específicos da marcha. Observa-se que a adequabilidade apresentou os menores escores em ambas as categorias, enquanto a confiabilidade obteve os maiores escores.
Nota: Os baixos escores de adequabilidade indicam que a maioria dos estudos não demonstrou suficientemente a adequação dos algoritmos às demandas clínicas específicas. O presente estudo buscou atender a esse aspecto ao utilizar como alvo de classificação um sistema clínico validado (GCS-CP) e ao avaliar o desempenho dos modelos com métricas clinicamente relevantes, como sensibilidade por classe e acurácia balanceada.
Fonte: *Elaborada pelo autor com base em Samadi Kohnehshahri F, Merlo A, Mazzoli D, Bò MC, Stagni R. Machine learning applied to gait analysis data in cerebral palsy and stroke: a systematic review. Gait Posture. 2024;111:105-121.*

2.8.2 Aplicações em Paralisia Cerebral

Diversos estudos têm investigado a aplicação de ML para a classificação de padrões da marcha em PC, conforme resumido na Tabela 7, que apresenta uma síntese cronológica dos principais estudos com seus respectivos tamanhos amostrais, populações estudadas e desempenho dos algoritmos.

Tabela 7 – Estudos de Aprendizado de Máquina na Classificação da Marcha em PC

Estudo	n	População	Classificação Alvo	Dados Utilizados	Melhor Algoritmo	Acurácia
Kamruzzaman & Begg (19)	156	PC diplegia espástica (88) + controles (68); 2-20 anos	PC vs. normal	Comprimento da passada + cadência	SVM	96,8%
Zhang & Ma (17)	200	PC diplegia espástica; 6-12 anos	4 padrões sagitais	Cinemática sagital	ANN	93,5%
Darbandi et al. (20)	132*	PC espástica; 5-19 anos	4 padrões (Rodda)	Cinemática sagital	Fuzzy	94%
Slijepcevic et al. (15)	302	PC espástica; crianças	4 padrões sagitais	Cinemática sagital (joelho/tornozelo)	RF	93,4%
Yoon et al. (47)	833	PC (333) + controles (500); GMFCS I-III; 5-65 anos	Níveis GMFCS	Cinemática multivariada (3D)	MV SFLDA	62-93%**

Nota: *132 membros de 66 pacientes (análise bilateral)

**93% para PC vs. controles; 62-67% para discriminação entre níveis GMFCS adjacentes

Fonte: *Elaborada pelo autor com base em Kamruzzaman & Begg* (19)*, Zhang & Ma* (17)*, Darbandi et al.* (20)*, Slijepcevic et al.* (15) *e Yoon et al.* (47)*.*

Kamruzzaman e Begg (19) foram pioneiros na aplicação de SVM para diferenciar crianças com PC de controles saudáveis, alcançando 96,8% de acurácia em uma amostra de 156 crianças utilizando apenas comprimento da passada e cadência normalizados. Zhang e Ma (17) compararam sete algoritmos de ML para classificar quatro padrões da marcha sagital (equino verdadeiro, marcha em salto, equino aparente e marcha agachada) em 200 crianças com diplegia espástica entre 6 e 12 anos, alcançando precisão de 93,5% com redes neurais artificiais (17). A Figura 10 ilustra as sete características cinemáticas extraídas do ciclo da marcha adotadas no presente estudo, adaptadas de Zhang e Ma (17), com dados dos participantes desta pesquisa.

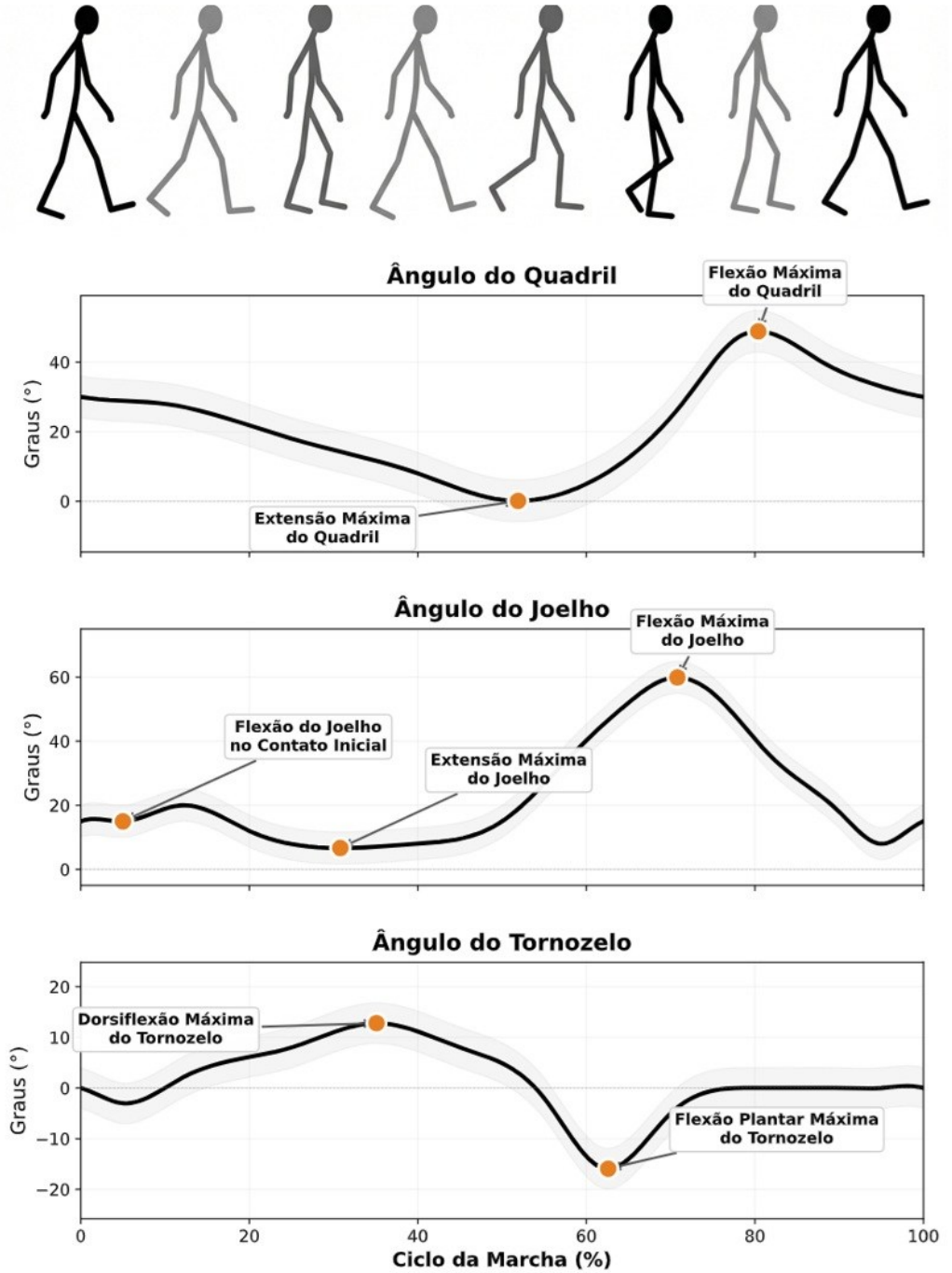

Figura 10 – Características cinemáticas extraídas do ciclo da marcha. Características cinemáticas extraídas do ciclo da marcha para classificação de padrões da marcha em crianças com paralisia cerebral. Os gráficos mostram os ângulos articulares do quadril, joelho e tornozelo no plano sagital ao longo do ciclo da marcha (%). As características incluem: extensão máxima do quadril, flexão máxima do quadril, flexão do joelho no contato inicial, flexão máxima do joelho, extensão máxima do joelho, flexão plantar máxima do tornozelo e dorsiflexão máxima do tornozelo.
Nota: "As curvas representam os valores médios dos participantes classificados com marcha normal (SN, n=553) no presente estudo. Por se tratar de crianças com paralisia cerebral, os valores diferem de referências normativas de crianças com desenvolvimento típico – por exemplo, a extensão máxima do quadril (0,4° ± 14,7°) indica limitação da extensão, mesmo nos padrões classificados como normais pelo GCS-CP. Estas sete características integram o conjunto de variáveis cinemáticas originais extraídas da análise tridimensional da marcha. Características definidas conforme Zhang e Ma (17).
Fonte: *Elaborada pelo autor.*

Darbandi et al. (20) desenvolveram um sistema de classificação *fuzzy* aplicado aos padrões de Rodda, traduzindo a experiência clínica em regras objetivas (n e acurácia na Tabela 7). A lógica fuzzy, ao contrário da lógica binária clássica, permite que um caso pertença parcialmente a mais de uma categoria, o que é adequado para padrões de marcha com sobreposição clínica. Slijepcevic et al. (15) demonstraram que dados cinemáticos superaram significativamente os dados de força de reação do solo na classificação dos padrões sagitais (93,4% vs 42,0% com Floresta Aleatória), e que abordagens tradicionais de ML alcançaram resultados superiores às redes neurais profundas, focando-se mais em regiões clinicamente relevantes.

Yoon et al. (47) conduziram o maior estudo até o momento, conforme detalhado na Tabela 7, avaliando níveis GMFCS através de análise discriminante linear funcional esparsa multivariada (MV SFLDA), técnica que trata os dados cinemáticos como funções contínuas e identifica apenas as fases do ciclo da marcha que contribuem para distinguir os níveis de gravidade. A discriminação entre PC e controles foi alta, mas caiu substancialmente entre níveis GMFCS adjacentes, sendo a classificação multiclasse a mais desafiadora.

Nota-se que os estudos apresentam heterogeneidade metodológica considerável, variando em tamanho amostral (117 a 833 participantes), faixa etária, sistemas de classificação utilizados (Rodda vs. GMFCS) e tipos de dados de entrada, o que dificulta comparações diretas de desempenho entre os algoritmos. Esta limitação já havia sido identificada por Dobson et al. (46) em revisão sistemática anterior sobre classificações da marcha em PC, que apontou qualidade metodológica variável entre os estudos e definições inadequadas de amostras e métodos de amostragem.

2.9 Algoritmos de Aprendizado de Máquina para Classificação da Marcha

A seleção do algoritmo de aprendizado de máquina mais apropriado para a classificação de padrões da marcha depende de múltiplos fatores, incluindo a natureza e dimensionalidade dos dados, o tamanho do conjunto de dados disponível, os recursos computacionais e a importância da interpretação para a aplicação clínica (15,17,20–22). Cada família de algoritmos apresenta vantagens e limitações específicas que devem ser consideradas no contexto da aplicação pretendida.

2.9.1 Máquinas de Vetores de Suporte ou "*Support Vector Machine"* (SVM)

As Máquinas de Vetores de Suporte são classificadores não probabilísticos que buscam construir um hiperplano de separação ótimo que maximize a margem entre classes em um espaço de características de alta dimensão (19). Utilizando funções *kernel*, as SVMs podem modelar relações não lineares através do mapeamento implícito dos dados para espaços de dimensão superior onde a separação linear é possível (19). Kamruzzaman e Begg (19) demonstraram que SVMs podem classificar crianças com PC com 96,8% de acurácia utilizando apenas comprimento da passada e cadência normalizados. Nesse estudo, a normalização dos parâmetros da marcha mostrou-se fundamental, elevando a acurácia de 83,3% (dados brutos) para 96,8% (dados normalizados). Os autores também compararam diferentes funções *kernel*, demonstrando que os *kernel*s polynomial e RBF apresentaram desempenho similar e superior aos *kernel*s linear e ANOVA *spline* (19).

2.9.2 Redes Neurais (ANN)

As Redes Neurais Artificiais (ANNs) são modelos computacionais inspirados na estrutura e função das redes neurais biológicas, compostos por camadas de neurônios interconectados que processam informação através de transformações não lineares (15,17). A arquitetura básica de uma ANN para classificação da marcha (Figura 11) consiste em três componentes: uma camada de entrada que recebe os dados cinemáticos obtidos pela análise 3D da marcha, camadas ocultas que extraem automaticamente padrões e características relevantes dos dados, e uma camada de saída que fornece a classificação do padrão de marcha. As conexões entre os neurônios possuem pesos sinápticos que são ajustados durante o treinamento para minimizar o erro de classificação. Zhang e Ma (17) reportaram que ANNs alcançaram a maior precisão (93,5%) entre os sete algoritmos testados para classificação da marcha em PC. A principal limitação é a natureza de 'caixa preta' que dificulta a interpretação das decisões (15). A arquitetura *feedforward* descrita acima, composta por camada de entrada, camadas ocultas e camada de saída totalmente conectadas, corresponde ao *Perceptron* Multicamadas (MLP), arquitetura de ANN empregada no presente estudo (Seção 3.7).

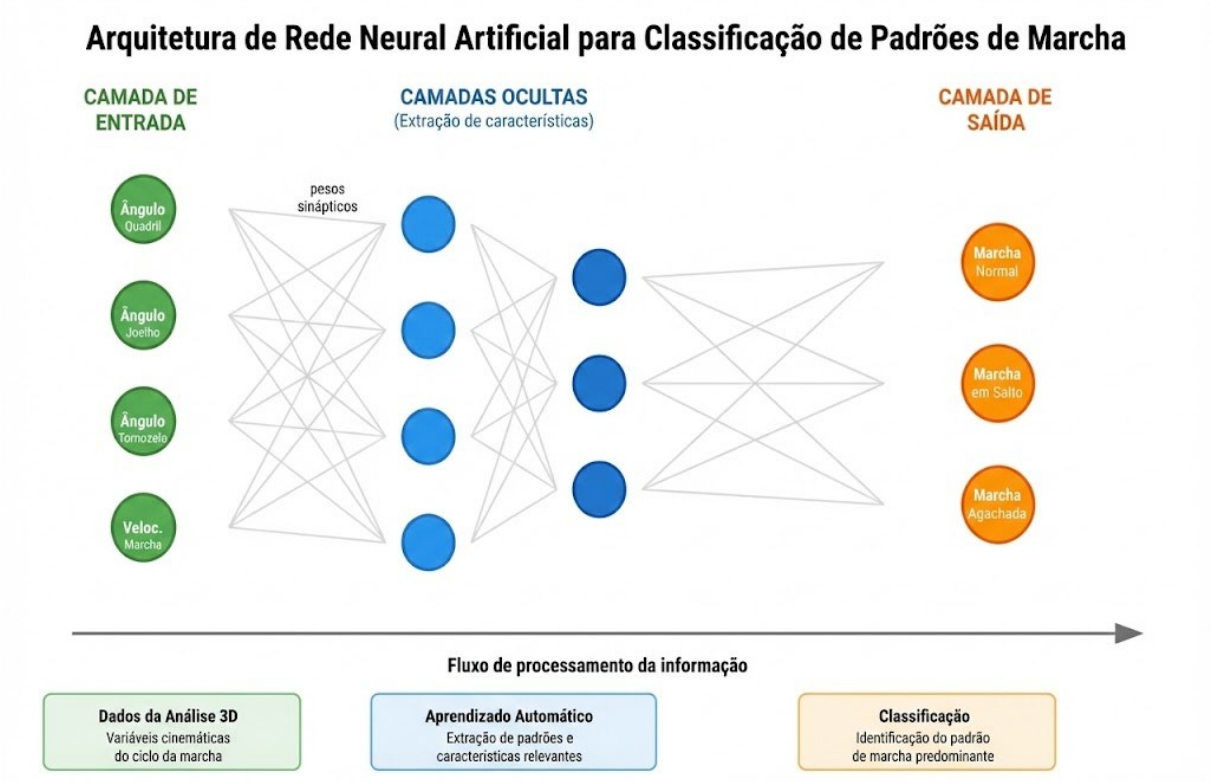

Figura 11 – Arquitetura básica de uma Rede Neural Artificial para classificação de padrões da marcha na paralisia cerebral. A camada de entrada (verde) recebe dados cinemáticos da análise tridimensional da marcha. As camadas ocultas (azul) processam as informações através de pesos sinápticos ajustáveis, extraindo características relevantes. A camada de saída (laranja) fornece a classificação final entre marcha normal, marcha em salto ou marcha agachada.
Fonte: *Elaborada pelo autor.*

2.9.3 Floresta Aleatória (*Random Forest*)

As Florestas Aleatórias são classificadores de conjunto que consistem em uma coleção de múltiplas árvores de decisão, cada uma treinada em diferentes subconjuntos dos dados (15,17). Slijepcevic et al. (15) demonstraram que Florestas Aleatórias superaram todos os outros modelos testados, incluindo redes neurais profundas, para classificação de padrões da marcha na PC, alcançando precisão de 93,4% com dados cinemáticos sagitais do joelho e tornozelo.

2.9.4 *Gradient Boosting* (LightGBM e XGBoost)

Os métodos de *Gradient Boosting Decision Tree* (GBDT) constroem o modelo de forma aditiva e sequencial, em que cada nova árvore é treinada para corrigir os erros residuais das anteriores (21,22). A Tabela 8 apresenta um comparativo dos principais algoritmos de ML utilizados na classificação da marcha. XGBoost (*eXtreme Gradient Boosting*) e LightGBM (*Light Gradient Boosting Machine*) representam implementações altamente otimizadas destes algoritmos, alcançando estado da arte em muitos problemas de aprendizado de máquina tabular (21,22). Chen e Guestrin (22) demonstraram que XGBoost supera outros métodos em diversos *benchmarks* de classificação. Ke et al. (21) mostraram que LightGBM pode ser mais de 20 vezes mais rápido que implementações convencionais de GBDT mantendo acurácia equivalente.

Tabela 8 – Comparativo de Algoritmos de aprendizado de máquina

Algoritmo	Vantagens	Desvantagens	Acurácia em PC
Naive Bayes	Simples, rápido, bom com dados pequenos	Assume independência de características; desempenho variável	~78-85%
Árvores de decisão	Alta interpretabilidade; captura não-linearidades	Propenso a *overfitting*; instável a variações nos dados	~85%
SVM	Eficaz em alta dimensionalidade; boa generalização	Sensível a *kernel*/hiperparâmetros; adaptação para multiclasse	84-97%
ANN	Aprende representações; captura relações complexas	Caixa preta; necessita muitos dados; ajuste complexo	84-93,5%
Florestas aleatórias	Robusta ao *overfitting*; importância de atributos	Menos interpretável que árvore única	83-93%
LightGBM/XGBoost	Alta acurácia; eficiente; lida com *missing values*	Risco de *overfitting*; interpretabilidade limitada	Alto potencial
Fuzzy Clustering	Traduz experiência clínica; resultados interpretáveis	Requer definição manual de regras	98%

Nota: As acurácias reportadas referem-se a tarefas de classificação distintas, detecção de PC versus controle saudável (Kamruzzaman e Begg), classificação de padrões sagitais segundo Rodda (Zhang & Ma) e classificação de padrões por agrupamento fuzzy (Darbandi et al.), entre outras, não sendo diretamente comparáveis entre si. Os valores devem ser interpretados como indicativos do desempenho relatado em cada contexto, e não como uma comparação controlada entre algoritmos.
Fonte: *Elaborada pelo autor com base em Zhang & Ma* (17)*, Slijepcevic et al.* (15)*, Darbandi et al.* (20)*, Kamruzzaman e Begg* (19)*, Ke et al.* (21) *e Chen & Guestrin* (22)*.*

2.9.5 Modelos de Fusão

Os modelos de fusão, ou *ensembles*, combinam múltiplos classificadores individuais para produzir uma predição final mais robusta do que qualquer modelo isolado. A combinação de classificadores tem demonstrado melhorar o desempenho em relação a modelos individuais, fundamentando-se na diversidade dos algoritmos (48). Dietterich (27) demonstrou que métodos de fusão superam classificadores individuais por três razões fundamentais: estatística (redução da variância ao combinar múltiplas hipóteses), computacional (mitigação de mínimos locais em algoritmos de otimização) e representacional (expansão do espaço de hipóteses além do alcance de modelos individuais).

A premissa fundamental é que diferentes algoritmos podem capturar aspectos complementares dos dados, e erros cometidos por classificadores individuais tendem a se cancelar quando suas predições são agregadas (27,28). Zhou (28) formalizou que a eficácia das fusão depende de dois requisitos: os classificadores base devem ser acurados (desempenho superior ao aleatório) e diversos (cometendo erros em diferentes subconjuntos dos dados). A Figura 12 ilustra a arquitetura dos três métodos de fusão empregados neste estudo.

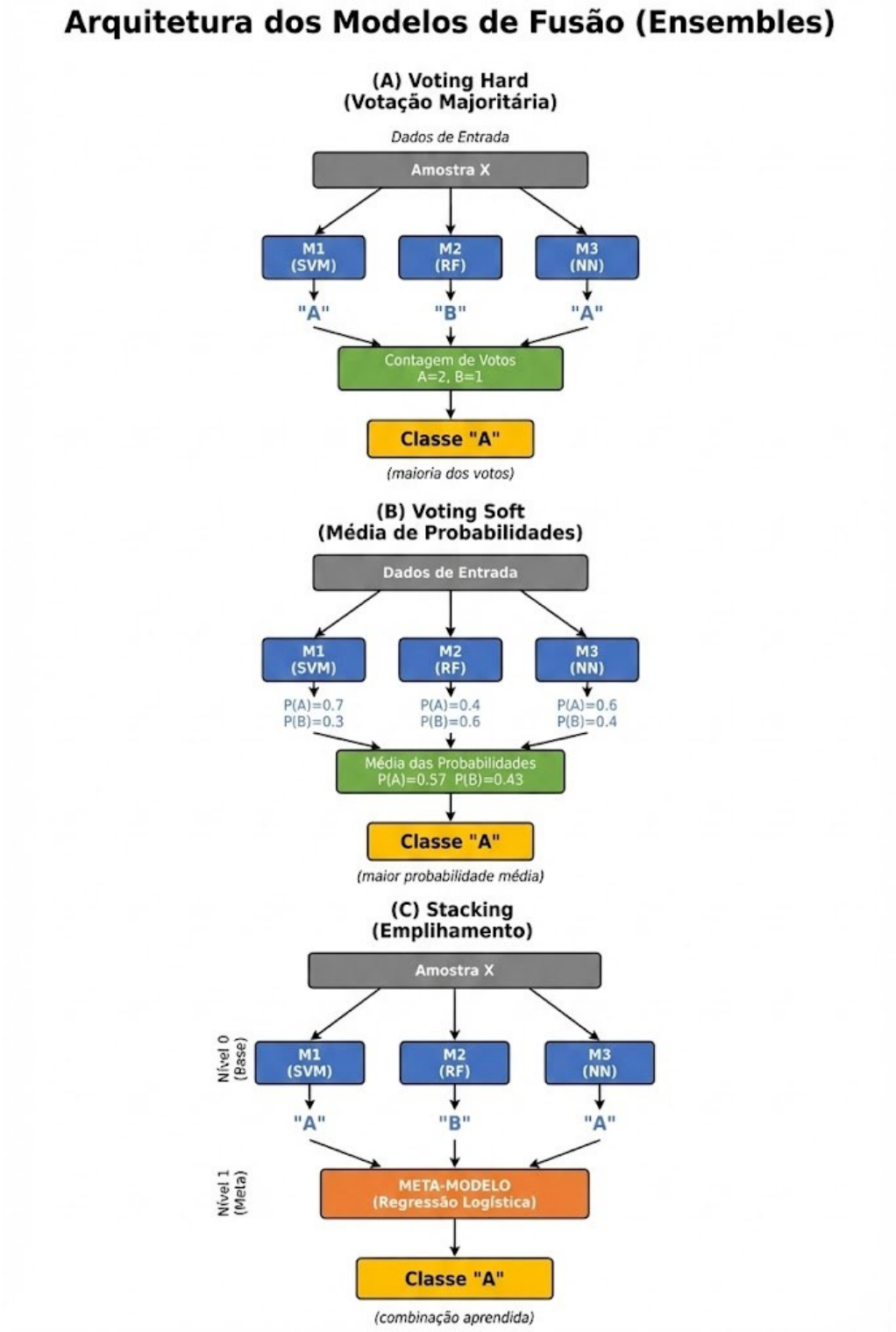

Figura 12 – Arquitetura dos modelos de fusão. (A) *Voting Hard*: cada classificador emite um voto categórico e a classe majoritária é selecionada. (B) *Voting Soft*: cada classificador fornece probabilidades e a média determina a predição. (C) *Stacking*: as predições dos classificadores base (Nível 0) alimentam um meta-modelo (Nível 1) que aprende a combiná-las. M1 = SVM; M2 = Floresta Aleatória; M3 = Rede Neural.
Fonte: *Elaborada pelo autor.*

Voting (Votação): O método de votação combina as predições de múltiplos classificadores através de regras de agregação (48). No *Voting Hard* (votação majoritária), cada classificador emite um voto para uma classe, e a classe com maior número de votos é selecionada como predição final. No *Voting Soft*, cada classificador fornece probabilidades para cada classe, e a predição final é baseada na média dessas probabilidades, favorecendo classificadores com maior confiança em suas predições (28,48). Kittler et al. (48) compararam diversas regras de combinação e demonstraram que a regra da soma (equivalente ao *Voting Soft*) apresenta maior robustez a erros de estimação de probabilidades. A análise de sensibilidade demonstrou que, enquanto a regra do produto amplifica erros de estimação, a regra da soma atenua esses erros através da agregação, o que explica seu desempenho consistentemente superior em dados reais.

Stacking (Empilhamento): O *Stacking*, proposto originalmente por Wolpert (49), representa uma abordagem mais sofisticada utilizando uma arquitetura de duas camadas. Na primeira camada (nível-0), múltiplos classificadores base geram predições a partir dos dados originais. Na segunda camada (nível-1), um meta-modelo aprende a combinar otimamente as predições dos classificadores base, identificando em quais situações confiar mais em cada modelo (28,49). A principal vantagem do *Stacking* sobre métodos de votação é sua capacidade de aprender pesos não-uniformes e relações não-lineares entre as predições dos classificadores base (49). Tipicamente, utiliza-se um algoritmo mais simples e regularizado como meta-modelo (como Regressão Logística) para evitar *overfitting* na segunda camada (28). A validação cruzada é empregada para gerar as predições do nível-0 que servirão de entrada para o meta-modelo, prevenindo vazamento de dados (28,49).

Em conjunto, esses avanços consolidam o aprendizado de máquina como uma alternativa robusta para a classificação automatizada de padrões da marcha em paralisia cerebral, com algoritmos individuais (notadamente Floresta Aleatória, SVM e Redes Neurais) e modelos de fusão (*Voting* e *Stacking*) demonstrando viabilidade técnica e potencial clínico (15,17,20,24). Esses trabalhos forneceram os fundamentos metodológicos sobre os quais o presente estudo se apoia, ao mesmo tempo em que evidenciam as lacunas que motivam a investigação conduzida, em especial a necessidade de análise por classe, a comparação sistemática entre estratégias de pré-processamento e a exploração de estratégias de otimização orientadas ao contexto clínico.

3 MATERIAIS E MÉTODOS

3.1 Desenho do Estudo e Aspectos Éticos

Este é um estudo transversal retrospectivo, desenvolvido a partir de dados do banco do Laboratório de Análise da Marcha do Centro de Reabilitação do Hospital Ana Carolina Moura Xavier, Curitiba, Brasil. O estudo foi aprovado pelo Comitê de Ética em Pesquisa local (CAAE: 2.447.001), seguindo as normas regulamentadoras brasileiras para pesquisa envolvendo seres humanos.

O desfecho de interesse deste estudo é a classificação automática dos padrões da marcha segundo o GCS-CP, organizada em três variáveis-alvo independentes: (I) Apoio Geral – três classes: Normal (SN), Salto (SS) e Agachado (SA); (II) Balanço Geral – duas classes: Normal (BN) e Rígido (BR); e (III) Transverso Geral – três classes: Normal (TN), Transverso Interno (TI) e Transverso Externo (TE). Todas as classificações foram atribuídas por dois ortopedistas com experiência em análise clínica da marcha, com base nos critérios cinemáticos definidos pelo GCS-CP (ângulos articulares de quadril, joelho e tornozelo ao longo do ciclo da marcha), conforme protocolo descrito em Melanda et al. (14). Em casos de divergência entre avaliadores, o consenso foi estabelecido por discussão clínica. Os detalhes da distribuição das classes são apresentados na seção 3.5.

O protocolo metodológico foi baseado no Sistema de Classificação da Marcha para Crianças com Paralisia Cerebral (GCS-CP) proposto por Davids e Bagley (9) e validado por Melanda et al. (14). O GCS-CP categoriza os padrões da marcha em dois planos de movimento: sagital (subdividido em fase de apoio e fase de balanço) e transverso, utilizando dados cinemáticos obtidos por análise tridimensional da marcha.

3.2 Participantes

A amostra inicial foi composta por 131 crianças e adolescentes com diagnóstico de Paralisia Cerebral espástica, provenientes do banco de dados de um laboratório de análise do movimento utilizado no estudo de validação do GCS-CP de Melanda et al. (14). Os critérios de inclusão foram: idade entre 5 a 19 anos, classificação funcional GMFCS níveis I a III, e capacidade de deambulação independente ou com dispositivo auxiliar. Foram excluídos

pacientes submetidos a procedimentos cirúrgicos nos 12 meses anteriores à avaliação ou que receberam aplicação de toxina botulínica tipo A nos 6 meses precedentes. O fluxograma de seleção dos participantes e divisão dos dados é apresentado na Figura 13.

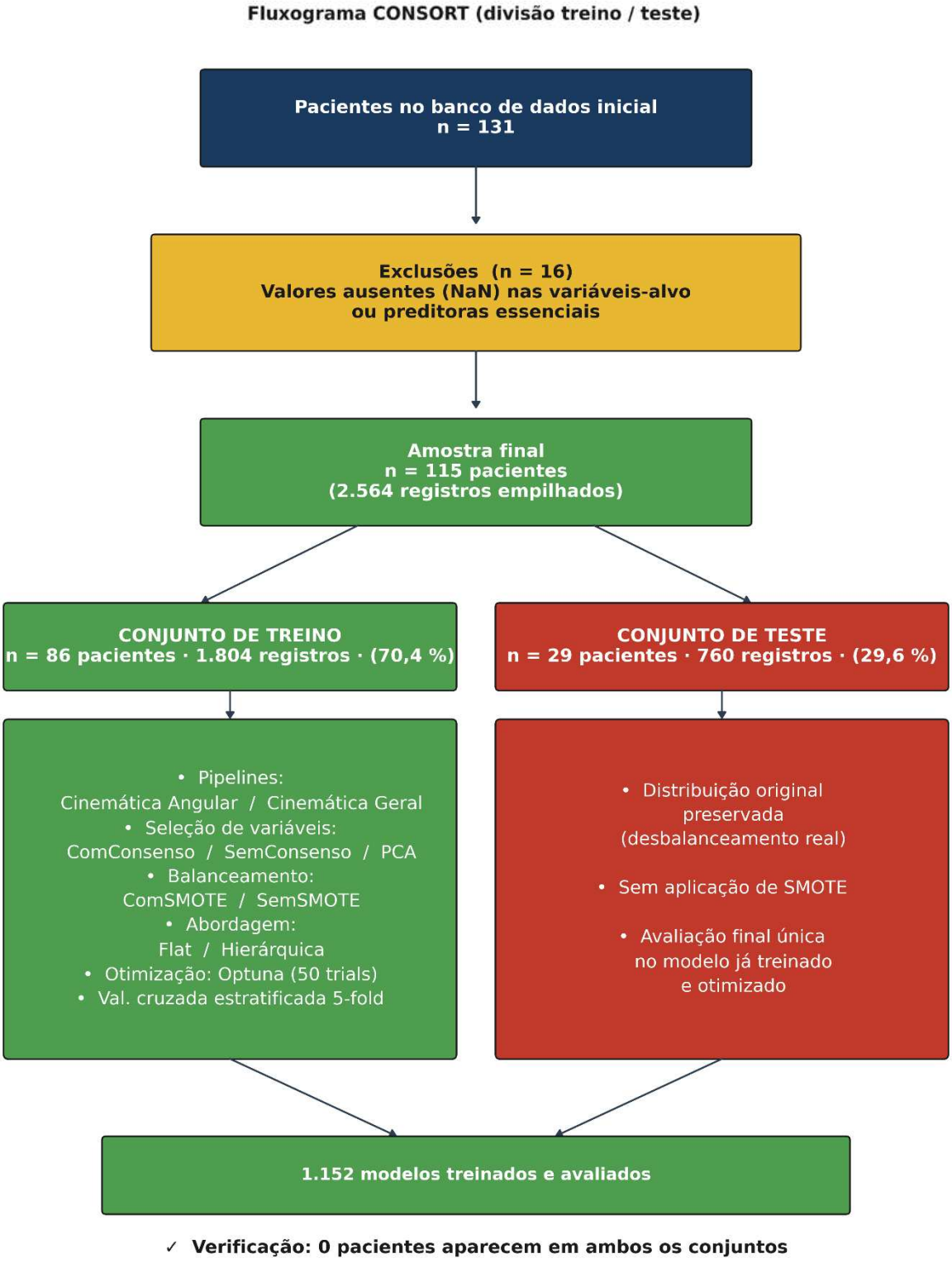

Figura 13 – Fluxograma de seleção dos participantes e divisão dos dados
Nota: A divisão entre conjuntos de treino e teste foi realizada por paciente, garantindo que nenhum participante aparecesse simultaneamente em ambos os conjuntos, prevenindo vazamento de dados. A técnica de sobreamostragem sintética da minoria (SMOTE) foi aplicada exclusivamente no conjunto de treino.
Fonte: *Elaborada pelo autor.*

Todos os participantes foram submetidos à análise tridimensional da marcha (3DGA) em condição descalça, com velocidade auto-selecionada, durante visita única ao laboratório de análise do movimento. O sistema de captura de movimento consistiu em 6 câmeras digitais de análise do movimento (Motion Analysis Corporation, Santa Rosa, CA) operando a 60 Hz, sincronizadas com 2 plataformas de força (AMTI Biomechanics Force Plate OR6-7 2000, Watertown, MA, USA) com frequência de amostragem de 1000 Hz. Utilizou-se o modelo biomecânico convencional da marcha (Helen-Hayes) com 26 marcadores reflexivos aplicados bilateralmente para captura de movimentos da pelve, quadril, joelho e tornozelo. O software Cortex e OrthoTrack 6.5.1 foram empregados para aquisição, processamento e cálculo da cinemática e dos momentos articulares.

Foram coletados dados de um mínimo de 7 ciclos da marcha para cada membro inferior avaliado, bilateralmente, com registro de vídeo sincronizado nas vistas sagital e coronal. As variáveis calculadas incluíram parâmetros espaço-temporais (comprimento do passo, comprimento da passada, cadência e velocidade de marcha), dados cinemáticos (ângulos articulares da pelve, quadril, joelho, tornozelo e ângulo de progressão do pé nos planos sagital, coronal e transverso), e dados cinéticos (momentos internos articulares do joelho no plano sagital). Para participantes classificados como GMFCS III que utilizaram dispositivos auxiliares da marcha (andadores, muletas ou bengalas), não foram coletados dados cinéticos. As posições das câmeras foram ajustadas para garantir que cada marcador fosse registrado por pelo menos duas câmeras em todos os momentos da coleta. As variáveis efetivamente submetidas ao *pipeline* de aprendizado de máquina, após aplicação da Regra Ipsilateral descrita na seção 3.3, estão detalhadas no Apêndice A.

3.3 Base de Dados e Estratégia de Empilhamento

Após o pré-processamento dos dados para o fluxo de processamento de aprendizado de máquina (*pipeline)*, foram excluídos 16 pacientes com valores ausentes (NaN) nas variáveis alvo ou variáveis preditoras essenciais, resultando em uma amostra final de 115 pacientes.

A base de dados original continha 1.282 registros, correspondentes a múltiplas tentativas de marcha por paciente. Cada registro consolida as variáveis cinemáticas e espaço-temporais de uma tentativa de marcha bilateral, calculadas a partir de no mínimo 7 ciclos da marcha coletados por membro inferior, conforme protocolo descrito na seção 3.2. O

número de tentativas variou entre os participantes, refletindo a prática clínica do laboratório onde múltiplas coletas são realizadas para garantir representatividade dos dados. Para maximizar a utilização dos dados, foi realizado um processo de empilhamento (*stacking*), tratando cada lado do corpo como uma observação independente, abordagem também adotada por Choisne et al. (50) em modelagem da marcha em crianças com PC, justificada pelo envolvimento bilateral assimétrico característico desta condição. Na diplegia, cada paciente contribuiu com múltiplas observações bilaterais. Na hemiplegia, incluiu-se ambos os lados: o afetado com classificação patológica e o contralateral com classificações normais nos três componentes do GCS-CP (Transverso Normal, Balanço Normal e Apoio Normal). A Tabela 9 apresenta a comparação entre as bases.

Tabela 9 – Comparação entre a base de dados original e empilhada

Característica	Original	Empilhada
Número de registros	1.282	2.564
Número de variáveis originais	70	70
Variáveis após Regra Ipsilateral	—	18 (Angular) / 25 (Geral)
Pacientes diplégicos (2 lados)	1.202	2.404
Pacientes hemiplégicos (2 lados)	80	160
Aumento de dados	—	+100%

Nota: Cada tentativa de marcha foi desdobrada em dois registros independentes após a aplicação do empilhamento bilateral (um para o lado direito e outro para o esquerdo), resultando na duplicação do número total de observações. Dos 2.564 registros empilhados, 2.404 (93,8%) correspondem a pacientes diplégicos e 160 (6,2%) a hemiplégicos. A amostra de 115 pacientes únicos é composta por 107 indivíduos diplégicos e 8 hemiplégicos (7 com hemiplegia direita e 1 com hemiplegia esquerda). Pacientes diplégicos contribuíram com classificações em todas as categorias do GCS-CP, incluindo padrões normais da marcha, refletindo a heterogeneidade funcional dentro desse grupo. A aplicação da Regra Ipsilateral (detalhada a seguir e no Apêndice A) limitou cada registro a 18 variáveis cinemáticas angulares (*Pipeline* Cinemática Angular) ou 25 variáveis com adição dos parâmetros espaço-temporais ipsilaterais (*Pipeline* Cinemática Geral).
Fonte: *Elaborada pelo autor.*

Cada membro inferior (direito e esquerdo) foi tratado como uma observação independente no *pipeline*, com sua própria variável-alvo correspondente ao padrão de marcha ipsilateral. Dessa forma, um paciente com diplegia espástica contribuiu com duas observações independentes: uma para o membro direito e outra para o esquerdo, cada qual com sua respectiva classificação GCS-CP.

Para garantir coerência metodológica com o princípio de classificação individual por membro do GCS-CP, foi aplicada a Regra Ipsilateral: cada observação utiliza

exclusivamente as variáveis cinemáticas do membro avaliado naquele ciclo, sem incluir variáveis do membro contralateral. Quando a linha corresponde ao membro direito, somente as variáveis do lado direito são utilizadas como preditoras, sendo o processo análogo para o membro esquerdo.

Esta decisão se justifica por três motivos. O primeiro, metodológico: a inclusão de variáveis contralaterais introduz multicolinearidade estrutural, pois os dois lados de um mesmo paciente são biologicamente correlacionados. O segundo, clínico: o GCS-CP classifica cada membro com base em sua própria cinemática, a classificação do membro direito não é definida pelo comportamento do esquerdo. O terceiro, empírico: a análise de concordância entre as classificações GCS-CP do lado direito e esquerdo, conduzida pelos autores nos próprios dados deste estudo, revelou assimetria substancial mesmo em pacientes diplégicos (n=107), grupo tradicionalmente presumido como simétrico, sendo que 26,2% apresentaram classe majoritária distinta entre os lados no alvo Transverso, 18,7% no alvo Apoio e 11,2% no alvo Balanço. Em pacientes hemiplégicos (n=8), as taxas de assimetria foram, como esperado, ainda maiores nos planos sagital de apoio (62,5%) e transverso (37,5%). Este achado demonstra que a classificação GCS-CP apresenta heterogeneidade bilateral clinicamente relevante, validando empiricamente o tratamento de cada lado como observação independente. A aplicação da Regra Ipsilateral resultou em dois conjuntos de atributos por registro: 18 variáveis cinemáticas angulares (*Pipeline* Cinemática Angular) e 25 variáveis com adição dos parâmetros espaço-temporais ipsilaterais (*Pipeline* Cinemática Geral), conforme detalhado no Apêndice A.

A Figura 14 apresenta o fluxograma completo do *pipeline* adotado neste estudo, ilustrando a sequência de etapas desde o banco de dados original até a avaliação final no conjunto de teste, com destaque para a separação rigorosa entre os dados de treinamento e teste em todas as etapas de pré-processamento.

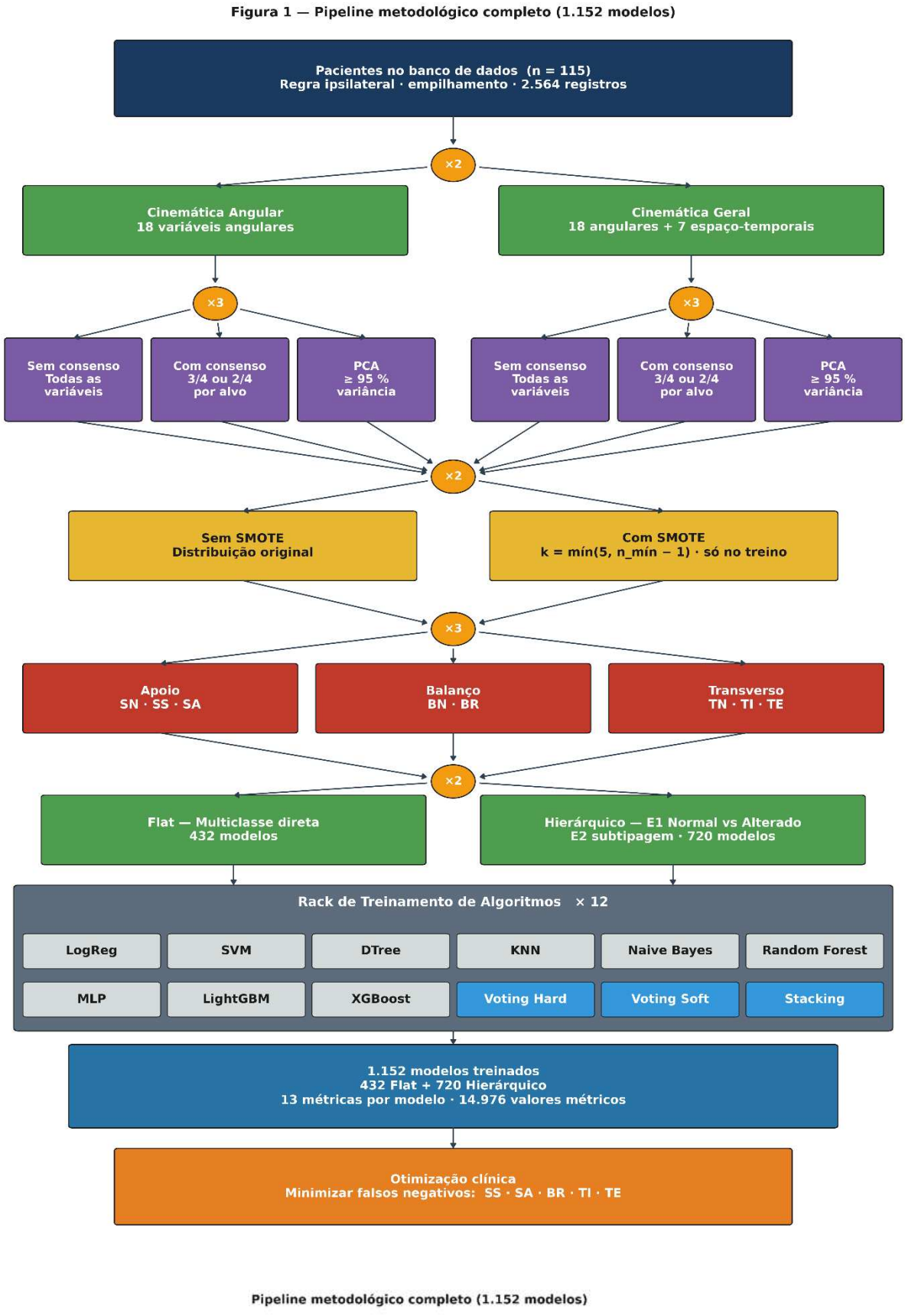

Figura 14 - Fluxograma do pipeline de aprendizado de máquina aplicado à classificação dos padrões de marcha do GCS-CP.
***Fonte**: Elaborada pelo autor.*

O espaço experimental foi estruturado em duas abordagens: *Flat* (multiclasse direta) e Hierárquica (cascata em duas etapas), contabilizando 432 modelos *Flat* e 720 Hierárquicos, totalizando 1.152 modelos treinados com validação cruzada estratificada de 5 dobras. A otimização bayesiana via Optuna (50 tentativas por modelo, com poda antecipada pelo HyperbandPruner) totalizou aproximadamente 57.600 treinamentos no espaço experimental principal, aos quais se somam 12.000 treinamentos da análise complementar de estabilidade estocástica para o alvo Balanço Hierárquico (Seção 3.9.4) e os treinamentos do retreinamento clínico orientado à sensibilidade (80 tentativas por configuração). A decomposição detalhada das combinações (*pipelines* × variantes × SMOTE × alvos × modelos) está apresentada na seção 4.1.

3.3.1 Pré-processamento dos Dados Cinemáticos

A inspeção sistemática das variáveis cinemáticas identificou 4.070 valores inadequados anatomicamente ou ausentes, distribuídos em 111 de 115 pacientes (96,5%). Esses valores foram classificados em três tipos: Tipo 1 – valores exatamente iguais a 0, 1 ou −1 (3.812 ocorrências; 93,7%), resultantes de falha de sensor ou preenchimento automático do software de processamento; Tipo 2 – valores fora dos limites anatômicos permissivos (246 ocorrências; 6,0%), concentrados predominantemente no tornozelo, articulação mais suscetível a fenômenos de *gimbal lock* (perda de um grau de liberdade rotacional quando dois eixos de rotação se alinham, gerando valores espúrios próximos a ±180°) e *sensor wrapping* (inversão abrupta do sinal angular quando o ângulo ultrapassa os limites de representação do sistema de captura); Tipo 3 – valores ausentes (NaN) nas variáveis espaço-temporais (12 ocorrências; 0,3%), restritos à variável *R Stride* de uma única participante por séries de captura insuficientes no lado direito.

Para o tratamento dos valores Tipo 2, foi aplicado um filtro anatômico permissivo, com limites definidos com margens de 20 a 40° acima dos valores de amplitude de movimento fisiologicamente alcançável documentados na literatura (51,52), acomodando manifestações severas de paralisia cerebral. Valores negativos de flexão do joelho, representando o fenômeno clínico do recurvato (hiperextensão do joelho), foram deliberadamente preservados, variando entre −1,1° e −27,4°, dentro do limite anatômico aceito de −60°, por constituírem padrão patológico biologicamente real que os modelos de classificação precisam identificar. O critério estatístico do intervalo interquartil (IQR 1,5×) foi

analisado e descartado por ser cego ao conhecimento clínico do domínio: aplicado a estes dados, classificaria como *outlier* 840 registros (32,8%), com impacto desproporcional sobre as classes mais graves, TE (58,1%), SA (44,8%) e TI (40,1%), que representam exatamente os padrões patológicos que os modelos devem aprender a identificar.

Os três tipos de valores foram tratados por imputação pontual pela mediana hierárquica em três níveis, aplicada separadamente para os lados Direito e Esquerdo. No primeiro nível, utilizou-se a mediana do próprio paciente no mesmo lado (≥ 5 observações válidas), cobrindo 72,0% das imputações angulares. No segundo nível, quando insuficiente, ampliou-se o conjunto para incluir ambos os lados do mesmo paciente (18,5% das imputações). No terceiro nível, recorreu-se à mediana da classe do alvo no mesmo lado, calculada sobre centenas de observações (9,5% das angulares e 100% das variáveis espaço-temporais). A mediana foi escolhida por ser robusta à presença de valores extremos biologicamente reais em dados de paralisia cerebral. A estratégia de imputação pontual, que substitui apenas o valor impossível, preservando todos os demais dados do registro, garantiu a integridade total da base: 100% dos 2.564 registros e 100% dos 115 pacientes foram preservados, sem exclusão de linhas em nenhuma etapa.

3.4 Divisão dos Dados e Prevenção de Vazamento

A separação entre treino e teste foi realizada por paciente, e não por ciclo de marcha ou por membro. Isso significa que todos os ciclos de um mesmo paciente (incluindo ambos os membros direito e esquerdo) foram alocados integralmente em um único conjunto, nunca sendo divididos entre treino e teste. Esta abordagem é essencial para evitar vazamento de dados, ou seja, se ciclos do mesmo paciente aparecessem nos dois conjuntos, o modelo poderia "memorizar" características biomecânicas individuais daquele paciente durante o treino e reconhecê-las no teste, produzindo estimativas de desempenho artificialmente otimistas. Verificou-se computacionalmente que zero pacientes aparecem em ambos os conjuntos. A Tabela 10 apresenta a divisão.

Tabela 10 – Divisão dos dados entre conjuntos de treino e teste

Conjunto	Registros		Pacientes	Proporção	Pacientes em ambos
Treino	1.804		86	70,4%	—
Teste	760		29	29,6%	—
Total	2.564		115	100%	—
Pacientes em ambos	—		**0**	0%	**✓ Confirmado**

Nota: A proporção refere-se aos registros. A amostra base compreende 2.564 registros (1.282 direito + 1.282 esquerdo), resultantes do empilhamento bilateral de 115 pacientes. O target Apoio conta com 2.549 registros válidos, em razão de 15 registros sem classificação disponível na variável-alvo; Balanço e Transverso utilizam os 2.564 registros completos. A célula destacada em verde confirma que nenhum paciente aparece simultaneamente em treino e teste, garantindo ausência total de vazamento de dados.
Fonte: *Elaborada pelo autor.*

Após o empilhamento bilateral, a base de dados continha 2.564 registros. No novo *pipeline*, os alvos tratados de forma independente, permitindo que cada modelo utilizasse o máximo de registros disponíveis para sua variável-alvo específica. Balanço e Transverso não apresentam valores ausentes, utilizando os 2.564 registros completos (1.804 treino / 760 teste). O Alvo Apoio conta com 2.549 registros válidos, em razão de 15 registros sem classificação disponível na variável-alvo, resultando em 1.795 registros no treino e 754 no teste. Em todos os casos, a divisão por paciente foi mantida e a ausência de sobreposição entre conjuntos foi verificada computacionalmente.

O processo de avaliação dos modelos seguiu um esquema de dois níveis. No primeiro nível, a validação cruzada estratificada de 5 dobras foi aplicada exclusivamente ao conjunto de treinamento, com o objetivo de estimar o desempenho de generalização de cada modelo e permitir a comparação entre algoritmos sem utilizar o conjunto de teste. Em cada dobra, 80% dos registros de treino foram utilizados para ajuste do modelo e 20% para validação interna, mantendo a proporção entre classes em cada partição (estratificação). A Acurácia Balanceada foi utilizada como métrica de referência para comparação entre modelos durante esta etapa. No segundo nível, o modelo final: treinado na totalidade do conjunto de treinamento, foi avaliado uma única vez no conjunto de teste fixo, que permaneceu completamente isolado durante todo o processo. Este esquema garante que os resultados reportados no conjunto de teste sejam imparciais e representativos do desempenho real esperado em novos pacientes.

3.5 Variáveis Alvo

O *pipeline* classifica, separadamente, três variáveis alvo baseadas no GCS-CP: Apoio Geral (Normal, Salto e Agachado), Balanço Geral (Normal e Rígido) e Transverso Geral (Normal, Transverso interno e Transverso Externo). A Tabela 11 apresenta a distribuição das classes, evidenciando o severo desbalanceamento.

Tabela 11 – Distribuição das classes por variável-alvo nos conjuntos de treino e teste

Alvo	Classe	Treino (n)	Treino (%)	Teste (n)	Teste (%)	Razão
Apoio	SS	1.105	61,6%	424	56,2%	**3,5:1**
	SN	375	20,9%	202	26,8%	
	SA	315	17,5%	128	17,0%	
Balanço	BN	1.596	88,5%	675	88,8%	**7,9:1**
	BR	208	11,5%	85	11,2%	
Transverso	TN	1.172	65,0%	509	67,0%	**8,9:1**
	TI	500	27,7%	194	25,5%	
	TE	132	7,3%	57	7,5%	

Nota: Células em amarelo indicam classes minoritárias com desbalanceamento severo. Razão = classe majoritária / classe minoritária. SS = Sagital Salto; SN = Sagital Normal; SA = Sagital Agachado; BN = Balanço Normal; BR = Balanço Rígido; TN = Transverso Normal; TI = Transverso Interno; TE = Transverso Externo.
Fonte: *Elaborada pelo autor.*

3.6 Seleção de Variáveis

O *pipeline* avaliou três variantes de seleção de atributos em paralelo, cada uma representando uma estratégia distinta de redução dimensional: (I) "Sem Consenso", com todos os atributos do *pipeline* sem filtro adicional, utilizada como linha de base; (II) "Com Consenso", com seleção por consenso entre quatro métodos independentes; e (III) PCA, com Análise de Componentes Principais retendo no mínimo 95% da variância acumulada.

A variante "Com Consenso" combina quatro métodos complementares: Limiar de Variância (*Variance Threshold*), Informação Mútua, *F-score* ANOVA e Importância por Floresta Aleatória, descritos em detalhe a seguir. Cada método supervisionado foi configurado para reter 60% das características candidatas ($K = \lceil 0{,}60 \times n \rceil$, em que n é o número

de características originais do *pipeline*). Um atributo é selecionado para o conjunto final quando aprovado por pelo menos o número mínimo de métodos definido pelo limiar de consenso, que varia conforme o número de classes do alvo: problemas com mais classes exigem seleção mais rigorosa para evitar atributos espúrios, conforme apresentado na Tabela 12. Esta estratégia, conhecida na literatura de aprendizado de máquina como seleção de variáveis por consenso (*ensemble feature selection)*, foi consolidada por Saeys et al. (25) como abordagem que combina múltiplos métodos de seleção para obter resultados mais robustos do que cada método isoladamente, mostrando-se particularmente vantajosa em cenários com tamanho amostral reduzido, característica relevante para os subconjuntos de pacientes com alteração confirmada utilizados na Etapa 2 da abordagem hierárquica (Seção 3.9.2).

Tabela 12 – Configuração da seleção de variáveis por consenso

Alvo	Limiar	Atributos Angular	Atributos Geral	Justificativa
Apoio	3/4 métodos (75%)	7	10	3 classes – consenso rigoroso para evitar *features* espúrias
Balanço	2/4 métodos (50%)	15	19	Problema binário – seleção mais permissiva para reter informação
Transverso	3/4 métodos (75%)	10	13	3 classes – mesmo critério do Apoio

Nota: Atributos Angular = número de atributos selecionados no *Pipeline* Cinemática Angular (18 atributos candidatos); Atributos Geral = número de atributos selecionados no *Pipeline* Cinemática Geral (25 atributos candidatos). O limiar diferenciado por alvo segue o princípio de que problemas com mais classes exigem seleção mais rigorosa.
Fonte: *Elaborada pelo autor.*

O primeiro método, Limiar de Variância (*Variance Threshold*), atua como filtro preliminar que descarta variáveis quase constantes. Se uma variável praticamente não muda entre os pacientes, tem pouca informação útil para distingui-los. É análogo a descartar, em uma anamnese, perguntas cuja resposta é praticamente a mesma para todos os pacientes, que não ajudam a diferenciar quadros clínicos. É computacionalmente leve e elimina redundâncias antes da aplicação de métodos mais sofisticados (53).

O segundo método, Seleção K-Melhores com Informação Mútua (*SelectKBest with Mutual Information*), mede o quanto cada variável reduz a incerteza sobre a classe a ser predita, ou seja, avalia se conhecer o valor de uma variável ajuda a adivinhar a qual

padrão de marcha o paciente pertence. É como, em uma consulta, perceber que saber se o paciente caminha em equino já reduz drasticamente as hipóteses diagnósticas plausíveis. Sua vantagem é capturar relações não lineares entre variáveis e classes, algo que métodos baseados em correlação linear deixam passar, característica especialmente relevante em dados biomecânicos, em que relações entre variáveis articulares e padrões da marcha raramente são lineares (53,54).

O terceiro método, Seleção K-Melhores com *F-score* ANOVA *(SelectKBest with ANOVA F-score*), parte de uma pergunta estatística clássica, análoga à que o clínico faz ao comparar grupos de pacientes: as médias dessa variável diferem significativamente entre as classes? O método atribui valor F mais alto às variáveis cujas médias se afastam claramente entre os grupos, identificando as de maior poder discriminativo. Enquanto a Informação Mútua captura padrões complexos, o ANOVA *F-score* se destaca em diferenças nítidas e diretas entre médias, os dois métodos se complementam (53).

O quarto método, Floresta Aleatória, adota estratégia diferente, análoga a uma junta médica: em vez de avaliar cada variável isoladamente, mede o quanto ela contribui para a precisão do modelo como um todo. O algoritmo treina centenas de árvores de decisão e embaralha aleatoriamente os valores de uma variável específica, observando o aumento da taxa de erro, quanto maior a piora, maior a importância daquela variável. Esse mecanismo de permutação *out-of-bag* é robusto a variáveis irrelevantes, captura interações entre variáveis (algo que os três métodos anteriores não fazem) e fornece estimativa já validada internamente, sem necessidade de conjunto de teste separado (55).

Uma variável foi selecionada apenas se identificada como relevante por um número mínimo de métodos, o limiar de consenso, conforme Tabela 12. Esta estratégia reduz o risco de seleção de variáveis espúrias que apenas um método isolado identificaria como relevantes, aumentando a robustez e a generalização do conjunto final de características.

Como abordagem comparativa de redução de dimensionalidade, foi também implementada a Análise de Componentes Principais (PCA), o método de extração de características mais utilizado em investigações de biomecânica da marcha (26). A ideia central do PCA é simples: em vez de escolher quais variáveis manter, ele cria variáveis novas, chamadas de componentes principais, formadas a partir de combinações lineares das variáveis originais. O resultado é um novo sistema de coordenadas em que cada eixo é ordenado pela quantidade de variância que explica nos dados (56). Matematicamente, esses componentes são obtidos pela decomposição em valores singulares (SVD) da matriz de dados, e têm uma propriedade importante: são ortogonais entre si, ou seja, capturam dimensões de variação

independentes umas das outras. O primeiro componente concentra a maior parcela possível da variância total; o segundo, a maior parcela da variância restante; e assim sucessivamente (56). As implicações dessa escolha metodológica para dados biomecânicos são discutidas na Seção 5.3.

O número de componentes retidos foi definido pelo critério de variância explicada acumulada ≥ 95%, ou seja, manteve-se o menor conjunto de componentes capaz de preservar ao menos 95% da informação original dos dados. Este critério resultou em 12 componentes para o *Pipeline* Cinemática Angular e 16 para o *Pipeline* Cinemática Geral, e é amplamente adotado em estudos de aprendizado de máquina aplicados a dados biomecânicos.

Para cada alvo, o PCA foi ajustado exclusivamente sobre o conjunto de treino e, em seguida, aplicado ao conjunto de teste com os parâmetros já fixados. Esse procedimento impede que informação do teste influencie a construção dos componentes, evitando vazamento de dados (*data leakage).*

3.7 Modelos de Aprendizado de Máquina

Os modelos de aprendizado de máquina foram implementados utilizando a biblioteca scikit-learn (53). Para os modelos de *gradient boosting*, foram utilizadas as bibliotecas LightGBM (21) e XGBoost (22). Todos os hiperparâmetros foram otimizados individualmente por meio da biblioteca Optuna (57), utilizando o algoritmo TPE (*Tree-structured Parzen Estimator*) multivariado com poda antecipada de tentativas pelo HyperbandPruner, conforme descrito na Seção 3.8.

Foram treinados nove algoritmos individuais de aprendizado de máquina supervisionado, cujos fundamentos conceituais foram apresentados na Seção 2.9 e cujos espaços de busca de hiperparâmetros estão detalhados na Tabela 13. Algumas configurações de implementação merecem destaque. A Regressão Logística (58) utilizou o *solver lbfgs* com balanceamento de classes. O SVM (59) empregou *kernel* RBF (*Radial Basis Function*), permitindo ao Optuna otimizar simultaneamente os hiperparâmetros C e gamma, o que cria uma superfície de busca mais rica do que o *kernel* linear. As Árvores de Decisão (60) e o K-NN (61), não cobertos individualmente na Seção 2.9, particionam recursivamente o espaço de atributos em regras hierárquicas e classificam pela classe predominante entre os K vizinhos mais próximos, respectivamente. O *Naive Bayes* Gaussiano (62) é um classificador probabilístico baseado no Teorema de Bayes que assume distribuição normal dos atributos por

classe. O MLP (63) foi implementado com o MLPClassifier do scikit-learn, função de ativação ReLU ou tanh, regularização L2 via hiperparâmetro alpha e *early stopping* (n_iter_no_change=10, validation_fraction=0,1).

Como estratégia complementar ao SMOTE, todos os modelos que permitem esse ajuste foram configurados com o parâmetro *class_weight='balanced'*. Na prática, essa configuração faz com que o algoritmo atribua um peso maior às classes menos frequentes durante o treinamento: errar uma amostra de uma classe rara passa a "custar" mais do que errar uma amostra de uma classe comum. Dessa forma, o modelo é incentivado a prestar mais atenção aos padrões minoritários, em vez de simplesmente favorecer as classes mais numerosas.

A ponderação de classes pertence a uma família de técnicas conhecida como aprendizado sensível ao custo (*cost-sensitive learning*), que, segundo a revisão de Araf, Idri e Chairi (32), pode ser implementada em três momentos do *pipeline*: antes do treinamento (pré-processamento), via métodos como *o MetaCost*; durante o treinamento, incorporando os custos diretamente na função objetivo do algoritmo, caso do *class_weight='balanced'* adotado neste estudo; e após o treinamento (pós-processamento), ajustando o limiar de decisão para corrigir o viés em favor da classe majoritária. A mesma revisão, ao analisar 173 estudos biomédicos, reportou melhoria de desempenho em 97,7% dos casos em que técnicas sensíveis ao custo foram aplicadas a dados desbalanceados (32).

Além das técnicas individuais descritas, foram avaliadas estratégias de combinação de modelos. Quatro opções de implementação específicas deste estudo merecem destaque: (I) no *Voting Hard*, empates retornam erro (−1) em vez da seleção por classe de menor índice (padrão da literatura), evitando inflar artificialmente o desempenho da classe Normal; (II) o *Voting Soft* foi implementado em sua configuração padrão, com a classe final determinada pela média não ponderada das probabilidades preditas pelos três classificadores-base; (III) o meta-modelo do *Stacking* foi a Regressão Logística regularizada com validação cruzada de 3 dobras; (IV) o MLP foi incluído nas fusões mesmo quando não figurava entre os três melhores, garantindo diversidade de hipóteses. A Tabela 13 sintetiza, por modelo, o espaço de busca dos hiperparâmetros otimizados pelo Optuna e a configuração de *class_weight* adotada.

Tabela 13 – Modelos e espaço de busca de hiperparâmetros

Modelo	Hiperparâmetros otimizados pelo Optuna	*class_weight*
LogReg	*C (log), solver (lbfgs/saga), max_iter=2000*	*balanced*
SVM	*C (log), gamma (log), kernel (rbf/poly), max_iter=5000*	*balanced*
DTree	*max_depth (3-15), min_samples_split (2-20), min_samples_leaf (1-10), criterion (gini/entropy)*	*balanced*
K-NN	*n_neighbors (3-15), weights (uniform/distance), metric (euclidean/manhattan) '*	N/A
Naive Bayes	*var_smoothing (log, 1e-12 a 1e-5)*	N/A
RandomForest	*n_estimators (100-500, step=50), max_depth (5-20), min_samples_leaf (1-8), max_features (sqrt/log2/0.5)*	*balanced*
MLP	*n_layers (1-4), n_units por camada [32, 64, 128, 256], activation (relu), learning_rate (log, 1×10^{-4} a 1×10^{-2}), alpha L2 (log, 1×10^{-5} a 1×10^{-2}), max_iter=500, early_stopping=True (validation_fraction=0.1, n_iter_no_change=10)*	N/A
LightGBM	*n_estimators (100-600, step=50), max_depth (3-10), num_leaves (15-127), learning_rate (log), subsample (0.5-1.0), colsample_bytree (0.5-1.0)*	*balanced*
XGBoost	*n_estimators (100-600, step=50), max_depth (3-10), learning_rate (log), subsample (0.5-1.0), colsample_bytree (0.5-1.0), gamma (0-5)*	N/A (*scale_pos_weight*)
Voting Hard	3 melhores modelos por CV – empate retorna -1 (erro)	N/A
Voting Soft	3 melhores modelos por CV – média das probabilidades	N/A
Stacking	Meta-modelo: *LogReg* (*balanced*), cv=3, 3 melhores modelos por CV + MLP forçado	N/A

Nota: (log) indica que o hiperparâmetro é amostrado em escala logarítmica pelo Optuna. Modelos em verde são fusões. *class_weight='balanced'* ajusta pesos inversamente proporcionais à frequência das classes. N/A = não aplicável. O XGBoost utiliza *scale_pos_weight* calculado automaticamente por classe. O MLP é forçado nas fusões mesmo quando não está no 3 melhores naturais, garantindo diversidade de hipóteses. Empate no *Voting Hard* retorna -1 (erro), tratamento metodologicamente mais honesto que seleção arbitrária.
Fonte: *Elaborada pelo autor.*

Os algoritmos foram aplicados conforme a natureza de cada alvo: na modalidade multiclasse para Apoio (SN/SS/SA) e Transverso (TN/TI/TE), que possuem três classes cada, e na modalidade binária para Balanço (BN/BR), que possui apenas duas. Como alguns dos algoritmos utilizados, notadamente a Regressão Logística e o SVM, foram originalmente formulados para problemas binários, sua aplicação a alvos de três classes exige uma estratégia de extensão. A literatura de aprendizado de máquina oferece três alternativas clássicas.

A primeira é a estratégia *One-vs-Rest* (OvR), em que se treinam tantos classificadores binários quantas forem as classes do problema. Para o alvo Apoio, por exemplo, são treinados três classificadores independentes: Normal vs. não-Normal, Salto vs. não-Salto e Agachado vs. não-Agachado. Cada um produz uma probabilidade de pertencimento à sua respectiva classe, e o padrão final é atribuído à classe com maior probabilidade.

A segunda é a softmax multinomial, em que um único modelo é generalizado para produzir simultaneamente as probabilidades de todas as classes, somando exatamente 1 entre elas. Em vez de decompor o problema em classificadores binários, o algoritmo aprende diretamente a fronteira entre as três classes em um único processo de treinamento.

A terceira é a *One-vs-One* (OvO), em que classificadores binários são treinados para cada par de classes possível, totalizando $n(n-1)/2$ classificadores para n classes. A diferença em relação ao OvR é fundamental: enquanto cada classificador OvR é treinado com todas as amostras (com a classe-alvo rotulada como "positivo" e todas as demais agrupadas como "negativo"), cada classificador OvO é treinado apenas com as amostras das duas classes em disputa, ignorando completamente as demais. Para o alvo Apoio, isso resulta em três classificadores: Normal vs. Salto (treinado apenas com amostras Normal e Salto, ignorando Agachado), Normal vs. Agachado (apenas Normal e Agachado, ignorando Salto) e Salto vs. Agachado (apenas Salto e Agachado, ignorando Normal). A predição final é determinada por votação majoritária entre os três classificadores, com a classe mais frequentemente apontada sendo atribuída à amostra.

No presente estudo, foram adotadas as configurações padrão das implementações nativas do scikit-learn (53), que diferem entre os dois algoritmos: a Regressão Logística (com *solver lbfgs*) utilizou softmax multinomial, e o SVM (SVC) utilizou *One-vs-One* internamente, com as saídas reorganizadas em formato compatível com OvR para fins de predição. As três estratégias são amplamente utilizadas na literatura de aprendizado de máquina e cumprem a mesma função: permitir que classificadores formulados para problemas binários operem em alvos multiclasse.

3.8 Tratamento de Desbalanceamento e Estratégias de Regularização

Modelos de aprendizado de máquina aplicados a dados desbalanceados estão sujeitos a duas dificuldades centrais: o domínio das classes majoritárias durante o aprendizado e o *overfitting*, particularmente em arquiteturas de alta capacidade de representação. Esta seção apresenta três frentes complementares para enfrentar essas dificuldades: balanceamento sintético da classe minoritária via SMOTE (Seção 3.8.1), regularização específica das redes neurais (Seção 3.8.2) e discussão e monitoramento do *overfitting* (Seção 3.8.3).

3.8.1 SMOTE: Balanceamento da Classe Minoritária e Prevenção de Vazamento de Dados

O SMOTE (*Synthetic Minority Over-sampling Technique*) gera amostras sintéticas por interpolação entre amostras existentes da classe minoritária e seus k vizinhos mais próximos, utilizando o valor padrão de k = 5 recomendado pelos autores originais (64). Diferentemente da simples replicação de amostras, que resulta em regiões de decisão cada vez mais específicas e propensas ao *overfitting*, o SMOTE força os classificadores a construírem regiões de decisão mais amplas e generalizáveis para a classe minoritária (64). Se aplicado antes da divisão treino/teste, amostras sintéticas poderiam conter "fragmentos" de informação do teste, constituindo vazamento de dados. Por esta razão, o SMOTE foi aplicado estritamente após a divisão dos dados, operando exclusivamente sobre o conjunto de treino. O conjunto de teste foi mantido com sua distribuição original (desbalanceada), refletindo a prevalência real das condições clínicas.

3.8.2 Regularização nas Redes Neurais

O MLP, implementado com o *MLPClassifier* da biblioteca *scikit-learn*, é particularmente suscetível ao *overfitting* devido à sua alta capacidade de representação, motivo pelo qual foram adotadas três estratégias complementares de regularização. A primeira é a penalidade L2 sobre os pesos da rede, controlada pelo hiperparâmetro alpha, otimizado pelo Optuna em escala logarítmica no intervalo de 1×10^{-5} a 5×10^{-3}. Esta penalização adiciona à função de custo um termo proporcional ao quadrado da norma dos pesos, desencorajando soluções em que conexões individuais assumem valores extremos e favorecendo representações

distribuídas mais robustas, princípio analisado formalmente por Krogh e Hertz (65) e integrado por padrão à implementação do MLPClassifier. A segunda estratégia é o *early stopping* ativado via *early_stopping=True*, no qual 10% do conjunto de treinamento é reservado como partição interna de validação e o treinamento é interrompido automaticamente quando a pontuação nesta partição deixa de melhorar por 10 épocas consecutivas (*valor default do n_iter_no_change do scikit-learn*). A terceira estratégia é a restrição da capacidade arquitetural, com o espaço de busca do Optuna limitado a no máximo três camadas ocultas e 256 unidades por camada, calibrado de acordo com o tamanho moderado do conjunto de treinamento ($n \approx 1.800$ registros de 86 pacientes), pois arquiteturas mais profundas ou mais largas tenderiam a memorizar padrões específicos do treino em vez de generalizar, conforme observado por Slijepcevic et al. (15) em população semelhante. A combinação dessas três estratégias substitui a necessidade de *dropout* (66), técnica não utilizada na implementação do MLPClassifier.

3.8.3 *Overfitting* e seu Monitoramento

O *overfitting* ocorre quando um modelo aprende não apenas os padrões genuínos dos dados de treino, mas também o ruído e particularidades específicas daquele conjunto, comprometendo a capacidade de generalização para novos dados. A Tabela 14 sintetiza as estratégias de prevenção e monitoramento adotadas neste estudo, integrando as técnicas detalhadas nas Seções 3.8.1 e 3.8.2 com as práticas de proteção contra vazamento de dados e a métrica de monitoramento descrita a seguir.

Tabela 14 – Estratégias de regularização e prevenção de *overfitting*

Estratégia	Aplicação	Justificativa
SMOTE exclusivo no treino	Após divisão treino/teste	Amostras sintéticas de teste 'vazariam' informação, inflando métricas artificialmente
Divisão por paciente	Nenhum paciente em ambos os conjuntos	Evita memorização de características biomecânicas individuais
***StandardScaler* isolado**	*fit*() treino, *transform*() teste	Parâmetros de normalização calculados apenas no treino
Validação cruzada 5 dobras	Otimização interna no treino	Avalia generalização antes de tocar o conjunto de teste
***HyperbandPruner* (Optuna)**	Otimização de hiperparâmetros	Interrompe tentativas ruins após a 2° dobra de validação cruzada, economizando 35-45% do tempo de processamento
Penalidade L2 (alpha)	Redes Neurais (MLP)	Regulariza os pesos da rede via termo quadrático na função de custo, otimizado pelo Optuna no intervalo 1×10^{-5} a 5×10^{-3}
early Stopping	Redes Neurais (MLP)	Interrompe o treino quando a pontuação na partição interna de validação (*validation_fraction*=0,1) não melhora por 10 iterações consecutivas (n_iter_no_change=10, *default* do scikit-learn)
***max_depth* otimizado pelo Optuna**	RF, LightGBM, XGBoost, Árvores	Restringe complexidade dentro de espaço calibrado, evitando divisões excessivamente específicas
class_weight=balanced	*LogReg*, SVM, RF, LightGBM, Árvores	Penaliza erros em classes minoritárias, complementando o SMOTE
Métrica de *Overfitting*	Bal.Acc_Treino - Bal.Acc_Teste	Monitora capacidade de generalização; diferenças acentuadas são investigadas

Nota: As estratégias SMOTE exclusivo no treino, Divisão por paciente e *StandardScaler* isolado são críticas para prevenção de vazamento de dados (*data leakage*).
Fonte: Elaborada pelo autor.

A capacidade de generalização foi monitorada pela métrica de *Overfitting* (Acurácia Balanceada Treino − Acurácia Balanceada Teste), incluída em todas as tabelas de ranqueamento. Valores próximos a zero indicam que o modelo aprende padrões genuínos dos dados; diferenças mais acentuadas sugerem que o modelo passou a capturar também o ruído e particularidades específicas do conjunto de treino, comportamento característico de *overfitting*. A literatura não estabelece limiares universais para essa interpretação, sendo dependente do contexto e da complexidade do problema.

Para os modelos treinados com SMOTE, parte da diferença entre treino e teste, frequentemente denominada lacuna de generalização (*generalization gap*), não constitui

overfitting problemático no sentido tradicional, mas sim o comportamento esperado de um protocolo metodológico correto. A aplicação do SMOTE exclusivamente no conjunto de treino introduz uma mudança intencional de distribuição (*dataset shift*) entre as etapas de treinamento e teste. Conforme formalizado por Moreno-Torres et al. (67), o *dataset shift* ocorre quando as distribuições conjuntas dos dados de treino e teste são diferentes. No presente estudo, o conjunto de treino com SMOTE opera em cenário artificialmente balanceado, enquanto o conjunto de teste preserva a distribuição original desbalanceada, refletindo a prevalência real dos padrões da marcha na população clínica.

Esta diferença de distribuição é parte integrante do protocolo metodológico correto para tratamento de dados desbalanceados. Em sua revisão sobre os 15 anos do SMOTE, Fernández et al. (68) ressaltam a importância de preservar a integridade distribucional na etapa de validação, argumento que se estende naturalmente ao princípio de manter o conjunto de teste livre de amostras sintéticas. Dessa forma, o esforço do pesquisador se concentra em modelar a realidade desbalanceada dos dados, e não em compensar artificialmente lacunas de generalização geradas por protocolos incorretos.

Para os modelos treinados sem SMOTE, esta justificativa não se aplica, e a interpretação do *overfitting* observado deve considerar outros fatores, como a complexidade arquitetural do modelo, o tamanho amostral moderado e a dificuldade de separabilidade entre classes patológicas com poucas amostras. Esses fatores são analisados em detalhe na Seção 5.7.

3.9 Análise Complementar: uma abordagem hierárquica

Além da abordagem multiclasse direta, foi avaliada uma estratégia alternativa baseada em classificação hierárquica (cascata). Esta análise complementar investigou se a decomposição do problema em duas etapas sequenciais poderia melhorar o desempenho na detecção de subtipos de alteração da marcha. Esta abordagem é conceitualmente alinhada ao raciocínio clínico, no qual o profissional primeiro identifica a presença de alteração (triagem) para então caracterizar o tipo específico de desvio (subtipagem), conforme o sistema de classificação hierárquica descrito por Davids e Bagley (9) e posteriormente validado por Melanda et al. (14).

3.9.1 Estrutura da Classificação Hierárquica

A classificação hierárquica foi estruturada em duas etapas sequenciais:
Etapa 1 (Triagem): Para os alvos Apoio Geral e Transverso Geral, as classes originais foram reagrupadas em duas categorias binárias. No alvo Apoio, as classes SN (Normal) permaneceram como "Normal", enquanto SA (Agachado) e SS (Salto) foram agrupadas como "Alterado". No alvo Transverso, TN (Transverso Normal) permaneceu como "Normal", enquanto TI (Transverso Interno) e TE (Transverso Externo) foram agrupadas como "Alterado". O alvo Balanço Geral, por já ser originalmente binário (BN vs BR), não necessitou de reagrupamento.

Etapa 2 (Subtipagem): Apenas os casos classificados como "Alterado" na Etapa 1 prosseguiram para a segunda etapa, onde foram discriminados entre os subtipos específicos: SA versus SS para o alvo Apoio, e TI versus TE para o alvo Transverso.

3.9.2 Seleção de Características por Etapa

A seleção de características foi realizada separadamente para cada etapa, aplicando-se em ambas o método "Com Consenso" descrito na Seção 3.6. Na Etapa 1, utilizou-se todo o conjunto de treino disponível. Na Etapa 2, a seleção foi conduzida apenas sobre o subconjunto de pacientes com alteração confirmada (classificados como "Alterado" pela Etapa 1), cenário em que a estratégia de seleção de variáveis por combinação de múltiplos métodos da Seção 3.6 foi adotada em razão do tamanho amostral reduzido.

3.9.3 Cálculo da Sensibilidade Efetiva (Propagação de Erros)

Um aspecto crítico da abordagem hierárquica é a propagação de erros entre as etapas. Um paciente só pode ser corretamente classificado no subtipo final se for primeiro corretamente identificado como "Alterado" na Etapa 1. Assim, a sensibilidade efetiva da cascata para cada subtipo de alteração foi calculada conforme a Equação (1):

$$Sens\_Efetiva(Subtipo) = Sens\_E1(Alterado) \times Sens\_E2(Subtipo) \quad (1)$$

Onde Sens_E1(Alterado) representa a sensibilidade da Etapa 1 para detectar casos alterados, e Sens_E2(Subtipo) representa a sensibilidade da Etapa 2 para classificar corretamente o subtipo específico (SA, SS, TI ou TE). Para as classes normais (SN e TN), a

sensibilidade efetiva corresponde diretamente à sensibilidade da Etapa 1 para a classe "Normal".

A Equação (1) calcula a sensibilidade efetiva como o produto da sensibilidade da Etapa 1 (identificar o caso como alterado) pela sensibilidade da Etapa 2 (atribuir o subtipo correto), refletindo que ambas as etapas precisam acertar para que a classificação final esteja correta. Essa multiplicação pressupõe independência entre as etapas: como a sensibilidade da Etapa 2 foi estimada sobre todos os casos verdadeiramente alterados, assume-se que o acerto da Etapa 2 ocorre na mesma proporção independentemente de a Etapa 1 ter acertado ou errado aquele caso. Caso os erros das duas etapas estejam correlacionados, isto é, casos difíceis para a Etapa 1 também sejam difíceis para a Etapa 2, o produto tende a subestimar o desempenho real da cascata, uma vez que os casos que passam pela Etapa 1 são, em média, os mais fáceis e tendem a ser também os mais bem classificados pela Etapa 2; os valores reportados constituem, assim, limites inferiores conservadores. Esse pressuposto é avaliado à luz dos resultados na Seção 5.2.

3.9.4 Análise de Estabilidade Estocástica

A otimização bayesiana de hiperparâmetros via *Optuna* apresenta componente estocástico inerente ao algoritmo TPE (*Tree-structured Parzen Estimator),* que pode levar a configurações finais distintas em execuções independentes mesmo sobre os mesmos dados. Para investigar empiricamente em que medida o desempenho dos modelos é sensível a essa variabilidade, foi conduzida análise complementar de estabilidade estocástica restrita ao alvo Balanço, em que a abordagem Hierárquica e a abordagem *Flat* apresentaram diferença de 2,8 pontos percentuais em Acurácia Balanceada na varredura preliminar. O experimento consistiu na reexecução completa do *pipeline* Hierárquico para esse alvo com cinco *seeds* aleatórias distintas (0, 1, 7, 42 e 123), preservando todas as demais condições (divisão treino/teste, pré-processamento, espaços de busca de hiperparâmetros e número de tentativas do Optuna). A *seed* 42, valor convencional na comunidade de aprendizado de máquina utilizado para garantir reprodutibilidade do estudo principal, foi incluída para verificar se o resultado original era replicável. Foram avaliadas as quatro combinações *Pipeline* × SMOTE (Cinemática Angular × "Com SMOTE"; Angular × "Sem SMOTE"; Geral × "Com SMOTE"; Geral × "Sem SMOTE") para os 12 algoritmos do estudo, totalizando 240 execuções. Para cada combinação, reportou-se a média, o desvio-padrão e a faixa de variação da Acurácia

Balanceada no conjunto de teste entre as cinco *seeds*, permitindo distinguir desempenho sistematicamente superior de variabilidade amostral. Os resultados desta análise são reportados na Seção 4.8.

3.10 Métricas de Avaliação

A avaliação de modelos de classificação em dados desbalanceados requer métricas que reflitam adequadamente a capacidade do modelo em detectar todas as classes, não apenas as majoritárias. Neste estudo foram utilizadas 13 métricas de avaliação, apresentadas na Tabela 15.

Tabela 15 – Métricas de avaliação utilizadas no estudo

Métrica	Descrição / Fórmula	Interpretação	Ideal
Bal.Acc	Média aritmética dos recalls de cada classe	MÉTRICA PRINCIPAL; corrige desbalanceamento	1,0
G-Mean	Média geométrica dos recalls: $(R_1 \times R_2 \times ... \times R_n)^{(1/n)}$	Zero se qualquer classe tiver recall=0	1,0
MCC	Matthews Correlation Coefficient	Melhor métrica única para dados desbalanceados; varia de -1 a +1; 0 = aleatório	1,0
Kappa	Cohen's Kappa: concordância corrigida pelo acaso	<0,00=fraca; 0,00-0,20=leve; 0,21-0,40=razoável; 0,41-0,60=moderado; 0,61-0,80=substancial; 0,81-1,00=quase perfeito (45) | 1,0	1,0
Acurácia	(VP+VN) / Total	Enganosa em dados desbalanceados; reportada para comparação com literatura	1,0
Precision	VP / (VP+FP) — *macro*	Dos preditos positivos, quantos são corretos	1,0
Recall	VP / (VP+FN) — *macro*	Dos reais positivos, quantos foram detectados	1,0
F1-Score	*2×(Precision×Recall) / (Precision+Recall) — macro*	Média harmônica entre *Precision* e *Recall*	1,0
Sens*	Sensibilidade média (*macro*-averaged): média dos recalls por classe	Equivalente *à Bal.Acc*; reportada por classe individualmente nas matrizes de confusão	1,0
Spec*	Especificidade média (*macro*-averaged): VN/(VN+FP) por classe	Média das especificidades por classe (*one-vs-rest*)	1,0
AUC	Área sob a curva ROC (*OvR macro*)	0,5=sem discriminação; 0,7-0,8=aceitável; 0,8-0,9=excelente; >0,9=excepcional (58). Independe do ponto de corte de decisão	1,0
Log_Loss	$-\Sigma[y \times \log(p)]$	Qualidade das probabilidades; menor = melhor calibração	0,0
ECE	$\Sigma m\ (\|Bm\|/n) \times \|acc(Bm) - conf(Bm)\|$	Erro de Calibração Esperado; três faixas operacionais adotadas neste estudo: < 0,05 = bem calibrado; 0,05–0,10 = razoável; > 0,10 = mal calibrado	0,0
Overfitting	Bal.Acc_Treino - Bal.Acc_Teste	Valores próximos a zero indicam boa generalização; diferença esperada quando SMOTE aplicado só no treino (*dataset shift* intencional)	0,0

Nota: O asterisco (*) indica métricas calculadas como média aritmética dos valores por classe (*macro*-averaged), conforme recomendado para problemas multiclasse com dados desbalanceados. VP = Verdadeiros Positivos; VN = Verdadeiros Negativos; FP = Falsos Positivos; FN = Falsos Negativos; OvR = *One-vs-rest*. Bal.Acc = Acurácia Balanceada; G-Mean = Média Geométrica dos recalls; MCC = *Matthews Correlation Coefficient*; AUC = Área sob a curva ROC; ECE = Erro de Calibração Esperado. Para o ECE, B_m representa a m-ésima caixa de probabilidades; $acc(B_m)$ representa a frequência observada de positivos na caixa; $conf(B_m)$ representa a probabilidade média predita na caixa.
Fonte: *Elaborada pelo autor.*

Acurácia Balanceada (*Balanced Accuracy*): Em problemas de classificação com dados desbalanceados, a acurácia tradicional pode ser enganosa, pois um modelo que simplesmente prediz sempre a classe majoritária pode alcançar valores elevados sem detectar as classes minoritárias (69). A Acurácia Balanceada resolve este problema calculando a média aritmética das sensibilidades (*recalls)* de cada classe, atribuindo peso igual a todas as classes independentemente de sua frequência no conjunto de dados (29,69), justificando sua adoção como métrica principal neste estudo.

Para avaliar a estabilidade da estimativa de Acurácia Balanceada, foi construído um intervalo de confiança de 95% pelo método *bootstrap* não paramétrico por percentil, técnica clássica descrita por Efron e Tibshirani (70) e aplicada em contexto biomédico por Chong e Choo (71). Em cada uma das 1.000 replicações, os pares (predição, rótulo) do conjunto de teste foram reamostrados com reposição, e a Acurácia Balanceada foi recalculada para cada amostra simulada. Os limites inferior e superior do intervalo foram obtidos pelos quantis 2,5% e 97,5% da distribuição empírica resultante.

Para garantir a reprodutibilidade, o gerador de números pseudoaleatórios foi inicializado com *seed* fixa (*random_state*=42). A largura do intervalo foi adotada como medida sintética da precisão de cada estimativa pontual: quanto menor a largura, maior a precisão. Os valores absolutos das larguras são reportados nos resultados, permitindo a comparação direta entre modelos.

Para interpretação dos valores de Kappa, adotou-se a escala proposta por Landis e Koch (45), apresentada na Tabela 15. Já para os valores de AUC, adotou-se a escala proposta por Hosmer e Lemeshowowow (58), igualmente apresentada na Tabela 15.

Sensibilidade representa a capacidade do modelo em identificar corretamente os casos positivos, enquanto especificidade representa a capacidade de identificar corretamente os casos negativos. Contudo, em problemas multiclasse, como os alvos Transverso e Apoio deste estudo, cada um com três classes, surge a questão de como reportar estas métricas de forma consolidada, uma vez que cada classe possui sua própria sensibilidade e especificidade (30).

Adotou-se a abordagem de média macro (*macro-averaged*), recomendada por Sokolova e Lapalme (31) e Grandini, Bagli e Visani (30) para classificadores em dados desbalanceados: a sensibilidade e a especificidade são calculadas individualmente por classe (*one-vs-rest*) e depois agregadas por média aritmética simples, atribuindo peso igual a todas as classes. A definição operacional consta da Tabela 15.

Esta escolha justifica-se pelo contexto clínico. Conforme observam Sokolova e Lapalme (31), a média macro atribui peso igual a todas as classes, enquanto média ponderada e *micro* favoreceriam as classes majoritárias. Em análise da marcha, é igualmente importante detectar TN, TI e TE independentemente de sua prevalência na amostra.

A implementação computacional utilizou a função *balanced_accuracy_score* da biblioteca *scikit-learn* (53). Para a especificidade média, cálculo análogo foi implementado por classe *(one-vs-rest)* e depois agregado por média aritmética.

Nas tabelas de ranqueamento (Tabelas 18, 20 e 22), as colunas Sens* e Spec* representam as médias macro das sensibilidades e especificidades calculadas para cada classe. O asterisco (*) foi adotado para distinguir estas métricas agregadas dos valores individuais de cada classe, que podem ser derivados diretamente das matrizes de confusão (Figuras 21, 24, 27 e 28) e são analisados em detalhe na Seção 5.4.

A matriz de confusão complementa as métricas agregadas ao detalhar, classe a classe, a distribuição das predições do modelo em relação aos rótulos verdadeiros. Cada linha representa uma classe verdadeira e cada coluna, uma classe predita, sendo os valores na diagonal correspondentes aos acertos e os valores fora da diagonal aos erros de classificação entre pares específicos de classes. Este formato permite identificar quais classes são sistematicamente confundidas entre si, informação particularmente relevante em contextos clínicos com classes patológicas heterogêneas. As matrizes de confusão dos modelos vencedores para cada alvo (Apoio, Balanço e Transverso) são apresentadas nas Figuras 21, 24, 27 e 28 da Seção 4.

Erro de Calibração Esperado (ECE): mede se as probabilidades preditas pelo modelo correspondem às frequências reais das classes (72). Para cada classe, as predições do conjunto de teste são agrupadas em M faixas (*bins*) no intervalo [0, 1]; em cada faixa, comparam-se a probabilidade média predita pelo modelo (confiança) com a proporção real de casos positivos (acurácia local). O ECE é a média ponderada das diferenças absolutas entre essas duas quantidades. Valores próximos de zero indicam que as probabilidades preditas refletem fielmente as chances reais. Para fins operacionais nesta dissertação, adotaram-se três faixas: valores abaixo de 0,05 indicam boa calibração; valores entre 0,05 e 0,10 indicam calibração razoável; valores acima de 0,10 indicam má calibração probabilística, situação que compromete o uso direto das probabilidades em decisões clínicas por limiar e em estratégias de combinação por médias. O número de faixas foi definido de forma adaptativa entre 2 e 8, conforme o número de valores únicos de probabilidade predita em cada classe, com divisão uniforme via função *calibration_curve* da biblioteca *scikit-learn* (53).

3.11 Estratégias de Otimização Clínica

Para minimizar falsos negativos nas cinco classes patológicas do estudo (SS, SA, BR, TI e TE), foram avaliadas três estratégias complementares de otimização clínica. As três estratégias correspondem às três categorias reconhecidas na taxonomia de aprendizado sensível ao custo consolidada por Araf, Idri e Chairi (32): seleção do modelo por métrica orientada ao custo clínico (E1), ajuste de limiar de decisão por pós-processamento (E2) e incorporação direta dos custos durante o retreinamento (E3).

A Estratégia E1 (Seleção por sensibilidade mínima) substitui o critério de seleção do modelo vencedor: em vez da Acurácia Balanceada, seleciona-se o *pipeline* que maximiza a sensibilidade da classe patológica com menor desempenho, mantendo o ponto de corte padrão de 0,50. Esta estratégia não requer nenhuma modificação no modelo, apenas uma mudança de critério de seleção.

A Estratégia E2 (Ajuste de ponto de corte) mantém o mesmo modelo base, mas reduz o limiar de decisão abaixo de 0,50, tornando o modelo mais sensível à classe patológica. O objetivo é atingir sensibilidade $\geq 90\%$ para a classe alvo, com o efeito sobre a especificidade reportado nos resultados. Esta abordagem alinha-se com a literatura recente, que indica o limiar padrão de 0,50 como subótimo em contextos médicos desbalanceados e o limiar ótimo como tendendo a acompanhar a prevalência da classe patológica (73).

A Estratégia E3 (Retreinamento clínico) reotimiza o modelo a partir do zero com o Optuna, redefinindo o objetivo da otimização para maximizar a sensibilidade mínima entre as classes patológicas em substituição à Acurácia Balanceada. No presente estudo, esta estratégia foi aplicada exclusivamente ao algoritmo Floresta Aleatória, com o objetivo de demonstrar a viabilidade técnica da estratégia e seu impacto sobre o desempenho clínico, sem pretensão de comparação sistemática entre classificadores. O modelo resultante prioriza a sensibilidade das classes patológicas como critério de otimização.

3.12 Uso de Inteligência Artificial Generativa

O desenvolvimento deste estudo contou com o auxílio de modelos de inteligência artificial generativa como ferramentas de suporte metodológico e operacional. O modelo de linguagem Claude Opus 4.7 (Anthropic, 2026) foi utilizado para: (I) idealização e

estruturação da arquitetura do *pipeline* de aprendizado de máquina; (II) discussão e refinamento de estratégias metodológicas, incluindo abordagens de validação, tratamento de desbalanceamento e seleção de métricas; (III) geração, revisão e depuração de código Python para implementação dos modelos e análises; (IV) auxílio na redação, revisão e aprimoramento do texto científico; e (V) elaboração e formatação de tabelas e figuras. Adicionalmente, o modelo Gemini Pro (Google DeepMind, 2025) foi utilizado como ferramenta auxiliar na edição e no tratamento visual de figuras.

Todas as decisões metodológicas, interpretações dos resultados, análises clínicas e conclusões foram elaboradas pelos autores. Os resultados gerados pelos assistentes de inteligência artificial foram sistematicamente revisados, validados e, quando necessário, corrigidos pelo autor principal sob supervisão dos orientadores. O uso das ferramentas foi aprovado pelos orientadores, e os autores assumem integral responsabilidade pelo conteúdo final apresentado.

4 RESULTADOS

4.1 Seleção de Atributos e Configuração dos Modelos

O *pipeline* treinou 1.152 modelos no total, divididos em duas abordagens. Cada modelo corresponde a um algoritmo ajustado a uma combinação específica de pré-processamento, alvo e etapa de classificação, razão pela qual a cascata hierárquica, com triagem e subtipagem, contribui com dois modelos por alvo de três classes. Na abordagem *Flat*, foram 432 modelos: 2 *pipelines* × 3 variantes × 2 SMOTE × 3 alvos × 12 algoritmos. Na abordagem Hierárquica, foram 720 modelos, número maior porque a cascata tem duas etapas sequenciais. Para os alvos Apoio e Transverso (3 classes), cada configuração de pré-processamento treina 24 modelos, 12 para a triagem (Normal vs Alterado) e 12 para a subtipagem. Para o alvo Balanço, que é binário, são apenas 12 modelos por configuração. Multiplicando pelas 12 configurações de pré-processamento (2 *pipelines* × 3 variantes × 2 SMOTE), chega-se a 288 + 288 + 144 = 720 modelos. Os hiperparâmetros foram otimizados individualmente pelo Optuna com 50 tentativas por modelo, conforme descrito na seção 3.7.

Para o ranqueamento global (seção 4.3), a abordagem Hierárquica é consolidada em uma única linha por combinação (*pipeline* × variante × SMOTE × algoritmo × alvo), utilizando a sensibilidade efetiva (Equação 1) como métrica das duas etapas da cascata. Assim, os 720 modelos hierárquicos reduzem-se a 432 linhas, sem descarte de modelos: as duas etapas da cascata apenas passam a ser avaliadas por uma métrica única O ranqueamento opera, portanto, sobre 864 configurações de comparação (432 *Flat* + 432 Hierárquicas). Os 1.152 modelos foram todos treinados; o número menor reflete apenas a granularidade do ranqueamento, em que as duas etapas da cascata hierárquica são avaliadas como uma única configuração.

A seleção por consenso ("Com Consenso") resultou em subconjuntos distintos por *pipeline* e alvo, conforme apresentado na Tabela 12. No *Pipeline* Cinemática Angular (18 atributos candidatos), foram selecionados 7 atributos para Apoio, 15 para Balanço e 10 para Transverso. No *Pipeline* Cinemática Geral (25 atributos candidatos), foram selecionados 10 para Apoio, 19 para Balanço e 13 para Transverso. O limiar mais permissivo do Balanço (2/4 votos) explica o maior número de atributos retidos nesse alvo. A variante PCA reteve 12 componentes principais para o *Pipeline* Angular e 16 para o *Pipeline* Geral, ambos

preservando 95% da variância acumulada. As Figuras 15 a 20 apresentam os gráficos de votos por método para cada combinação *pipeline* × alvo no "Com Consenso".

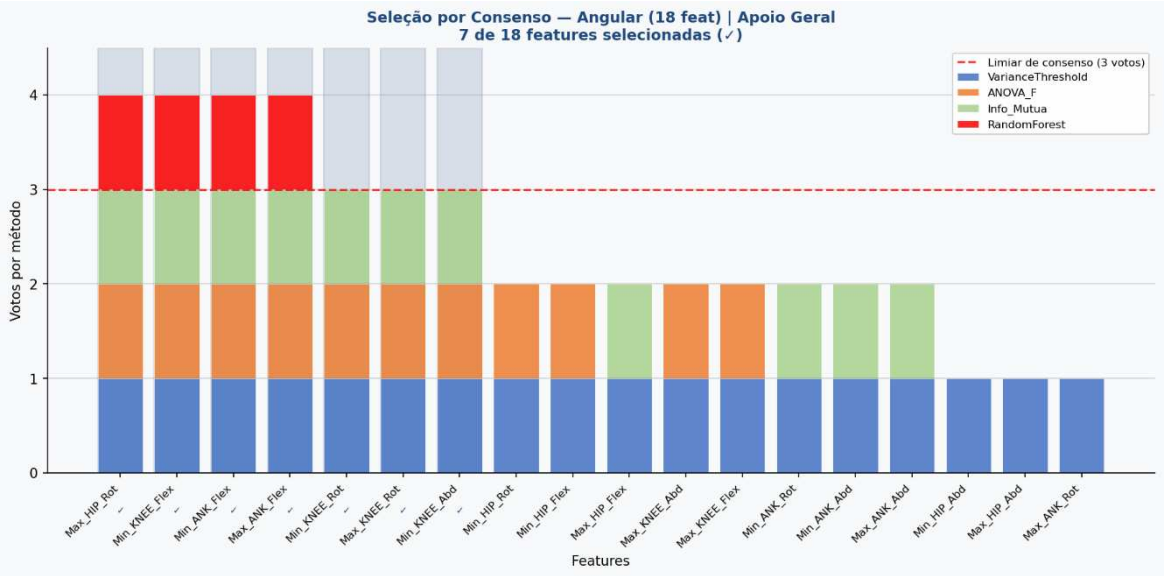

Figura 15 – Seleção por Consenso: *Pipeline* Cinemática Angular | Apoio Geral (7 de 18 atributos selecionados)
Nota: Barras coloridas representam os votos de cada método (azul = *VarianceThreshold*, laranja = ANOVA-F, verde = Informação Mútua, vermelho = *Random Forest*). A linha tracejada vermelha indica o limiar de consenso (3 votos). *Features* acima do limiar foram selecionadas para o *pipeline*.
Fonte: *Elaborada pelo autor.*

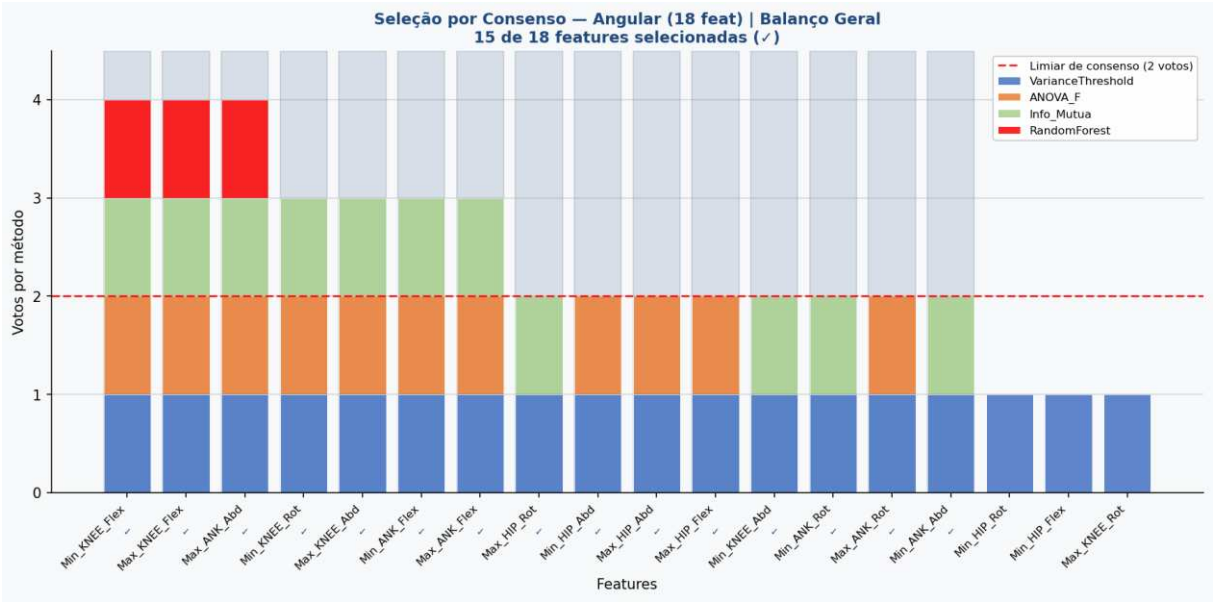

Figura 16 – Seleção por Consenso: *Pipeline* Cinemática Angular | Balanço Geral (15 de 18 atributos selecionados)
Nota: Limiar de consenso = 2 votos (problema binário – critério mais permissivo). Atributos com altura de barra igual ou superior ao limiar foram selecionadas.
Fonte: *Elaborada pelo autor.*

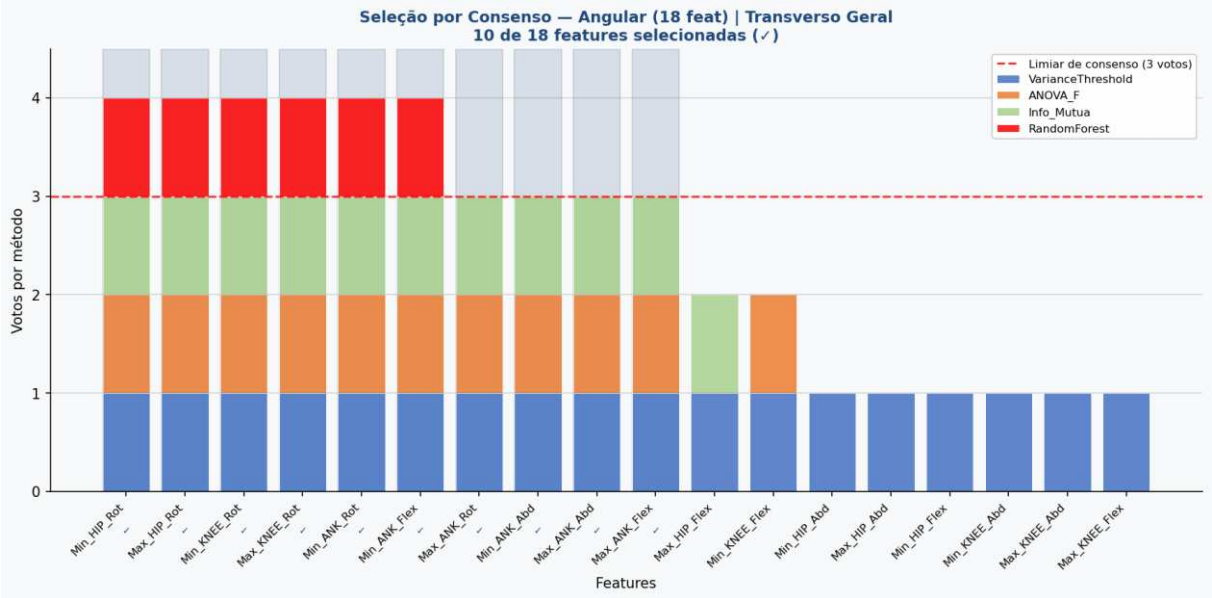

Figura 17 – Seleção por Consenso: *Pipeline* Cinemática Angular | Transverso Geral (10 de 18 atributos selecionados)
Nota: Limiar de consenso = 3 votos. A predominância de variáveis de rotação articular entre os selecionados é coerente com a natureza do alvo (plano transverso).
Fonte: *Elaborada pelo autor.*

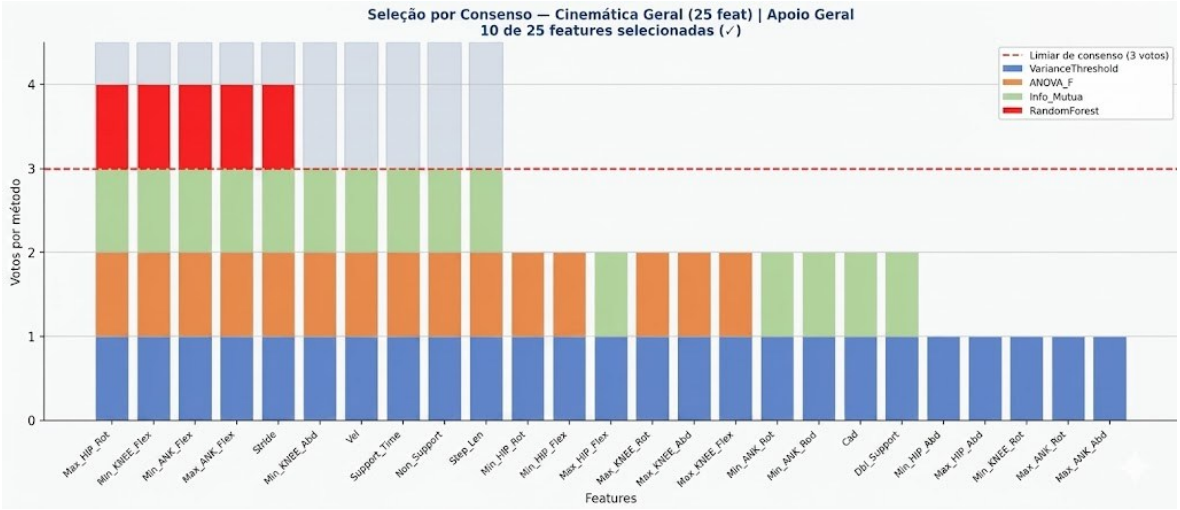

Figura 18 – Seleção por Consenso: *Pipeline* Cinemática Geral | Apoio Geral (10 de 25 atributos selecionados)
Nota: Limiar de consenso = 3 votos. Além das variáveis cinemáticas angulares, parâmetros espaço-temporais (*Stride, Vel, Support_Time, Non_Support, Step_Len*) atingiram o limiar de consenso.
Fonte: Elaborada pelo autor.

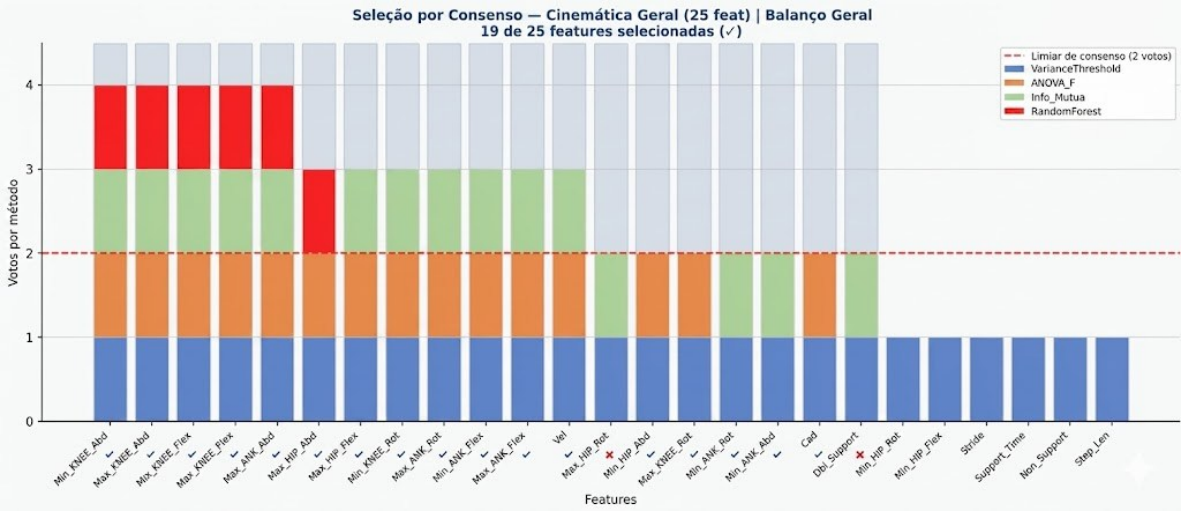

Figura 19 – Seleção por Consenso: *Pipeline* Cinemática Geral | Balanço Geral (19 de 25 atributos selecionados)
Nota: Limiar de consenso = 2 votos, critério mais permissivo adotado para problemas binários.
Fonte: *Elaborada pelo autor.*

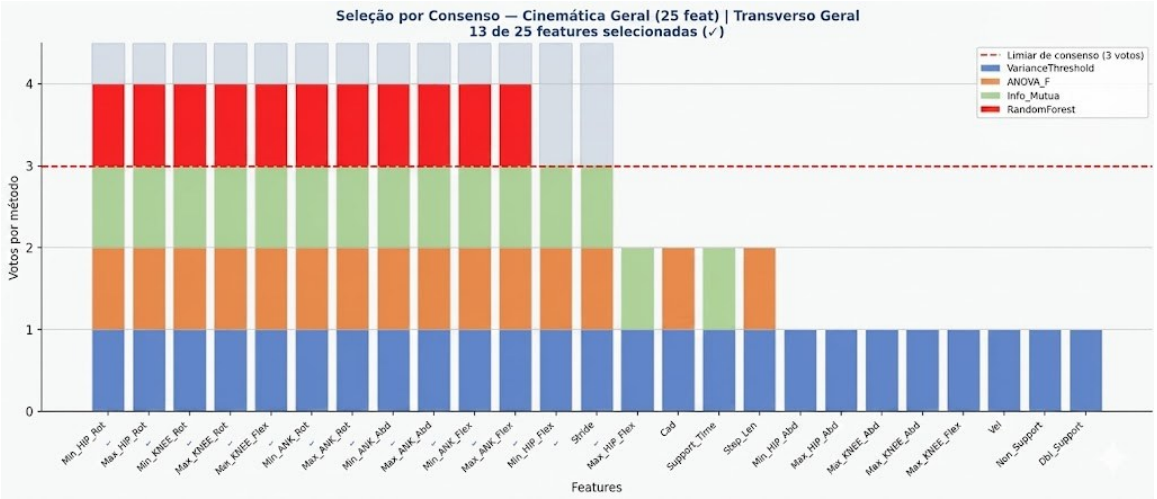

Figura 20 – Seleção por Consenso: *Pipeline* Cinemática Geral | Transverso Geral (13 de 25 atributos selecionados)
Nota: Limiar de consenso = 3 votos. As variáveis de rotação articular do quadril, joelho e tornozelo dominaram a seleção, confirmando a relevância biomecânica do plano transverso para este alvo.
Fonte: *Elaborada pelo autor.*

Três atributos foram selecionados em todas as seis configurações avaliadas: Max_ANK_Flex, Max_HIP_Rot e Min_ANK_Flex. Outras duas foram selecionadas em cinco das seis configurações: Min_KNEE_Flex e Min_KNEE_Rot. Esta consistência possui fundamentação biomecânica direta: a flexão do tornozelo (Max/Min_ANK_Flex) é determinante nos padrões Salto e Agachado; a rotação máxima do quadril (Max_HIP_Rot) é a variável mais discriminativa para transverso interno versus rotação externa no plano transverso;

e a flexão mínima do joelho (Min_KNEE_Flex) define classicamente o padrão rígido na fase de balanço.

4.1.1 Variáveis Mais Relevantes por Alvo

A análise de relevância das variáveis foi conduzida com base na consistência de seleção pelo método "Com Consenso" entre os quatro métodos independentes. Quanto maior o número de métodos que concordam sobre a relevância de uma variável, mais robusto é esse achado. A Tabela 16 apresenta as variáveis selecionadas pelos quatro métodos em ambos os fluxos de processamento para cada variável-alvo, com sua respectiva fundamentação clínica.

Tabela 16 – Principais variáveis selecionadas por alvo e fundamentação clínica

Alvos	Variáveis	Fundamentação Clínica
Apoio	Max_ANK_Flex / Min_ANK_Flex	Flexão máxima e mínima do tornozelo – distingue equino (SS) de calcâneo (SA)
Apoio	Min_KNEE_Flex	Flexão mínima do joelho – extensão excessiva em SS vs flexão aumentada em SA
Apoio	Max_HIP_Rot	Rotação máxima do quadril – compensação rotacional nos padrões SS e SA
Balanço	Min_KNEE_Flex / Max_KNEE_Flex	Flexão mínima e máxima do joelho – marcadores principais do padrão Rígido (BR)
Balanço	Min_KNEE_Abd / Max_KNEE_Abd	Abdução do joelho – associada ao padrão de marcha em tesoura na PC espástica
Transverso	Min_HIP_Rot / Max_HIP_Rot	Rotação mínima e máxima do quadril – marcadores de TI e TE respectivamente
Transverso	Min_KNEE_Rot / Max_KNEE_Rot	Rotação do joelho – desvio rotacional no plano transverso
Transverso	Min_ANK_Rot / Max_ANK_Rot	Rotação do tornozelo – compensação rotacional distal

Nota: Atributos aprovados pelos quatro métodos de consenso (*VarianceThres hold*, ANOVA-F, Informação Mútua e Floresta Aleatória) em ambos os *pipelines*. HIP = quadril; KNEE = joelho; ANK = tornozelo; Flex = flexão; Rot = rotação; Abd = abdução; SS = Sagital Salto; SA = Sagital Agachado; BR = Balanço Rígido; TI = Transverso interno; TE = Transverso Externo.
Fonte: *Elaborada pelo autor.*

4.2 Resultados do Conjunto de Teste

Os resultados apresentados nas seções a seguir representam o desempenho dos modelos aplicados ao conjunto de teste fixo (760 registros de 29 pacientes para Balanço e Transverso; 754 registros para Apoio), que permaneceu completamente isolado durante todo o processo de treinamento. Estes constituem os resultados principais do estudo, sendo a base para as conclusões sobre aplicabilidade clínica.

4.3 Ranqueamento Global dos Modelos

O ranqueamento global ordenou as 864 configurações de modelagem pela Acurácia Balanceada no conjunto de teste, correspondentes às 432 combinações *Flat* e às 432 combinações Hierárquicas (cada uma agregando as etapas E1 e E2 da cascata via sensibilidade efetiva, conforme Equação 1). O alvo Transverso ocupou integralmente as dez melhores colocações. Entre os modelos de fusão*, Voting Soft* e *Voting Hard* apareceram em 4 das 10 primeiras posições.

Como a concentração das posições superiores em um único alvo limita o valor informativo de uma tabela com os dez melhores modelos globais, optou-se por apresentar na Tabela 17 o melhor modelo de cada variável-alvo, permitindo a comparação direta entre alvos.

Tabela 17 – Melhor modelo por variável-alvo no conjunto de teste

Alvo	*Pipeline*	Variante	SMOTE	Modelo	Abord.	Bal.Acc	MCC	AUC	*Overfit*
Transverso	Cin. Geral	PCA	"Sem SMOTE"	*LogReg*	*Flat*	0,791	0,507	0,874	0,036
Apoio	Cin. Angular	"Com Consenso"	"Com SMOTE"	*Naive Bayes*	*Flat*	0,717	0,519	0,813	0,082
Balanço	Cin. Angular	"Com Consenso"	"Sem SMOTE"	*DTree*	Hierárquico	0,702	0,359	0,731	0,263

Nota: Melhor combinação por variável-alvo ordenada por Acurácia Balanceada no conjunto de teste. Bal.Acc = Acurácia Balanceada; MCC = *Matthews Correlation Coefficient*; AUC = Área sob a curva ROC (*macro* OvR); Overfit. = $\text{Bal.Acc}_{\text{Treino}} - \text{Bal.Acc}_{\text{Teste}}$. Cin. = Cinemática; Abord. = Abordagem.
Fonte: Elaborada pelo autor.

O melhor desempenho global foi obtido pela Regressão Logística com PCA para o alvo Transverso (Acurácia Balanceada = 0,791, G-*Mean* = 0,778). As Seções 4.4 a 4.6 apresentam o ranqueamento detalhado por variável-alvo.

4.4 Ranqueamento por alvo: Transverso Geral

O alvo Transverso apresentou o melhor desempenho absoluto entre os três alvos avaliados. O melhor modelo foi a Regressão Logística com *Pipeline* Cinemática Geral, PCA e "Sem SMOTE" (abordagem *Flat*), alcançando Acurácia Balanceada de 0,791, G-Mean de 0,778, AUC de 0,874, MCC de 0,507 e Kappa de 0,491 com *overfitting* de apenas 0,036. Os cinco melhores modelos para este alvo são apresentados na Tabela 18.

Tabela 18 – 5 Melhores: Transverso Geral por Acurácia Balanceada no conjunto de teste

Pos.	Variante	SMOTE	Modelo	Bal.Acc	MCC	Kappa	AUC	Sens TN	Sens TI	Sens TE
1°	PCA	"Sem SMOTE"	*LogReg*	0,791	0,507	0,491	0,874	70,9%	66,5%	100,0%
2°	PCA	"Com SMOTE"	*LogReg*	0,791	0,507	0,491	0,870	70,7%	66,5%	100,0%
3°	"Sem Consenso"	"Com SMOTE"	*LogReg*	0,780	0,474	0,461	0,849	69,2%	66,5%	98,2%
4°	"Sem Consenso"	"Sem SMOTE"	*Voting Soft*	0,772	0,506	0,504	0,836	77,6%	62,9%	91,2%
5°	"Sem Consenso"	"Sem SMOTE"	*LogReg*	0,772	0,455	0,440	0,849	66,8%	66,5%	98,2%

Nota: Todos os modelos referem-se ao *Pipeline* Cinemática Geral, abordagem *Flat*. Bal.Acc = Acurácia Balanceada; MCC = Matthews Correlation Coefficient; AUC = Área sob a curva ROC (*macro* OvR); Sens = Sensibilidade por classe. TN = Transverso Normal; TI = Transverso interno; TE = Transverso Externo.
Fonte: Elaborada pelo autor.

Vale destacar que as duas primeiras posições do ranqueamento (PCA + "Sem SMOTE" e PCA + "Com SMOTE", ambas com Regressão Logística) apresentam desempenho estatisticamente equivalente em todas as métricas avaliadas, com diferenças inferiores a 0,001 em Acurácia Balanceada, MCC e Kappa.

A análise por classe revelou assimetria importante na sensibilidade da classe Transverso Externo (TE) atingiu sensibilidade de 100% (57/57 casos corretamente classificados), enquanto Transverso Normal (TN) e Transverso interno (TI) apresentaram sensibilidades de 70,9% e 66,5%, respectivamente. A distribuição entre as classes no conjunto de teste foi de 509 registros para TN, 194 para TI e 57 para TE. A matriz de confusão do melhor modelo é apresentada na Figura 21.

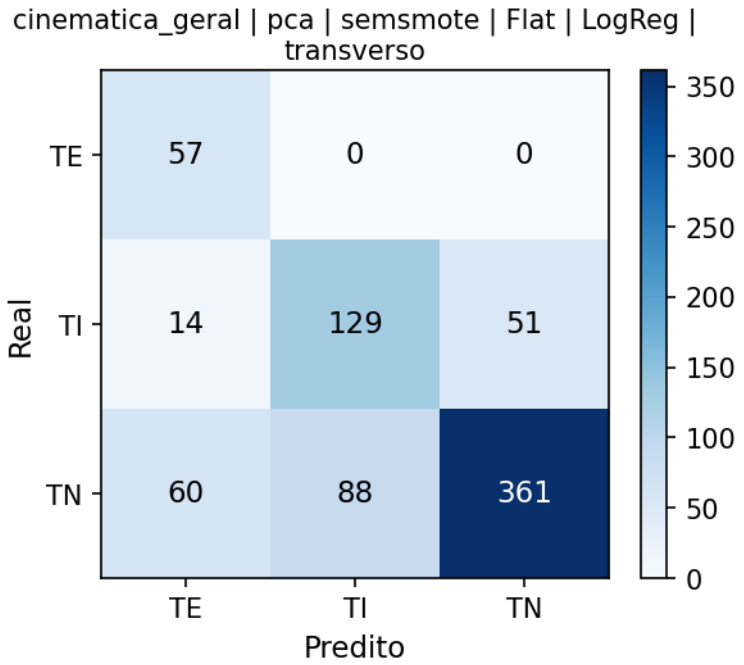

Figura 21 – Matriz de confusão do modelo vencedor para o alvo Transverso "Sem SMOTE"
Nota: A matriz apresenta a distribuição das predições do modelo Regressão Logística (*Pipeline* Cinemática Geral, variante PCA, "Sem SMOTE", abordagem *Flat*) no conjunto de teste. Linhas representam as classes reais e colunas as classes preditas. Os valores na diagonal principal indicam classificações corretas. TN = Transverso Normal; TI = Transverso interno; TE = Transverso externo.
Fonte: *Elaborada pelo autor.*

A Figura 22 apresenta os intervalos de confiança (IC 95%, 1.000 replicações) dos dez melhores modelos para o alvo Transverso. As larguras dos intervalos variaram entre 0,051 e 0,080. O melhor modelo (*LogReg* / Geral / PCA / "Sem SMOTE") atingiu Acurácia Balanceada = 0,791, com IC 95% entre 0,764 e 0,818 (largura = 0,054). A Figura 23 apresenta a AUC ROC *macro* dos dez melhores modelos com o melhor resultado de 0,874, situado na faixa considerada muito boa (0,80–0,90).

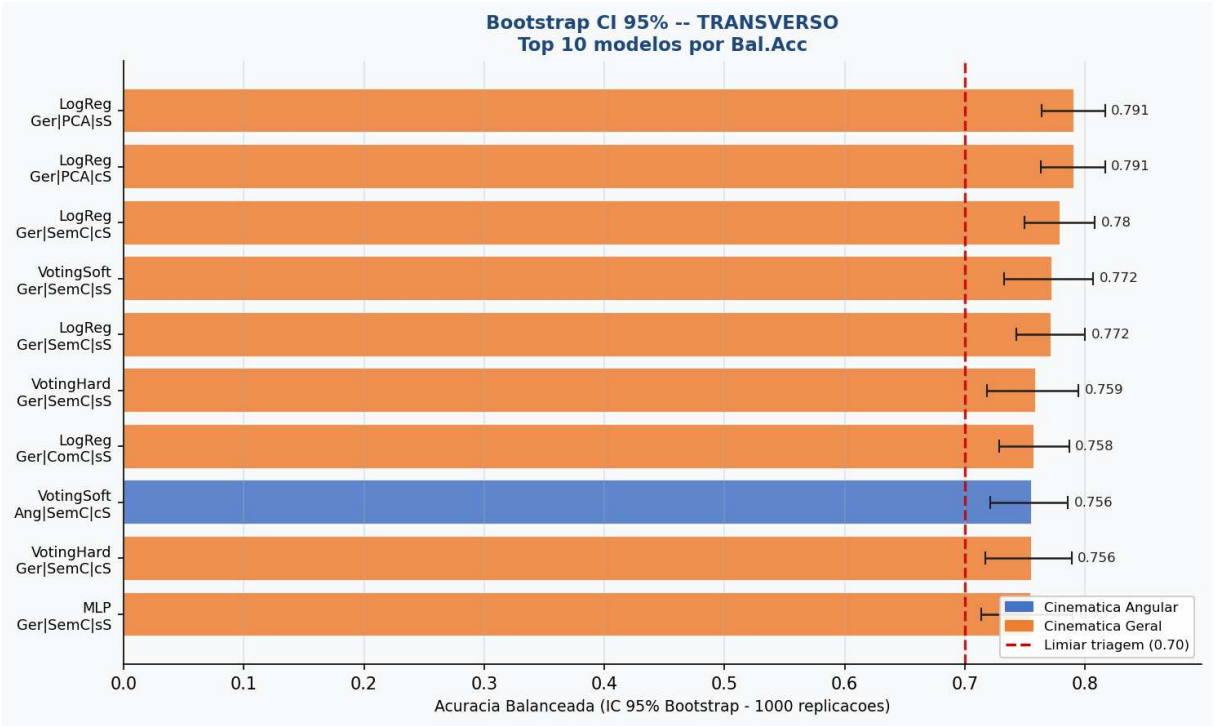

Figura 22 – Intervalos de confiança *Bootstrap* 95%, dez melhores modelos: Transverso
Nota: Barras representam a Acurácia Balanceada observada. As linhas de erro indicam o intervalo de confiança de 95% calculado por *Bootstrap* com 1.000 replicações. A linha tracejada vermelha indica o limiar mínimo para triagem clínica (0,70). Azul = Pipeline Cinemática Angular; Laranja = Pipeline Cinemática Geral. Ang = Cinemática Angular; Ger = Cinemática Geral; SemC = "Sem Consenso"; ComC = "Com Consenso"; PCA = Análise de Componentes Principais; cS = "Com SMOTE"; sS = "Sem SMOTE".
Fonte: *Elaborada pelo autor.*

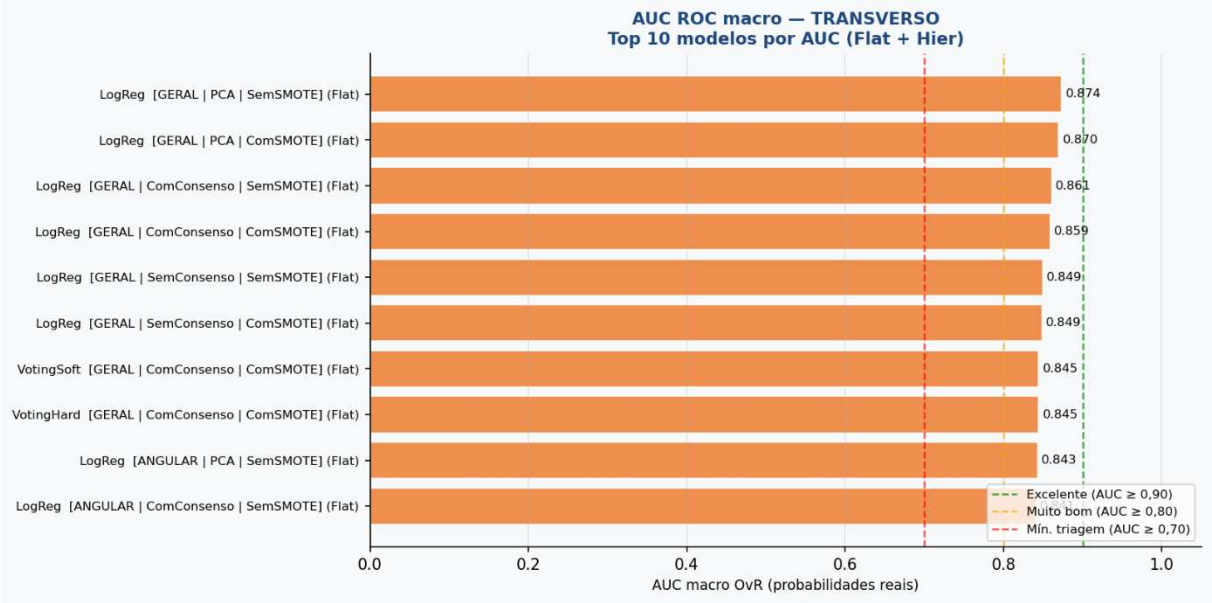

Figura 23 – AUC ROC macro – Transverso Geral
Nota: Barras representam a AUC ROC macro (*One-vs-Rest*) no conjunto de teste. A linha tracejada vermelha indica o limite inferior da faixa aceitável (AUC ≥ 0,70); a linha tracejada laranja indica desempenho excelente (AUC ≥ 0,80); a linha tracejada verde indica desempenho excepcional (AUC ≥ 0,90), conforme escala de Hosmer e Lemesh. Laranja = modelos com AUC ≥ 0,80. Todos os dez modelos do top 10 são da abordagem *Flat* e estão na faixa de desempenho muito bom (0,841–0,874), com predomínio da Regressão Logística no *Pipeline* Cinemática Geral (8 das 10 posições). A ausência de modelos Hierárquicos no top 10 é coerente com a maior separabilidade linear dos padrões transversais (Rotação Externa, Neutra e Interna) no espaço cinemático, que favorece classificadores multiclasse diretos sobre arquiteturas em cascata. A AUC é independente do limiar de decisão, complementando a Acurácia Balanceada como métrica de avaliação. Ang = Cinemática Angular; Ger = Cinemática Geral; "Sem Consenso" = seleção por consenso quando ≥ 3 dos 4 métodos concordam; "Com Consenso" = seleção quando ≥ 2 dos 4 métodos concordam; PCA = Análise de Componentes Principais; "Com SMOTE" = balanceamento sintético aplicado; "Sem SMOTE" = sem balanceamento; *Flat* = abordagem multiclasse direta.
Fonte: *Elaborada pelo autor.*

Nos 10 melhores do Transverso, a Regressão Logística foi o modelo mais frequente (5 posições), seguida por *Voting Soft* e *Voting Hard* (2 cada) e MLP (1). O *Pipeline* Cinemática Geral dominou com 9 das 10 posições, com apenas uma configuração baseada em Cinemática Angular. A variante "Sem Consenso" foi a mais frequente (7 de 10), seguida por PCA (2 de 10) e "Com Consenso" (1 de 10).

Tabela 19 – Análise Comparativa de Desempenho por Classe: Transverso

Abordagem	TN	TI	TE	Bal.Acc
PCA	**70,9%**	**66,5%**	**100,0%**	**0,791**
"Sem Consenso"	69,2%	**66,5%**	98,2%	0,780
"Com Consenso"	67,8%	61,3%	98,2%	0,758
Hierárquico	**72,9%**	54,6%	58,0%	0,618

Nota: valores de TN, TI e TE correspondem à sensibilidade por classe, em porcentagem. Células em verde indicam melhor desempenho por coluna. TN = Transverso Normal, TI = Transverso Interno, TE = Transverso Externo. Bal.Acc = média das sensibilidades por classe.
Fonte: *Elaborada pelo autor.*

4.5 Ranqueamento por alvo: Apoio Geral

O alvo Apoio apresentou desempenho inferior ao Transverso, porém o melhor modelo superou o limiar mínimo de triagem clínica (Acurácia Balanceada ≥ 0,70). O melhor modelo foi o *Naive Bayes* com *Pipeline* Cinemática Angular, variante "Com Consenso" e "Com SMOTE" (abordagem *Flat*), alcançando Acurácia Balanceada de 0,717, G-Mean de 0,708, AUC de 0,813, MCC de 0,519 e Kappa de 0,490, com *overfitting* de 0,082. Os cinco melhores modelos para este alvo são apresentados na Tabela 20.

Tabela 20 – 5 Melhores: Apoio por Acurácia Balanceada no conjunto de teste

Pos.	Variante	SMOTE	Modelo	Bal.Acc	MCC	Kappa	AUC	Sens SN	Sens SS	Sens SA
1°	"Com Consenso"	"Com SMOTE"	*Naive Bayes*	0,717	0,519	0,490	0,813	81,2%	56,6%	77,3%
2°	"Sem Consenso"	"Com SMOTE"	*Naive Bayes*	0,673	0,461	0,432	0,793	84,2%	52,1%	65,6%
3°	"Com Consenso"	"Sem SMOTE"	*Naive Bayes*	0,664	0,468	0,465	0,814	66,8%	69,8%	62,5%
4°	"Com Consenso"	"Com SMOTE"	*Naive Bayes**	0,651	0,444	0,408	0,819	82,2%	55,9%	57,2%
5°	"Sem Consenso"	"Sem SMOTE"	*LogReg* †	0,650	0,419	0,408	0,783	74,3%	57,3%	63,3%

Nota: Posições 1–4: *Pipeline* Cinemática Angular, abordagem *Flat*. *Posição 4: abordagem Hierárquica. †Posição 5: *Pipeline* Cinemática Geral, abordagem *Flat*. Bal.Acc = Acurácia Balanceada; MCC = *Matthews Correlation Coefficient*; AUC = Área sob a curva ROC (*macro* OvR); Sens = Sensibilidade por classe. SN = Sagital Normal; SS = Sagital Salto; SA = Sagital Agachado.

Fonte: Elaborada pelo autor.

A análise por classe revelou um padrão de sensibilidades inversamente proporcional ao tamanho das classes: a classe SS, majoritária no conjunto de teste com 424 registros (56,2% do total), apresentou a menor sensibilidade (56,6%), enquanto SN (202 registros; 26,8%) e SA (128 registros; 17,0%) atingiram 81,2% e 77,3%, respectivamente. A matriz de confusão do modelo vencedor do alvo Apoio (*Naive Bayes* / *Pipeline* Cinemática Angular / "Com Consenso" / "Com SMOTE" / *Flat*) é apresentada na Figura 24.

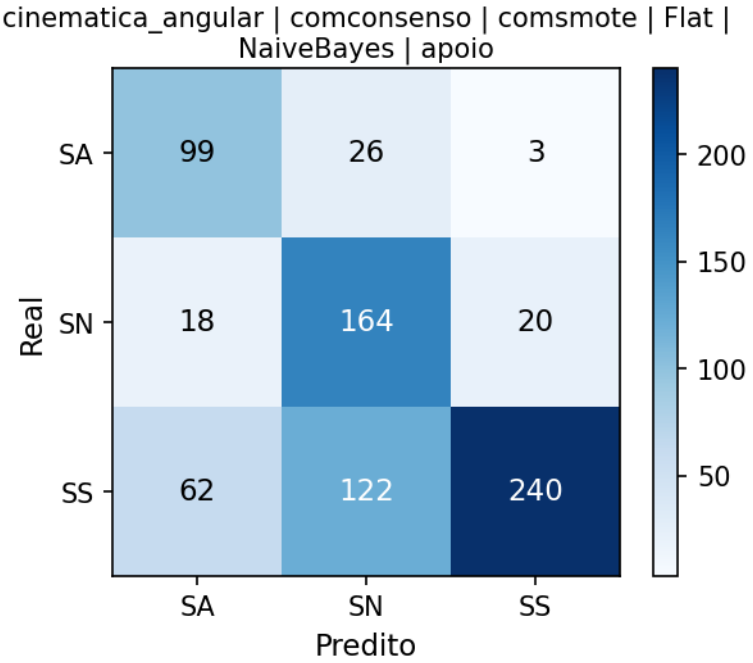

Figura 24 – Matriz de confusão do modelo vencedor para o alvo Apoio
Nota: A matriz apresenta a distribuição das predições do modelo *Naive Bayes* (*Pipeline* Cinemática Angular, variante "Com Consenso", "Com SMOTE", abordagem *Flat*) no conjunto de teste. Linhas representam as classes reais e colunas as classes preditas. Os valores na diagonal principal indicam classificações corretas. SN = Sagital Normal; SS = Sagital Salto; SA = Sagital Agachado.
Fonte: *Elaborada pelo autor.*

A Figura 25 apresenta os intervalos de confiança *Bootstrap* (IC 95%, 1.000 replicações) dos dez melhores modelos para o alvo Apoio. As larguras dos intervalos variaram entre 0,068 e 0,081. O melhor modelo (*Naive Bayes* / Angular / "Com Consenso" / "Com SMOTE") atingiu Acurácia Balanceada de 0,717, com IC 95% entre 0,684 e 0,752 (largura = 0,068).

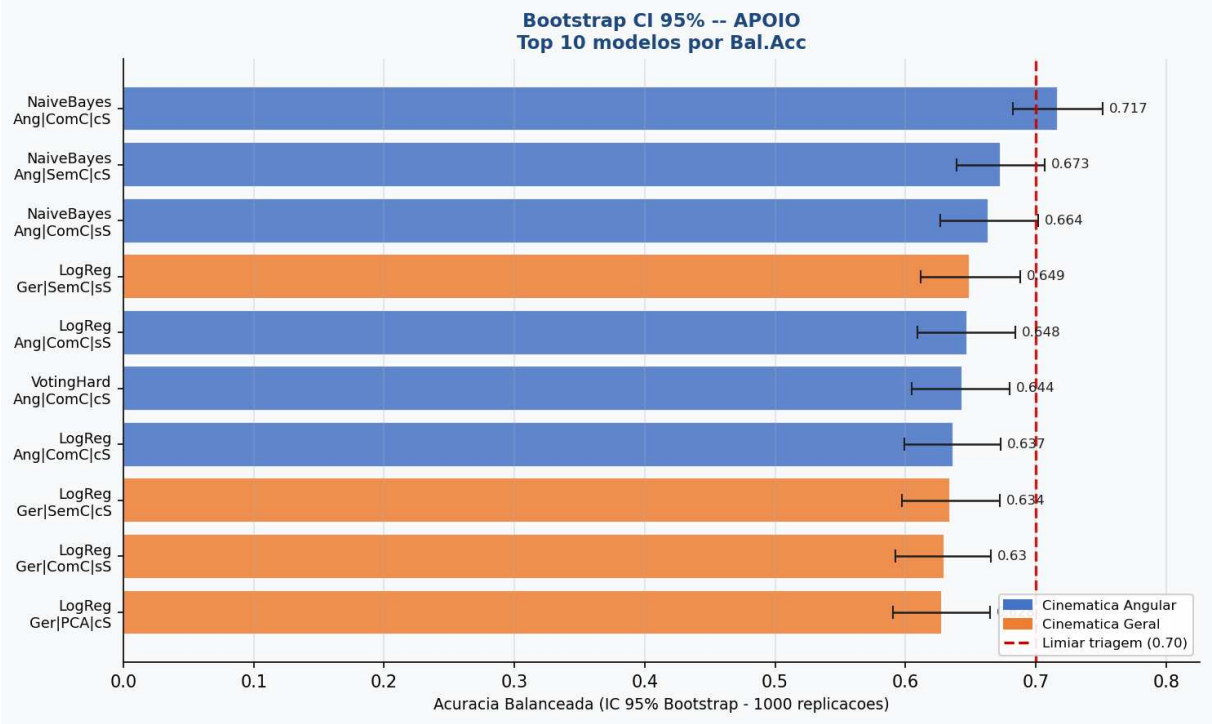

Figura 25 – Intervalos de confiança *Bootstrap* 95%, dez melhores modelos: Apoio
Nota: Barras representam a Acurácia Balanceada observada. As linhas de erro indicam o intervalo de confiança de 95% calculado por *Bootstrap* com 1.000 replicações. A linha tracejada vermelha indica o limiar mínimo para triagem clínica (0,70). Azul = Pipeline Cinemática Angular; Laranja = *Pipeline* Cinemática Geral. Ang = Cinemática Angular; Ger = Cinemática Geral; ComC = "Com Consenso"; SemC = "Sem Consenso"; PCA = Análise de Componentes Principais; cS = "Com SMOTE"; sS = "Sem SMOTE".
Fonte: *Elaborada pelo autor.*

Nos 10 melhores do alvo Apoio por Acurácia Balanceada, a Regressão Logística foi o modelo mais frequente em número absoluto (5 posições), embora o *Naive Bayes* ocupe as quatro primeiras colocações. O *Voting Hard* aparece em uma posição. O *Pipeline* Cinemática Angular dominou com 7 das 10 posições, e a variante "Com Consenso" foi a mais frequente (7 de 10), seguida por "Sem Consenso" (3 de 10). Notavelmente, a variante PCA não apareceu entre os dez melhores modelos por Acurácia Balanceada para este alvo, em contraste com o alvo Transverso. A Figura 26 apresenta a AUC ROC macro dos dez melhores modelos para este alvo. O modelo de maior Acurácia Balanceada (*Naive Bayes* / Angular / "Com Consenso" / "Com SMOTE", Acurácia Balanceada = 0,717) apresentou AUC = 0,813, situada na faixa considerada muito boa (0,80–0,90), confirmando boa capacidade discriminativa mesmo em cenário de classes desbalanceadas. Considerando o ranqueamento por AUC, o melhor valor foi 0,885 (*Naive Bayes* / Geral / "Sem Consenso" / "Com SMOTE" / Hierárquico), embora esse modelo tenha apresentado Acurácia Balanceada inferior (0,579). O *Naive Bayes* apareceu em 8 das 10 posições do ranqueamento por AUC.

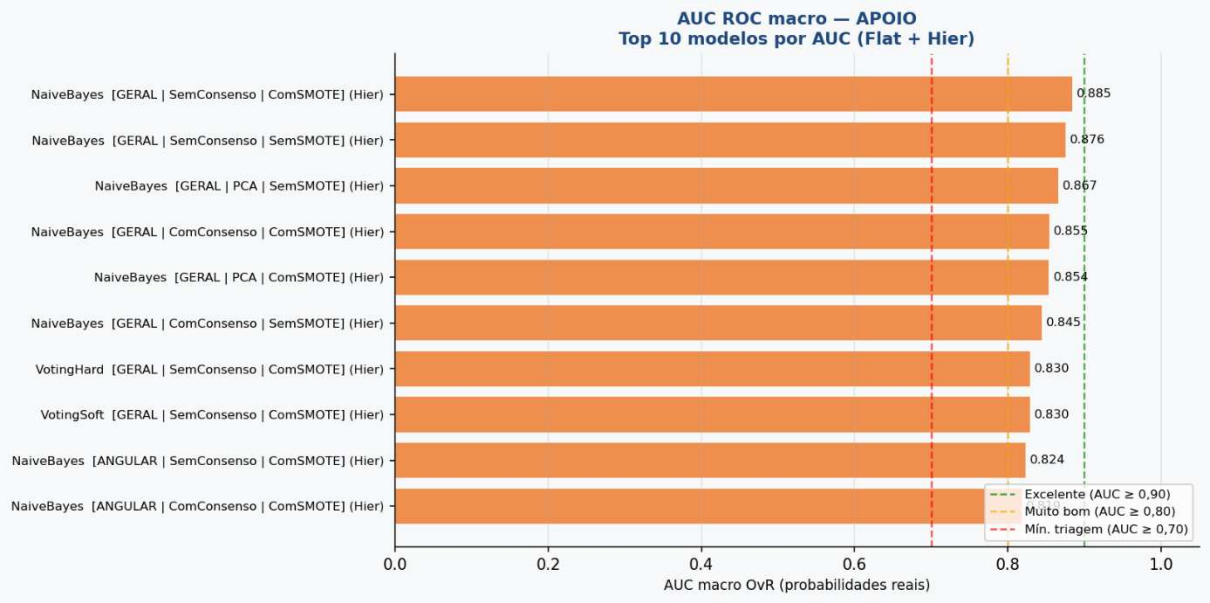

Figura 26 – AUC ROC macro – Apoio
Nota: Barras representam a AUC ROC macro (*One-vs-Rest*) no conjunto de teste. A linha tracejada vermelha indica o limite inferior da faixa aceitável (AUC ≥ 0,70); a linha tracejada laranja indica desempenho excelente (AUC ≥ 0,80); a linha tracejada verde indica desempenho excepcional (AUC ≥ 0,90), conforme escala de Hosmer e Lemesh. Verde = modelos com AUC ≥ 0,70; Laranja = AUC ≥ 0,80; Azul = AUC ≥ 0,90; Vermelho = abaixo de 0,70. Todos os dez modelos do top 10 são da abordagem Hierárquica e estão na faixa de desempenho muito bom (0,824–0,885), com domínio do *Naive Bayes* (8 das 10 posições), confirmando a adequação desse classificador às características de distribuição das variáveis do alvo Apoio. A AUC é independente do limiar de decisão, complementando a Acurácia Balanceada como métrica de avaliação. Ang = Cinemática Angular; Ger = Cinemática Geral; "Sem Consenso" = seleção por consenso quando ≥ 3 dos 4 métodos concordam; "Com Consenso" = seleção quando ≥ 2 dos 4 métodos concordam; PCA = Análise de Componentes Principais; "Com SMOTE" = balanceamento sintético aplicado; "Sem SMOTE" = sem balanceamento; Hier = abordagem hierárquica em cascata.
Fonte: *Elaborada pelo autor.*

Complementando o ranqueamento por Acurácia Balanceada (Tabela 20) e a matriz de confusão do modelo vencedor (Figura 24), a Tabela 21 apresenta a decomposição do desempenho por classe para o alvo Apoio.

Tabela 21 – Análise Comparativa de Desempenho por Classe: Apoio

Abordagem	SN	SS	SA	Bal.Acc
"Com Consenso"	81,2%	56,6%	**77,3%**	**0,717**
"Sem Consenso"	**84,2%**	52,1%	65,6%	0,673
PCA	71,8%	56,4%	60,2%	0,628
Hierárquico	82,2%	55,9%	57,2%	0,651

Nota: valores de SN, SS e SA correspondem à sensibilidade por classe, em porcentagem. Células em verde indicam melhor desempenho por coluna. SN = Sagital Normal, SS = Sagital Salto, SA = Sagital Agachado. Bal.Acc = média das sensibilidades por classe.

Fonte: *Elaborada pelo autor.*

4.6 Ranqueamento por alvo: Balanço Geral

O alvo Balanço apresentou o menor número de classes entre os três alvos avaliados (problema binário entre Balanço Normal (BN) e Balanço Rígido (BR), porém o maior desbalanceamento relativo, especialmente em cenários com forte desbalanceamento entre classes como o do alvo Balanço (razão aproximada de 7,9:1 entre BN e BR). O melhor modelo foi a Árvore de Decisão com *Pipeline* Cinemática Angular, variante "Com Consenso" e "Sem SMOTE" (abordagem Hierárquica), alcançando Acurácia Balanceada de 0,702, G-*Mean* de 0,674, AUC de 0,731, MCC de 0,359 e Kappa de 0,354, com *overfitting* de 0,263. Os cinco melhores modelos para este alvo são apresentados na Tabela 20.

Tabela 22 – 5 Melhores: Balanço por Acurácia Balanceada no conjunto de teste

Pos.	Variante	SMOTE	Abordagem	Modelo	Bal.Acc	MCC	Kappa	AUC	Sens BN	Sens BR
1°	"Com Consenso"	"Sem SMOTE"	Hierárquico	*DTree*	0,702	0,359	0,354	0,731	89,8%	50,6%
2°	PCA	"Sem SMOTE"	Hierárquico	*LogReg*	0,674	0,256	0,230	0,639	79,6%	55,3%
3°	PCA	"Sem SMOTE"	*Flat*	*LogReg*	0,674	0,256	0,230	0,639	79,6%	55,3%
4°†	PCA	"Sem SMOTE"	Hierárquico	*LogReg*	0,667	0,245	0,220	0,632	79,3%	54,1%
5°†	PCA	"Sem SMOTE"	*Flat*	*LogReg*	0,667	0,245	0,220	0,631	79,3%	54,1%

Nota: Posições 1–3: *Pipeline* Cinemática Angular. †Posições 4–5: *Pipeline* Cinemática Geral. Bal.Acc = Acurácia Balanceada; MCC = *Matthews Correlation Coefficient*; AUC = Área sob a curva ROC (*macro* OvR); Sens = Sensibilidade por classe. BN = Balanço Normal; BR = Balanço Rígido. *Overfitting* do melhor modelo = 0,263. Posições 2 e 3, e posições 4 e 5, apresentam métricas idênticas por equivalência metodológica entre abordagens *Flat* e Hierárquica em problemas binários.

Fonte: *Elaborada pelo autor.*

A análise por classe revelou assimetria acentuada na sensibilidade. A classe BN atingiu 89,8% (606/675 casos corretamente classificados), enquanto BR apresentou apenas 50,6% (43/85 casos). As matrizes de confusão do melhor modelo *Flat* (*LogReg* / Angular / PCA / "Sem SMOTE", Figura 27) e do melhor modelo Hierárquico (*DTree* / Angular / "Com Consenso" / "Sem SMOTE", Figura 28) são apresentadas a seguir. O *DTree* Hierárquico acertou 606 casos de BN (vs. 537 do *LogReg Flat*), e o *LogReg Flat* acertou 47 casos de BR (vs. 43 do *DTree* Hierárquico).

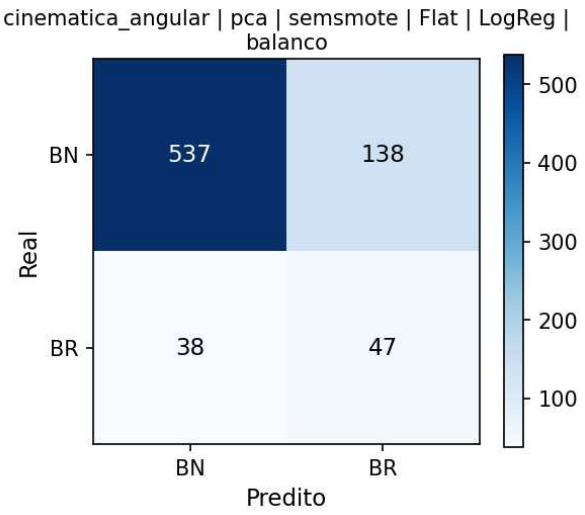

Figura 27 – Matriz de confusão do melhor modelo *Flat* para o alvo Balanço
Nota: A matriz apresenta a distribuição das predições do modelo Regressão Logística (*Pipeline* Cinemática Angular, variante PCA, "Sem SMOTE", abordagem *Flat*) no conjunto de teste. Linhas representam as classes reais e colunas as classes preditas. Os valores na diagonal principal indicam classificações corretas. BN = Balanço Normal; BR = Balanço Rígido.
Fonte: *Elaborada pelo autor.*

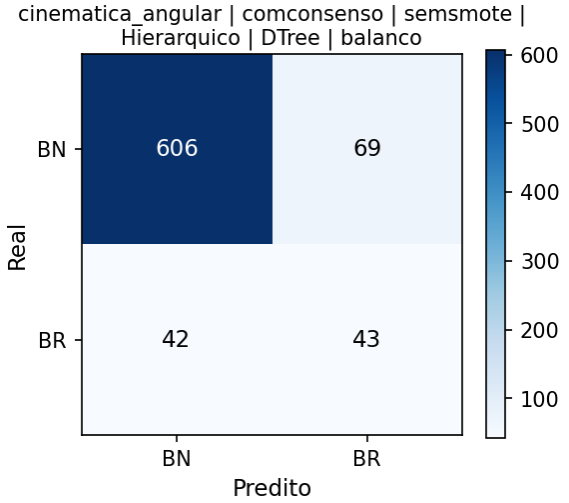

Figura 28 – Matriz de confusão do melhor modelo Hierárquico para o alvo Balanço
Nota: A matriz apresenta a distribuição das predições do modelo Árvore de Decisão (*Pipeline* Cinemática Angular, variante "Com Consenso", "Sem SMOTE", abordagem Hierárquica) no conjunto de teste. Linhas representam as classes reais e colunas as classes preditas. Os valores na diagonal principal indicam classificações corretas. BN = Balanço Normal; BR = Balanço Rígido.
Fonte: *Elaborada pelo autor.*

A Figura 29 apresenta os intervalos de confiança (IC 95%, 1.000 replicações) dos dez melhores modelos para o alvo Balanço, incluindo as abordagens *Flat* e Hierárquica. O

modelo vencedor (*DTree* / Angular / "Com Consenso" / "Sem SMOTE" / Hierárquico) com Acurácia Balanceada de 0,702 é o único a superar o limiar de triagem clínica; entretanto, seu IC 95% (0,645–0,756; largura de 0,111, o mais amplo entre os três alvos) contém o valor de 0,70, indicando que a superação do limiar não é robusta. A abordagem Hierárquica concentrou 6 das 10 posições do ranqueamento, enquanto a redução por PCA predominou na etapa de redução de dimensionalidade, respondendo por 8 das 10 posições – apenas o modelo vencedor (*DTree* / "Com Consenso") e a sexta posição (*DTree* / Geral / "Com Conseno") escaparam a esse padrão. A variante "Sem Consenso" não apareceu entre os 10 melhores para o alvo Balanço. Observa-se também que as posições 2 e 3, assim como 4 e 5 da Tabela 22, apresentam métricas idênticas entre as abordagens *Flat* e Hierárquica.

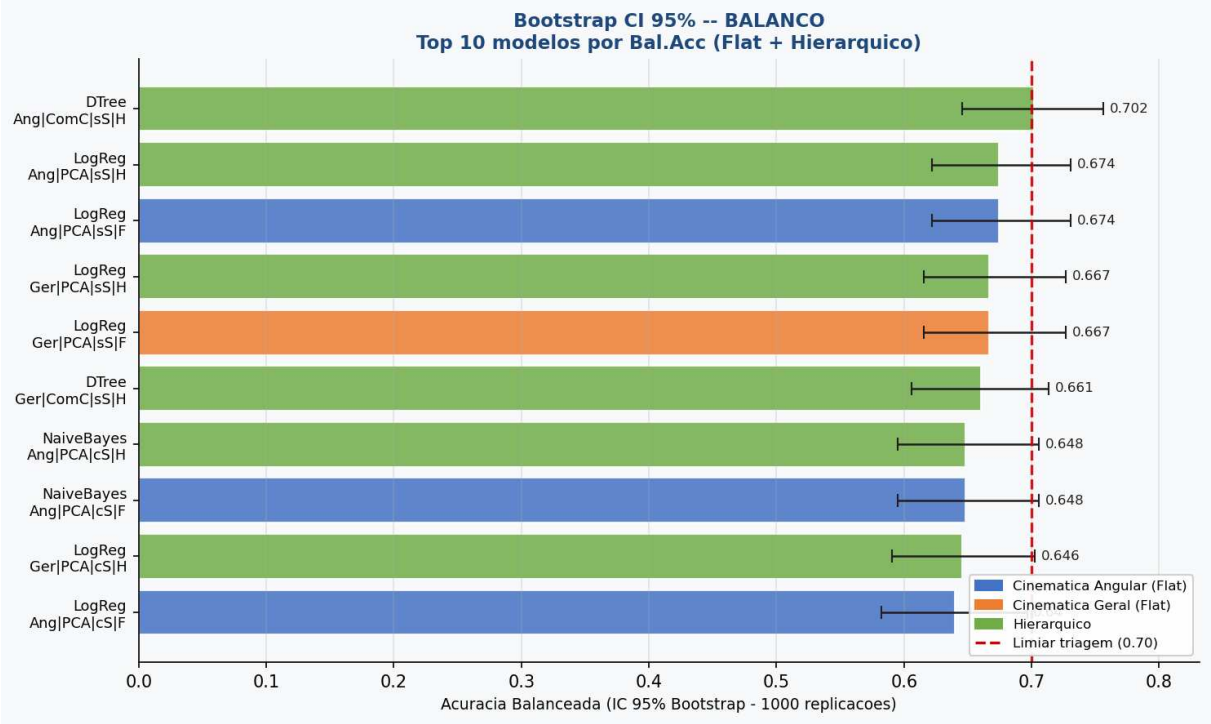

Figura 29 – Intervalos de confiança *Bootstrap* 95%, dez melhores modelos: Balanço
Nota: Barras representam a Acurácia Balanceada observada. As linhas de erro indicam o intervalo de confiança de 95% calculado por *Bootstrap* com 1.000 replicações. A linha tracejada vermelha indica o limiar mínimo para triagem clínica (0,70). Azul = *Pipeline* Cinemática Angular (*Flat*); Laranja = *Pipeline* Cinemática Geral (*Flat*); Verde = abordagem Hierárquica. Ang = Cinemática Angular; Ger = Cinemática Geral; ComC = "Com Consenso"; SemC = "Sem Consenso"; PCA = Análise de Componentes Principais; cS = "Com SMOTE"; sS = "Sem SMOTE"; F = *Flat*; H = Hierárquico.
Fonte: *Elaborada pelo autor.*

A Figura 30 apresenta a AUC ROC macro dos dez melhores modelos por AUC para o alvo Balanço, combinando as abordagens *Flat* e Hierárquica e complementando o ranqueamento principal por Acurácia Balanceada. O melhor desempenho do Top 10 coincide com o modelo vencedor por Acurácia Balanceada (*DTree* / Angular / "Com Consenso" / "Sem SMOTE" / Hierárquico, AUC = 0,731 e Acurácia Balanceada = 0,702), reforçando a robustez desse modelo em ambas as métricas. Os demais nove modelos da seleção são fortemente dominados por *gradient boosting* e Floresta Aleatória, *LightGBM, XGBoost* e *RandomForest* correspondem a 8 das 10 posições, com AUC entre 0,701 e 0,724, todos na faixa aceitável (0,70–0,80), mas Acurácia Balanceada substancialmente inferior (entre 0,49 e 0,59) à do modelo vencedor. Esse padrão é metodologicamente revelador. Indica que múltiplos modelos *gradient boosting* atingem boa capacidade probabilística de discriminação entre BN e BR, mas não convertem essa discriminação em classificação balanceada no ponto de corte padrão de 0,5. Reforça-se, assim, a importância de avaliar AUC e Acurácia Balanceada como métricas complementares e não substitutas, especialmente em cenários com forte desbalanceamento entre classes como o do alvo Balanço (razão aproximada de 7,9:1 entre BN e BR no conjunto de teste).

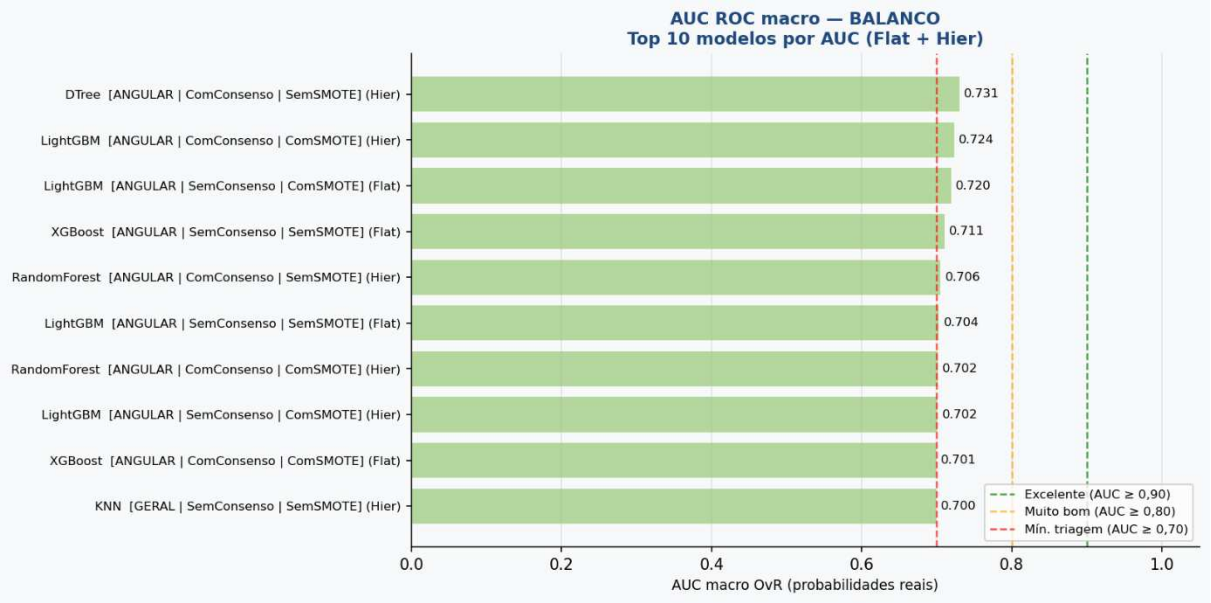

Figura 30 – AUC ROC macro – Balanço
Nota: Barras representam a AUC ROC *macro* (*One-vs-Rest*) no conjunto de teste. A linha tracejada vermelha indica o limite inferior da faixa aceitável (AUC ≥ 0,70); a linha tracejada laranja indica desempenho excelente (AUC ≥ 0,80); a linha tracejada verde indica desempenho excepcional (AUC ≥ 0,90), conforme escala de Hosmer e Lemeshowow. Verde = modelos com AUC ≥ 0,70; Vermelho = modelos abaixo de 0,70. O top 10 apresenta-se concentrado na faixa de triagem aceitável (0,700–0,731), com cinco modelos Hierárquicos e cinco *Flat*, sem alcançar a faixa de desempenho muito bom. O modelo de maior AUC (*DTree* / Angular / "Com Consenso" / "Sem SMOTE" / Hierárquico, AUC = 0,731) coincide com o modelo vencedor por Acurácia Balanceada para este alvo. A proximidade dos valores e a sobreposição dos intervalos de confiança (vide análise *bootstrap*) caracterizam o ranqueamento do top 10 para este alvo. A AUC é independente do limiar de decisão, complementando a Acurácia Balanceada como métrica de avaliação. Ang = Cinemática Angular; Ger = Cinemática Geral; "Sem Consenso" = todos os atributos candidatos são mantidos, sem filtro de consenso; "Com Consenso" = atributos selecionados quando atingem o limiar de votos dos quatro métodos (≥ 2/4 para o alvo Balanço, ≥ 3/4 para Apoio e Transverso); "Com SMOTE" = balanceamento sintético aplicado; "Sem SMOTE" = sem balanceamento; *Flat* = abordagem multiclasse direta; Hier = abordagem hierárquica em cascata.
Fonte: *Elaborada pelo autor.*

Nos 10 melhores do Balanço por Acurácia Balanceada, a Regressão Logística foi o modelo mais frequente (6 posições), seguida por *DTree* e Naive Bayes (2 posições cada). O *Pipeline* Cinemática Angular foi o mais representado (6 de 10 posições), e a variante PCA dominou amplamente (8 de 10), contrastando com os alvos Transverso e Apoio. No Transverso, a variante "Sem Consenso" prevaleceu (7 de 10), enquanto no Apoio o "Com Consenso" dominou (7 de 10). Este padrão alvo-dependente reforça a tese central deste estudo, segundo a qual nenhuma estratégia de redução dimensional é universalmente superior. O melhor modelo do alvo Balanço utilizou redução de dimensionalidade por PCA.

Tabela 23 – Análise Comparativa de Desempenho por Classe: Balanço

Abordagem	BN	BR	Bal.Acc
"Com Consenso" (Hier.)	**89,8%**	50,6%	**0,702**
PCA	79,6%	**55,3%**	0,674
"Sem Consenso"	89,2%	37,6%	0,634

Nota: Células em verde indicam melhor desempenho por coluna. BN = Balanço Normal; BR = Balanço Rígido. "Com Consenso" (Hier.) = melhor resultado por "Com Consenso", obtido na abordagem Hierárquica. Bal.Acc = média das sensibilidades por classe.
Fonte: *Elaborada pelo autor.*

Consolidando os melhores modelos identificados para cada alvo, a Tabela 24 apresenta o desempenho comparativo das quatro variantes de pré-processamento (Com Consenso, Sem Consenso e PCA em abordagem *Flat*, e Com Consenso em abordagem Hierárquica) por alvo.

Tabela 24 – Resumo comparativo das abordagens por alvo

Variante / Abordagem	Apoio	Balanço	Transverso
"Com Consenso" (*Flat*)	**0,717**	0,619	0,758
"Sem Consenso" (*Flat*)	0,673	0,628	0,780
PCA (*Flat*)	0,628	0,674	**0,791**
Hierárquica ("Com Consenso")	0,651	**0,702**	0,594

Nota: Células em verde indicam o melhor resultado por alvo. Os valores representam a Acurácia Balanceada do melhor modelo em cada combinação variante/abordagem no conjunto de teste. Hierárquica = abordagem em cascata com variante "Com Consenso".
Fonte: Elaborada pelo autor.

4.7 Comparação *Flat* vs Hierárquico

A comparação sistemática entre as abordagens *Flat* (classificação multiclasse direta) e Hierárquica (cascata em duas etapas) revelou que a abordagem *Flat* foi superior em dois dos três alvos avaliados, enquanto a abordagem Hierárquica venceu exclusivamente para o alvo Balanço. Os resultados completos são apresentados na Tabela 25 e ilustrados na Figura 31.

Tabela 25 – Comparação entre abordagens *Flat* e Hierárquica por alvo

Alvo	Abord.	*Pipeline*	Variante	SMOTE	Modelo	Bal.Acc	MCC	Kappa	AUC	Venc.
Transverso	***Flat***	**Geral**	**PCA**	**Sem**	***LogReg***	**0,791**	**0,507**	**0,491**	**0,874**	**SIM**
	Hier.	Angular	PCA	Com	*Naive Bayes*	0,618	0,320	0,318	0,688	
Apoio	***Flat***	**Angular**	**"Com Consenso"**	**Com**	***Naive Bayes***	**0,717**	**0,519**	**0,490**	**0,813**	**SIM**
	Hier.	Angular	"Com Consenso"	Com	*Naive Bayes*	0,651	0,444	0,408	0,819	
Balanço	*Flat*	Angular	PCA	Sem	*LogReg*	0,674	0,256	0,230	0,639	
	Hier.	**Angular**	**"Com Consenso"**	**Sem**	***DTree***	**0,702**	**0,359**	**0,354**	**0,731**	**SIM**

Nota: Linhas destacadas em verde correspondem ao modelo vencedor por alvo. Bal.Acc = Acurácia Balanceada; MCC = *Matthews Correlation Coefficient*; AUC = Área sob a curva ROC *(macro OvR)*; Abord. = Abordagem; Hier. = Hierárquica; Venc. = Vencedor do alvo. Angular = Cinemática Angular; Geral = Cinemática Geral.
Fonte: Elaborada pelo autor.

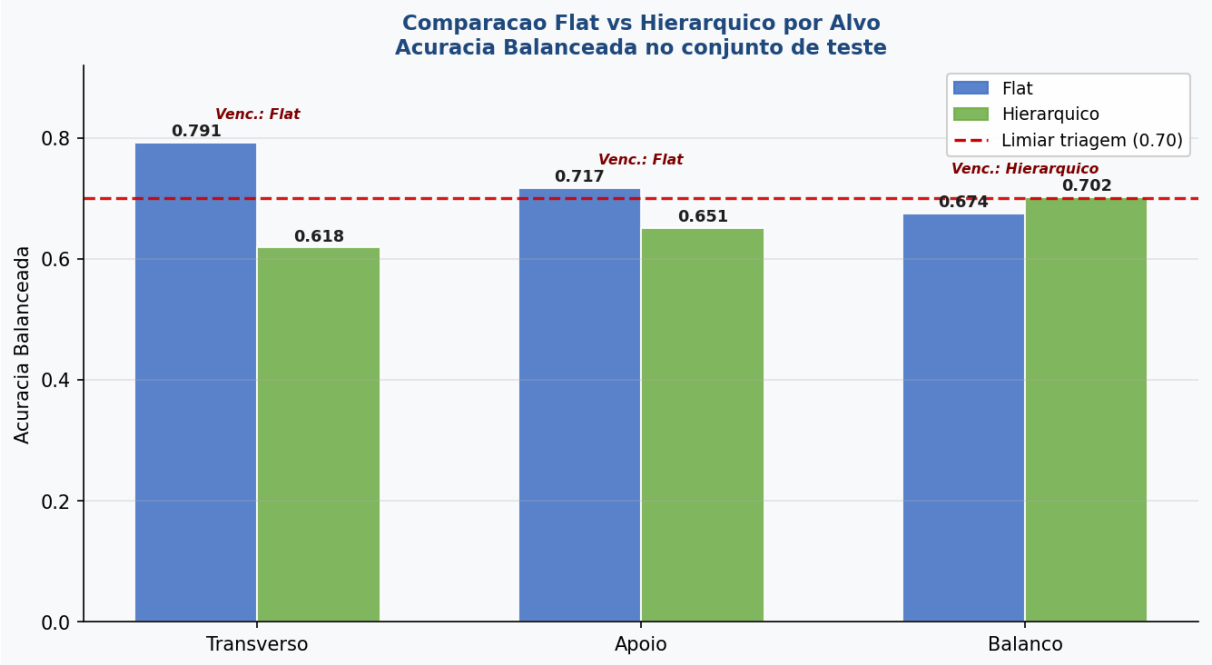

Figura 31 – Acurácia Balanceada: *Flat* vs. Hierárquica por alvo
Nota: Barras azuis representam o melhor modelo *Flat* e barras verdes o melhor modelo Hierárquico por alvo. A linha tracejada vermelha indica o limiar mínimo para triagem clínica (0,70). O rótulo "Venc." indica a abordagem vencedora em cada alvo.
Fonte: Elaborada pelo autor.

Para o alvo Transverso, a abordagem *Flat* foi superior em 17,3 pontos percentuais (0,791 vs 0,618), com o melhor modelo sendo a Regressão Logística com *Pipeline* Cinemática Geral, PCA e "Sem SMOTE". Na cascata Hierárquica, a Etapa 1 apresentou sensibilidade de 75,1% para a classe Alterado, e a sensibilidade efetiva para TE caiu de 100% (*Flat*) para 58,0% (Hier.; *Naive Bayes* / Angular / PCA / "Com SMOTE"), perda de 42,0 pontos percentuais. O *overfitting* do melhor modelo *Flat* foi o menor do estudo (0,036).

Para o alvo Apoio, a abordagem *Flat* foi superior em 6,6 pontos percentuais (0,717 vs 0,651), com o mesmo modelo base, *Naive Bayes* / Angular / "Com Consenso" / "Com SMOTE", vencendo em ambas as abordagens. A Etapa 1 da cascata atingiu sensibilidade de 94,7% para Alterado. Por classe, SA caiu de 77,3% (*Flat*) para 57,2% (Hierarquia: −20,2 pontos percentuais), SS manteve-se estável (56,6% vs 55,9%) e SN também (81,2% vs 82,2%).

Para o alvo Balanço, a abordagem Hierárquica superou a *Flat* em 2,8 pontos percentuais (0,702 vs 0,674). O modelo vencedor *Flat* foi a Regressão Logística (Angular / PCA / "Sem SMOTE") e o vencedor Hierárquico foi a Árvore de Decisão (Angular / "Com Consenso" / "Sem SMOTE"). A análise de estabilidade estocástica conduzida sobre o alvo Balanço (Seção 3.9.4; cinco *seeds*: 0, 1, 7, 42 e 123; 240 execuções) revelou que a configuração vencedora *DTree* apresentou acurácia balanceada média de 0,580 ± 0,111 (mínimo 0,499; máximo 0,702), com distribuição bimodal, duas *seeds* atingiram 0,702 e três caíram para 0,499.

A Regressão Logística manteve-se estável em 0,618 ± 0,000 nas cinco *seeds*. A *seed* 42 reproduziu o valor original de 0,702, confirmando a reprodutibilidade do *pipeline*. Em síntese, a abordagem *Flat* venceu nos dois alvos multiclasse (Transverso e Apoio) e foi superada por margem estreita no alvo binário Balanço.

4.8 Comparação Angular vs Geral

A comparação entre os dois *pipelines* de variáveis avaliados, Cinemática Angular (18 variáveis angulares ipsilaterais) e Cinemática Geral (25 variáveis, acrescidas de 7 parâmetros espaço-temporais ipsilaterais), revelou que o *Pipeline* Cinemática Angular foi superior em dois dos três alvos, enquanto o *Pipeline* Cinemática Geral venceu exclusivamente para o alvo Transverso. Os resultados são apresentados na Tabela 26 e ilustrados na Figura 32.

Tabela 26 – Comparação Cinemática Angular vs. Geral por alvo

Alvo	*Pipeline*	Variante	SMOTE	Abord.	Modelo	Bal.Acc	MCC	Kappa	AUC	Venc.
Transverso	Geral	PCA	Sem	*Flat*	*LogReg*	0,791	0,507	0,491	0,874	SIM
	Angular	"Sem Consenso"	Com	*Flat*	*Voting Soft*	0,756	0,445	0,436	0,831	
Apoio	Angular	"Com Consenso"	Com	*Flat*	*Naive Bayes*	0,717	0,519	0,490	0,813	SIM
	Geral	"Sem Consenso"	Sem	*Flat*	*LogReg*	0,649	0,419	0,408	0,783	
Balanço	Angular	"Com Consenso"	Sem	Hier.	*DTree*	0,702	0,359	0,354	0,731	SIM
	Geral	PCA	Sem	Hier.	*LogReg*	0,667	0,245	0,220	0,632	

Nota: Linhas destacadas em verde correspondem ao *pipeline* vencedor por alvo. Bal.Acc = Acurácia Balanceada; MCC = *Matthews Correlation Coefficient*; AUC = Área sob a curva ROC (*macro OvR*); Abord. = Abordagem; Hier. = Hierárquica; Venc. = Vencedor do alvo.
Fonte: *Elaborada pelo autor.*

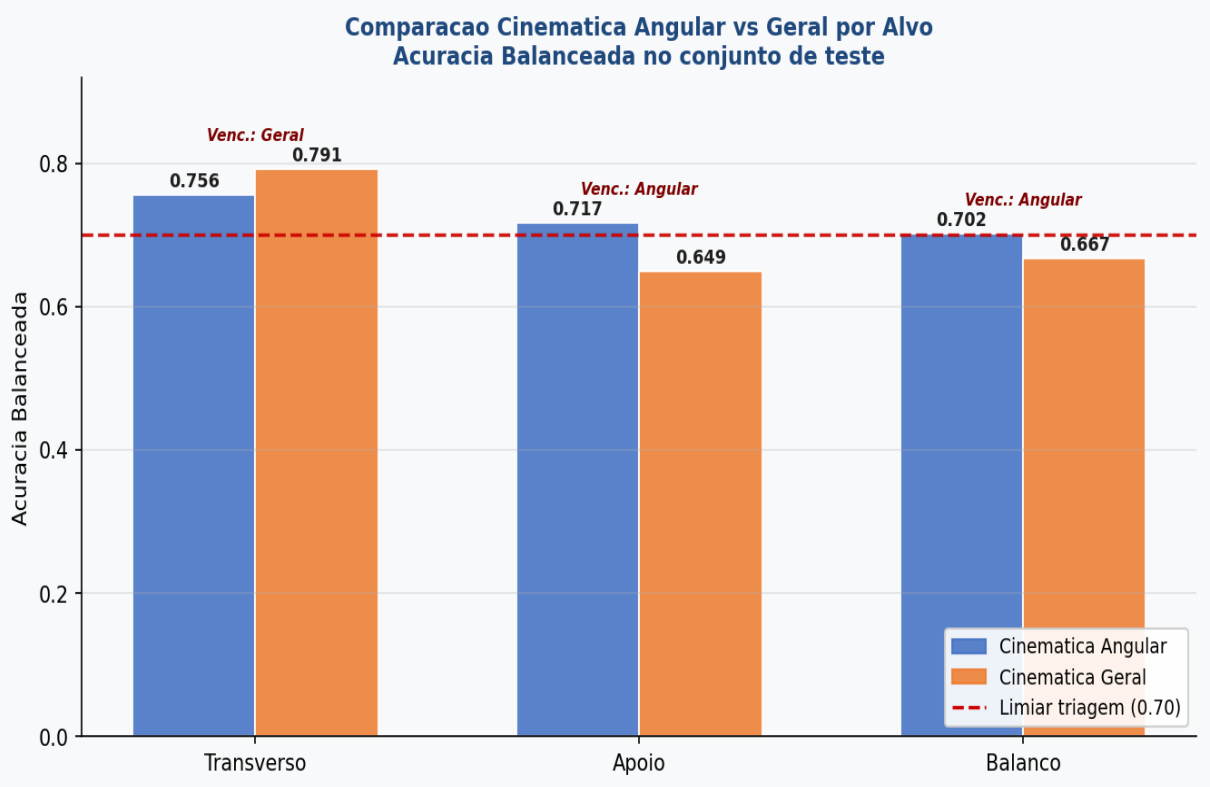

Figura 32 – Acurácia Balanceada: Cinemática Angular vs. Geral por alvo
Nota: Barras azuis representam o melhor modelo do *Pipeline* Cinemática Angular e barras laranja o melhor modelo do Pipeline Cinemática Geral por alvo. A linha tracejada vermelha indica o limiar mínimo para triagem clínica (0,70). O rótulo "Venc." indica o pipeline vencedor em cada alvo.
Fonte: *Elaborada pelo autor.*

Para o alvo Transverso, o *Pipeline* Cinemática Geral foi superior em 3,5 pontos percentuais (0,791 vs 0,756), com o melhor modelo sendo a Regressão Logística com PCA e "Sem SMOTE".

Para o alvo Apoio, o *Pipeline* Cinemática Angular foi superior em 6,8 pontos percentuais (0,717 vs 0,649), com o melhor modelo sendo o *Naive Bayes* com "Com Consenso" e "Com SMOTE". Para o alvo Balanço, o *Pipeline* Cinemática Angular também foi superior em 3,5 pontos percentuais (0,702 vs 0,667). Em síntese, o *Pipeline* Cinemática Angular foi superior nos alvos Apoio e Balanço, e o *Pipeline* Cinemática Geral foi superior apenas no alvo Transverso.

4.9 Comparação “Com SMOTE” vs “Sem SMOTE”

Para avaliar se o balanceamento sintético da classe minoritária via SMOTE produz ganhos sistemáticos de desempenho, comparou-se a frequência de cada variante (Com SMOTE e Sem SMOTE) entre os cinco melhores modelos de cada alvo, bem como o desempenho dos melhores modelos de cada variante. A Tabela 27 sintetiza o resultado, separado por alvo.

Tabela 27 – Comparação entre variantes "Com SMOTE" e "Sem SMOTE" por alvo

Alvo	Vencedor	Bal.Acc	Δ vs. variante alternativa	5 melhores (Sem/Com)
Transverso	**"Sem SMOTE"**	0,791	0,000 (empate; "Sem SMOTE" em 1° por AUC superior: 0,874 vs 0,870)	3 / 2
Apoio	**"Com SMOTE"**	0,717	+0,053 vs. "Sem SMOTE" (0,664)	2 / 3
Balanço	**"Sem SMOTE"**	0,702	≥ +0,035 vs. "Com SMOTE"[1]	5 / 0

Nota: Bal.Acc = Acurácia Balanceada do melhor modelo no conjunto de teste para a variante vencedora (referência: Tabelas 18 a 20). Δ = diferença em pontos absolutos de Bal.Acc entre o melhor modelo de cada variante; valor positivo indica vantagem da variante destacada na coluna 'Vencedor'. Distribuição nos 5 melhores = número de modelos "Sem SMOTE" / "Com SMOTE" entre os cinco melhores por Bal.Acc no conjunto de teste. Sombreamento verde indica vantagem clara de uma variante; sombreamento amarelo indica empate em Bal.Acc com desempate por AUC. [1] "Com SMOTE" não figurou no Top 5 do alvo Balanço; o valor de Δ reportado representa a diferença mínima verificável (Bal.Acc do vencedor "Sem SMOTE" = 0,702; menor Bal.Acc do Top 5 = 0,667).
Fonte: Elaborada pelo autor.

4.10 Análise Clínica: Minimização de Falsos Negativos (FN)

Esta seção apresenta os resultados das três estratégias de otimização clínica (E1, E2 e E3) definidas na Seção 3.11, aplicadas às cinco classes patológicas do estudo (SS, SA, BR, TI e TE), comparadas com o desempenho do modelo padrão (ponto de corte = 0,50). Os resultados são sumarizados na Tabela 28.

Tabela 28 – Comparação de Falsos Negativos por estratégia de otimização clínica

Classe	Nome Completo	Alvo	FN Padrão	FN E1	FN E2	FN E3	Melhor Estratégia	Redução FN
SS	Sagital Salto	Apoio	82	51	42	33	**E3**	**-49**
SA	Sagital Agachado	Apoio	81	26	11	0	**E3**	**-81**
BR	Balanço Rígido	Balanço	70	38	21	66	**E2**	**-49**
TI	Transverso interno	Transverso	97	64	26	79	**E2**	**-71**
TE	Transverso externo	Transverso	33	0	4	14	**E1**	**-33**

Nota: FN = Falsos Negativos (crianças com padrão patológico classificadas incorretamente como normais). FN Padrão = RandomForest com ponderação clínica (RF_PesoClinico), avaliado no ponto de corte 0,50, por classe no pipeline indicado — *mesmo* baseline da Tabela 29. E1 = melhor *pipeline* que maximiza a sensibilidade mínima patológica, sem ajuste de ponto de corte; E2 = ajuste de ponto de corte para atingir sensibilidade ≥ 90%; E3 = Floresta Aleatória retreinado com Optuna otimizando sensibilidade patológica mínima. "Melhor" = estratégia de menor FN; "Redução FN" = FN(padrão) − FN(melhor). SS = Sagital Salto; SA = Sagital Agachado; BR = Balanço Rígido; TI = Transverso Interno; TE = Transverso Externo.

Fonte: Elaborada pelo autor.

A Estratégia E1 (Seleção por sensibilidade mínima) produziu ganhos em todas as cinco classes patológicas: eliminou completamente os falsos negativos do TE (de 33 para zero), e reduziu os de SA em 55 casos (de 81 para 26), TI em 33 (de 97 para 64), BR em 32 (de 70 para 38) e SS em 31 (de 82 para 51). Para TE, a E1 constitui a estratégia de melhor resultado entre todas as avaliadas (Tabela 28), dispensando ajuste de ponto de corte ou retreinamento. Para as demais quatro classes (SS, SA, BR e TI), a E1 foi superada pelas Estratégias E2 ou E3, que produziram ganhos adicionais clinicamente substantivos e são detalhadas a seguir.

As curvas de otimização do ponto de corte (Estratégia E2) para cada classe patológica são apresentadas nas Figuras 33 a 37. Cada figura exibe dois painéis: à esquerda, a evolução da sensibilidade e especificidade em função do ponto de corte; à direita, o número de falsos negativos resultante para cada valor de ponto de corte. A linha verde tracejada indica o ponto de corte ótimo selecionado e a linha vermelha tracejada indica o alvo de 90% de sensibilidade.

A Estratégia E2 apresentou desempenho variável conforme a classe. Para TE (Transverso Externo, Figura 33), o ajuste do ponto de corte de 0,50 para 0,18 reduziu os falsos

negativos de 33 para 4 (redução de 88%), resultado inferior ao da E1 (0 FN), mas incluído aqui para completude da análise comparativa entre as três estratégias. Para SS (Figura 34), o ponto de corte de 0,06 reduziu os falsos negativos de 82 para 42 com sensibilidade de 90,1%, mas com colapso da especificidade (43,3% → 0,0%). Para SA (Figura 35), o ponto de corte de 0,02 reduziu os falsos negativos de 81 para 11 com sensibilidade de 91,4% e especificidade de 90,1% → 70,3%. Para TI (Transverso interno, Figura 36) e BR (Balanço Rígido, Figura 37), os pontos de corte ótimos caíram para o limite inferior do espaço de busca (0,01). Em TI, a sensibilidade subiu de 50,0% para 86,6% e a especificidade caiu de 97,2% para 12,5% (perda de 84,6 pontos percentuais); em termos absolutos, dos 566 casos sem TI no conjunto de teste, 71 permaneceram classificados como não-TI e 495 foram classificados como TI (87,5%). Em BR, a sensibilidade subiu de 17,6% para 75,3% e a especificidade caiu de 94,7% para 27,6% (perda de 67,1 pontos percentuais); dos 675 casos de Balanço Normal, 489 foram classificados como Balanço Rígido (72,4%). A Tabela 29 sintetiza o compromisso sensibilidade × especificidade para as cinco classes patológicas avaliadas.

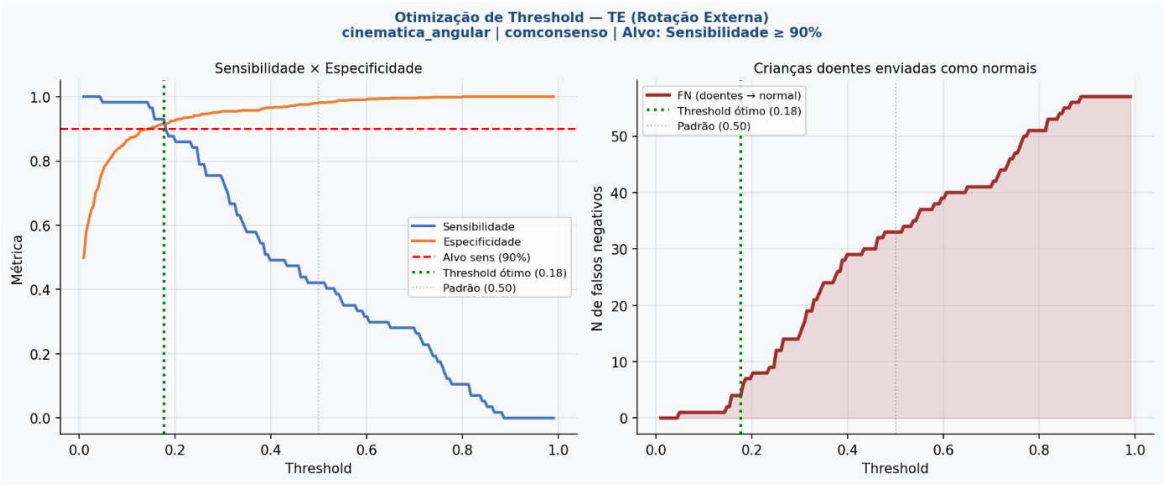

Figura 33 – Otimização de ponto de corte: classe TE (Transverso Externo).
Nota: Ponto de corte ótimo = 0,18; Sensibilidade atingida = 93,0%; FN reduzido de 33 para 4. *Pipeline*: Cinemática Angular, "Com Consenso".
Fonte: Elaborada pelo autor.

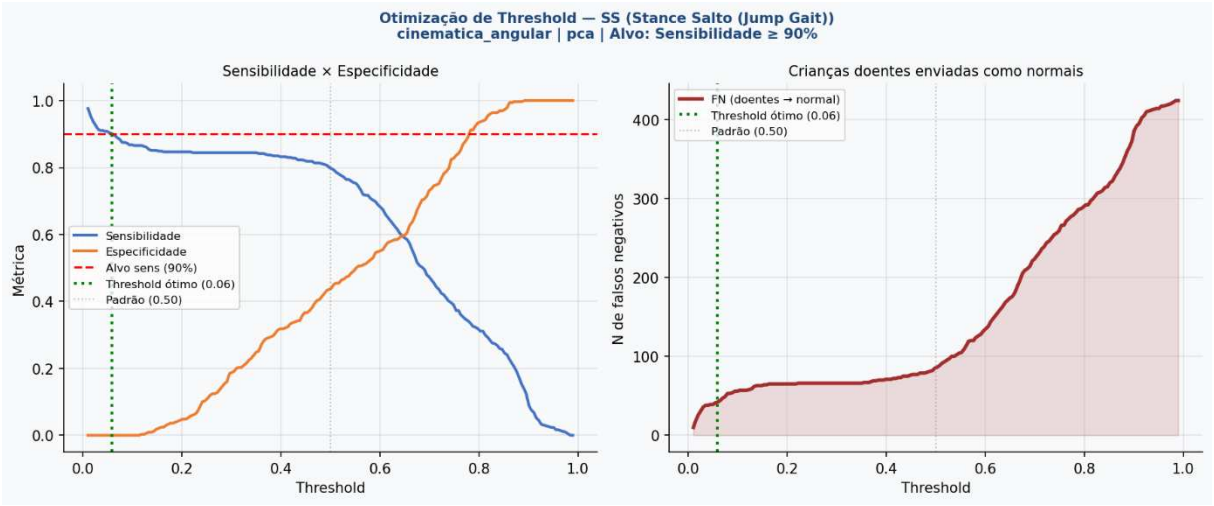

Figura 34 – Otimização de ponto de corte: classe SS (Sagital Salto)
Nota: Ponto de corte ótimo = 0,06; Sensibilidade atingida = 90,1%; FN reduzido de 82 para 42. *Pipeline*: Cinemática Angular, PCA.
Fonte: Elaborada pelo autor.

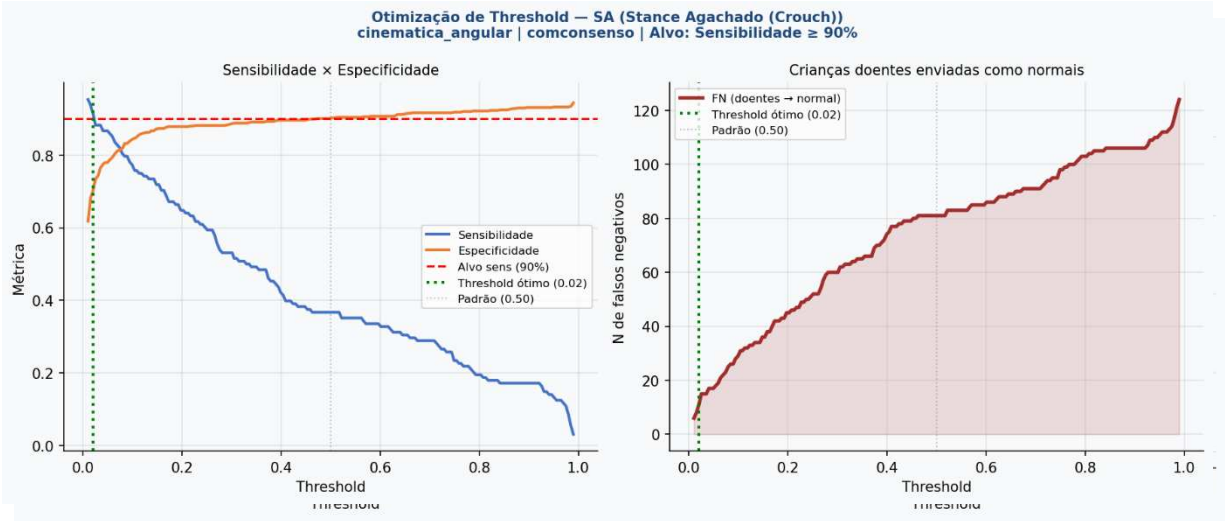

Figura 35 – Otimização de ponto de corte: classe SA (Sagital Agachado)
Nota: Ponto de corte ótimo = 0,02; Sensibilidade atingida = 91,4%; FN reduzido de 81 para 11. *Pipeline*: Cinemática Angular, "Com Consenso". Nota: a Estratégia E3 obteve resultado superior (FN = 0).
Fonte: Elaborada pelo autor.

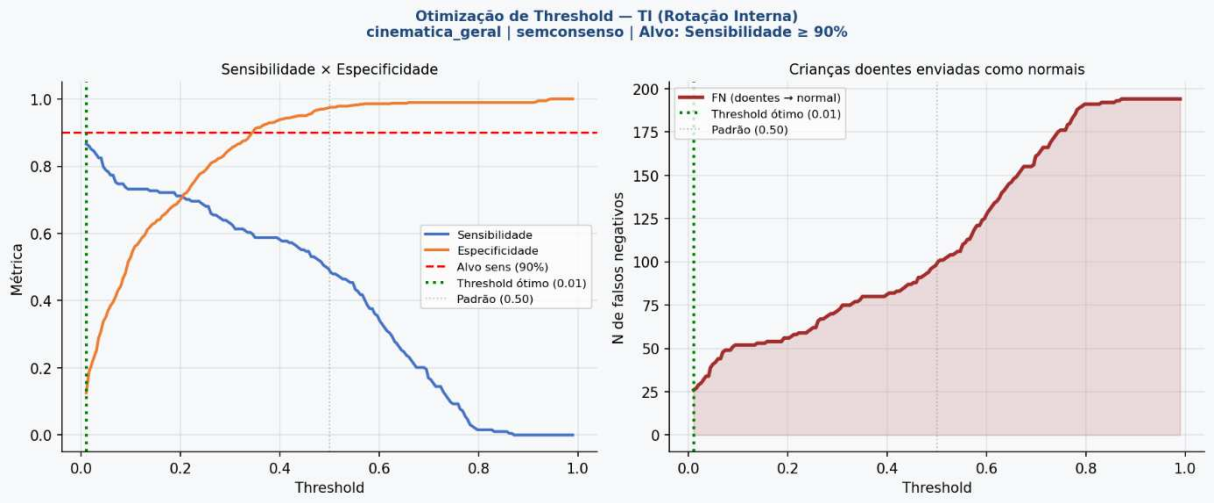

Figura 36 – Otimização de ponto de corte: classe TI (Transverso interno)
Nota: Ponto de corte ótimo = 0,01 (limite inferior do espaço de busca); Sensibilidade máxima atingida = 86,6%, alvo de 90% não foi alcançado; FN reduzido de 97 para 26; especificidade reduzida de 97,2% para 12,5% (perda de 84,6 pontos percentuais). *Pipeline*: Cinemática Geral, "Sem Consenso".
Fonte: Elaborada pelo autor.

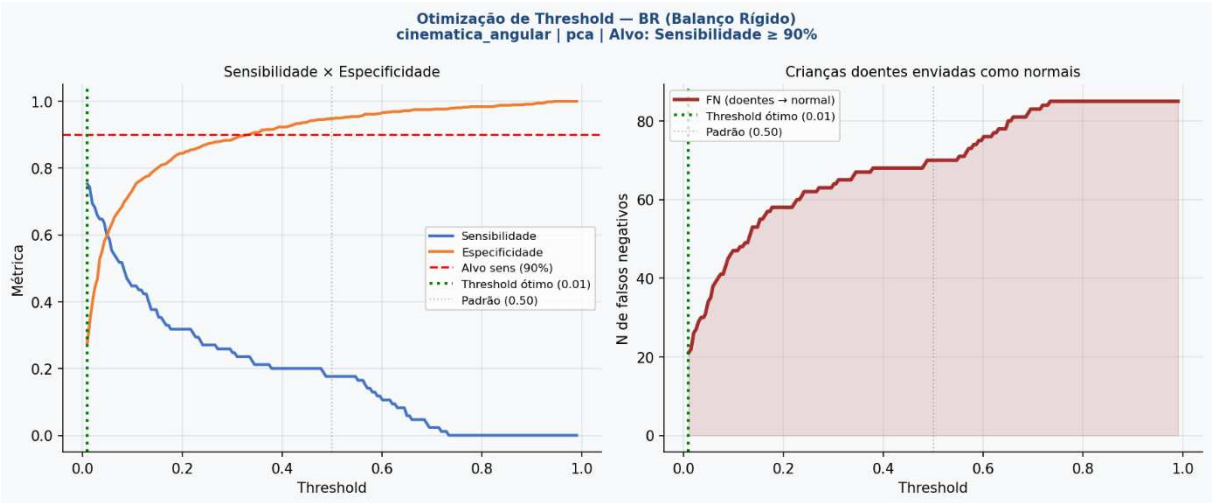

Figura 37 – Otimização de ponto de corte: classe BR (Balanço Rígido)
Nota: Ponto de corte ótimo = 0,01 (limite inferior do espaço de busca); Sensibilidade máxima atingida = 75,3% - alvo de 90% não foi alcançado; FN reduzido de 70 para 21; especificidade caiu de 94,7% para 27,6% (perda de 67,1 pontos percentuais), configurando colapso de especificidade com inviável taxa de falsos positivos para uso clínico. *Pipeline*: Cinemática Angular, PCA.
Fonte: Elaborada pelo autor.

Tabela 29 – Compromisso sensibilidade × especificidade (Estratégia E2) por classe

Classe	Ponto de corte ótimo	Sens (padrão → otimizado)	Spec (padrão → otimizado)	ΔSpec	Viável para triagem?
TE	0,18	42,1% → 93,0%	98,0% → 91,5%	**−6,5 pp**	**Sim**
SA	0,02	36,7% → 91,4%	90,1% → 70,3%	−19,8 pp	Marginal
SS	0,06	80,7% → 90,1%	**43,3% → 0,0%**	−43,3 pp	**Não**
BR	0,01	17,6% → 75,3%	**94,7% → 27,6%**	**−67,1 pp**	**Não**
TI	0,01	50,0% → 86,6%	**97,2% → 12,5%**	**−84,6 pp**	**Não**

Nota: Sens = sensibilidade; Spec = especificidade; ΔSpec = variação em pontos percentuais (pp) entre o ponto de corte padrão (0,50) e o ponto de corte otimizado. Pontos de corte iguais a 0,01 correspondem ao limite inferior do espaço de busca e configuram colapso de especificidade, no qual o modelo classifica praticamente todos os casos como patológicos. A coluna "Viável para triagem?" considera se compromisso entre sensibilidade obtida e especificidade remanescente é clinicamente aceitável para uso como ferramenta de rastreio; "Marginal" indica compromisso aceitável apenas sob supervisão especializada imediata. Entre as estratégias de ajuste de ponto de corte avaliadas, a classe TE apresentou o único compromisso plenamente compatível com triagem clínica (embora a Estratégia E1 produza resultado ainda superior para essa classe, conforme Tabela 28); SS e SA foram mais bem endereçadas pela Estratégia E3 (retreinamento); BR e TI permanecem limitadas pela separabilidade intrínseca no espaço cinemático, conforme discutido nas Seções 5.5.2 e 5.9. As cores de fundo das linhas indicam a viabilidade clínica do compromisso (verde = viável, amarelo = marginal, laranja = inviável).

Fonte: *Elaborada pelo autor.*

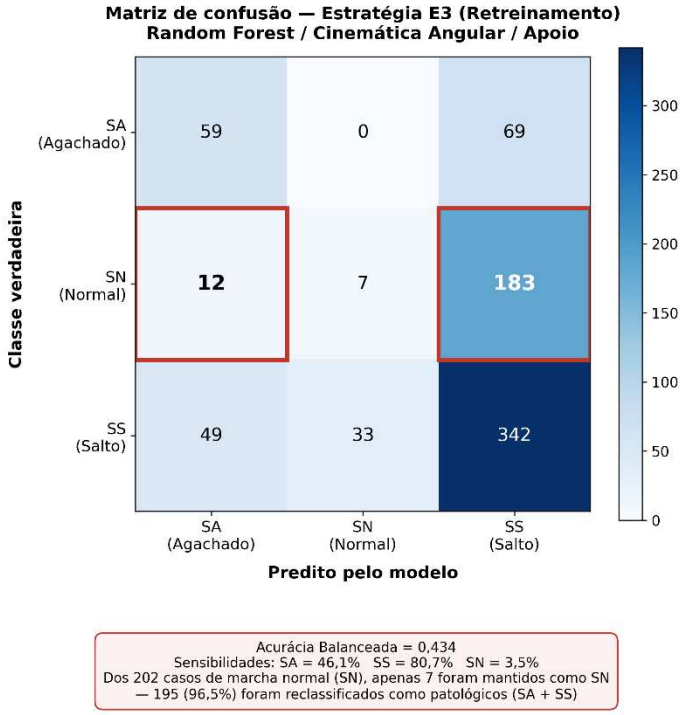

Figura 38 – Matriz de confusão da Estratégia E3 (Retreinamento), alvo Apoio
Nota: Valores representam o número de casos do conjunto de teste (n = 754). Linhas indicam classe verdadeira; colunas, classe predita. SA = Sagital Agachado; SN = Sagital Normal; SS = Sagital Salto. As células destacadas em vermelho correspondem a casos de marcha normal (SN) incorretamente classificados como patológicos, evidência direta do colapso de especificidade obtido pela estratégia E3 para este alvo: embora a meta de minimizar falsos negativos para SA tenha sido atingida (sensibilidade de 46,1% e zero casos SA → SN), o modelo reclassificou 195 dos 202 casos normais (96,5%) como patológicos, resultando em Acurácia Balanceada de apenas 0,434.
***Fonte**: Elaborada pelo autor.*

Para as classes do alvo Apoio, a Estratégia E3 (Retreinamento) obteve a maior redução de falsos negativos. Para SS (Sagital Salto), os falsos negativos reduziram de 82 para 33; para SA (Sagital Agachado), de 81 para zero (nenhum caso classificado como Normal), embora a sensibilidade efetiva específica para SA tenha sido de 46,1%, com os demais casos classificados como SS. A matriz de confusão correspondente é apresentada na Figura 38.

5 DISCUSSÃO

5.1 Síntese dos Principais Achados

Este estudo desenvolveu e avaliou um *pipeline* de aprendizado de máquina para classificação automática de padrões da marcha em crianças e adolescentes com paralisia cerebral espástica, com base no Sistema GCS-CP (9). Os melhores modelos identificados para cada alvo são apresentados na Tabela 24, junto com o desempenho das demais variantes; nas Seções 5.2 e 5.3, esses resultados são interpretados em profundidade. Esta seção sintetiza os achados centrais e avalia as hipóteses iniciais.

Um achado metodológico transversal aos três alvos foi a constatação de que a Acurácia Balanceada, embora adequada para seleção de modelos sob desbalanceamento, mascara limitações clinicamente relevantes quando interpretada isoladamente. Esta observação motivou a análise por classe e fundamenta a avaliação clínica detalhada conduzida na Seção 4.10.

À luz dos resultados consolidados (Tabela 24, Seção 4.6), as seis hipóteses iniciais podem ser avaliadas:

A Hipótese 1 (H1), que previa Acurácia Balanceada superior a 80% para os algoritmos avaliados, foi parcialmente refutada. Nenhum alvo atingiu o limiar agregado, embora sensibilidades por classe superiores a 80% tenham sido alcançadas em padrões específicos, indicando sustentação da hipótese em nível de classe, mas não como propriedade universal do *pipeline*.

A Hipótese 2 (H2), de superioridade da seleção de variáveis sobre o PCA, foi parcialmente confirmada. A seleção por consenso predominou em frequência, mas com vantagem expressiva apenas no alvo Apoio; nos demais, o PCA produziu os melhores modelos absolutos. A relação entre as duas estratégias mostrou-se alvo-dependente, contrariando expectativas de generalização a priori e estabelecendo achado metodológico interpretado em detalhe na Seção 5.3.

A Hipótese 3 (H3), de superioridade dos modelos de fusão, foi refutada. As fusões não superaram os melhores modelos individuais em nenhum dos três alvos avaliados, embora tenham se mostrado competitivas no alvo Transverso. O comportamento é compatível com a hipótese teórica de que fusões compensam fragilidades quando os classificadores-base

apresentam erros não-correlacionados, ganho que se torna menos relevante quando um classificador-base é claramente superior. Esta interpretação é aprofundada na Seção 5.6.

A Hipótese 4 (H4), de heterogeneidade de desempenho por classe, foi plenamente confirmada e constitui o achado metodológico central deste trabalho. As sensibilidades por classe variaram de patamares clinicamente inviáveis a 100%, validando a premissa de que métricas agregadas são insuficientes para avaliação de aplicabilidade clínica e justificando a abordagem de minimização de falsos negativos por classe (Seção 4.10). Os mecanismos biomecânicos e algorítmicos por trás dessa heterogeneidade são examinados nas Seções 5.4 e 5.5.

A Hipótese 5 (H5), de que estratégias de otimização clínica reduziriam substancialmente os falsos negativos, foi confirmada em termos de redução, mas qualificada por compromissos heterogêneos entre sensibilidade e especificidade. O resultado divide as classes patológicas em três categorias de viabilidade clínica: plenamente viável para TE, aceita com ressalvas para SA e SS, e inviável sem comprometimento inaceitável para BR e TI. Esses três regimes e seus mecanismos são detalhados na Seção 5.10 e sintetizados na Tabela 30.

A Hipótese 6 (H6), de que a classificação hierárquica em cascata apresentaria desempenho clínico por classe distinto da classificação multiclasse direta (*Flat*), especialmente nas classes patológicas minoritárias, foi refutada. A abordagem *Flat* mostrou-se superior nos dois alvos multiclasse (Transverso e Apoio), enquanto no alvo binário Balanço as duas abordagens são matematicamente equivalentes. A única vantagem aparente da cascata decorreu de variabilidade estocástica na otimização, e não de superioridade real, de modo que a estrutura em cascata não acrescentou poder discriminativo às classes minoritárias. Esta comparação é detalhada na Seção 5.2.

A caracterização alvo-dependente do compromisso entre seleção de variáveis e PCA, a relevância biomecânica das variáveis selecionadas e a análise das estratégias de otimização clínica constituem contribuições metodológicas centrais deste trabalho, e são discutidas em profundidade nas seções subsequentes. Estas contribuições devem ser interpretadas à luz das limitações por classe aqui identificadas.

5.2 Abordagem hierárquica versus abordagem multiclasse direta

A análise comparativa entre as abordagens *Flat* (classificação multiclasse direta) e Hierárquica (cascata em duas etapas) revelou uma importante relação de compromisso

entre a decomposição do problema em subproblemas mais simples e a propagação de erros entre as etapas. Embora revisões da literatura apontem que abordagens hierárquicas locais tendem a superar a classificação *Flat* quando há estrutura hierárquica natural entre as classes (34), os resultados do presente estudo indicam que essa vantagem não é universal e depende criticamente da estrutura de cada alvo.

Para o alvo Transverso, a ampla superioridade da abordagem *Flat* (Tabela 25, Seção 4.7) reflete a propagação de erros da cascata. Na Etapa 1 (triagem Normal vs Alterado), casos de TE classificados incorretamente como normais são eliminados antes de chegarem à subtipagem na Etapa 2, tornando sua recuperação impossível. O mesmo mecanismo penaliza sistematicamente a classe TI, confirmando que a estrutura em cascata afeta de forma desproporcional as classes patológicas deste alvo.

Para o alvo Apoio, o fato de o mesmo modelo base, *Naive Bayes* / Angular / "Com Consenso" / "Com SMOTE", vencer nas duas abordagens permite isolar o efeito da arquitetura. A queda observada concentrou-se na classe SA, enquanto SS e SN permaneceram praticamente estáveis (Tabela 25, Seção 4.7). Este padrão indica que a Etapa 1 da cascata classificou corretamente a maioria dos casos normais, mas introduziu erros adicionais na subtipagem entre SA e SS na Etapa 2. A propagação de erros não é uniforme entre as classes: penaliza seletivamente subtipos específicos.

A formulação multiplicativa da sensibilidade efetiva apresentada na Seção 3.9.3 pressupõe independência entre os erros das duas etapas hierárquicas. Na prática, é provável que exista correlação positiva entre esses erros: os casos mais ambíguos na triagem (os que a Etapa 1 tende a errar) costumam ser também os clinicamente limítrofes entre subtipos (os que a Etapa 2 tende a errar).

Sob essa dependência positiva, a sensibilidade efetiva real pode ser ligeiramente superior ao produto multiplicativo, uma vez que os casos que passam pela Etapa 1 são, em média, os mais fáceis e tendem a ser mais bem classificados pela Etapa 2, e os valores reportados devem ser interpretados como limites inferiores conservadores do desempenho real. Ainda assim, esse desempenho real permanece limitado pelo teto imposto pela sensibilidade da Etapa 1 e bem abaixo da abordagem *Flat* nos alvos multiclasse. O tratamento formal dessa correlação exigiria modelar conjuntamente as duas etapas, por exemplo via matriz de confusão aninhada de 3×3 condicionais, o que constitui refinamento metodológico para trabalhos futuros.

A formulação da propagação de erros em sistemas de classificação hierárquica evidência que o desempenho global é fortemente condicionado pela acurácia das decisões nos níveis iniciais, uma vez que erros ocorridos nas etapas superiores não podem ser

recuperados nas etapas subsequentes (34). Portanto, a abordagem hierárquica somente tende a ser vantajosa quando a Etapa 1 apresenta sensibilidade muito elevada para a detecção de casos alterados.

Para o alvo Balanço, a abordagem Hierárquica obteve resultado superior em 2,8 pontos percentuais (0,702 vs 0,674). Este resultado merece interpretação cuidadosa, pois pode induzir a uma conclusão incorreta de que a cascata foi vantajosa para este alvo.

O primeiro ponto a compreender é que o alvo Balanço é um problema binário: existem apenas duas classes, Balanço Normal (BN) e Balanço Rígido (BR). Em uma abordagem hierárquica com dois níveis, a Etapa 1 classifica os casos em "Normal vs Alterado" e a Etapa 2 faz a subtipagem entre os padrões alterados. Para o Balanço, entretanto, não existe Etapa 2, pois há apenas um padrão alterado (BR). Isso significa que o modelo Hierárquico e o modelo *Flat* resolvem exatamente o mesmo problema, com as mesmas classes e os mesmos dados. A estrutura em cascata não acrescenta nenhuma informação adicional.

O segundo ponto é entender por que os resultados diferem. O modelo vencedor *Flat* é uma Regressão Logística, enquanto o vencedor Hierárquico é uma Árvore de Decisão, dois algoritmos completamente distintos. A Árvore de Decisão é um algoritmo notoriamente instável, conforme demonstrado por Breiman (74), que introduziu o *Bagging*, combinação de múltiplas árvores treinadas em diferentes reamostragens dos dados, justamente como método de redução de variância para classificadores instáveis como árvores. Pequenas variações nos hiperparâmetros *max_depth, criterion e min_samples_leaf* podem produzir modelos com desempenhos muito diferentes. O processo de otimização via Optuna realiza 50 tentativas independentes com inicializações aleatórias distintas para cada abordagem. Uma analogia ilustra esse mecanismo: imagine dois grupos de pesquisadores testando combinações de ingredientes para uma receita, com 50 tentativas cada; é provável que um dos grupos, por acaso, encontre uma combinação ligeiramente melhor, não porque seu método seja superior, mas porque teve sorte nas tentativas exploradas. Foi exatamente o que ocorreu: o Optuna da abordagem Hierárquica encontrou por acaso uma configuração favorável da Árvore de Decisão que o Optuna da abordagem *Flat* não localizou.

A evidência desse argumento está nos algoritmos estáveis. A Regressão Logística obteve resultados idênticos nas duas abordagens (*Flat* = 0,674; Hierárquico = 0,674), com sensibilidades também idênticas para BN (79,6%) e BR (55,3%). Isso ocorre porque algoritmos estáveis e bem-comportados produzem soluções consistentes independentemente da inicialização do Optuna. A diferença observada entre os modelos vencedores absolutos das duas abordagens (*LogReg Flat* × *DTree* Hierárquico) reflete, portanto, a variabilidade do processo

de otimização estocástica, e não uma superioridade real da abordagem em cascata para este alvo. Ao comparar o vencedor *Flat* (*LogReg*, Sens_BN = 79,6%, Sens_BR = 55,3%) com o vencedor Hierárquico (*DTree*, Sens_BN = 89,8%, Sens_BR = 50,6%), nota-se que o ganho do *DTree* na Acurácia Balanceada foi obtido por melhoria em BN (+10,2 pontos percentuais) às custas de queda em BR (-4,7 pontos percentuais), exatamente o padrão esperado de uma Árvore de Decisão que encontrou uma partição favorável à classe majoritária.

A análise complementar de estabilidade estocástica conduzida sobre o alvo Balanço (Seção 3.9.4; números completos reportados na Seção 4.7) sustenta empiricamente essa interpretação. A configuração vencedora (*DTree* / Angular / "Sem SMOTE") apresentou distribuição bimodal entre *seeds*, enquanto a Regressão Logística na mesma configuração manteve desempenho perfeitamente estável, padrão consistente com a sensibilidade conhecida de Árvores de Decisão a variações de hiperparâmetros e com a robustez de modelos lineares regularizados. O valor 0,702 é reprodutível apenas em uma fração específica das condições estocásticas e não corresponde à expectativa central de desempenho da abordagem Hierárquica neste alvo. Para implementação clínica, algoritmos estáveis como a Regressão Logística oferecem maior previsibilidade de desempenho. A análise exaustiva com dezenas de *seeds*, seguindo as recomendações de rigor estatístico de Demšar (75), constitui uma direção para estudos futuros.

Em síntese, a abordagem *Flat* emergiu como a estratégia mais robusta para os alvos multiclasse (Transverso e Apoio), com base nas evidências empíricas reunidas no presente estudo; para o alvo Balanço, a aparente superioridade da abordagem Hierárquica deve ser interpretada com cautela, dado o contexto metodológico descrito anteriormente.

5.3 Seleção de Atributos, PCA e Pipelines de Variáveis

A comparação sistemática entre a seleção de variáveis por consenso e a Análise de Componentes Principais (PCA) revelou um padrão alvo-dependente, sem superioridade universal. O PCA produziu os melhores modelos absolutos em Transverso e Balanço, embora com margem estreita sobre a variante "Sem Consenso" no Transverso, enquanto o "Com Consenso" superou amplamente o PCA no Apoio. Cada estratégia prevalece em alvos com características biomecânicas distintas, padrão discutido em detalhe a seguir.

Para o alvo Transverso, o PCA produziu o melhor modelo absoluto do estudo (*LogReg* / Geral / PCA / "Sem SMOTE", Acurácia Balanceada = 0,791), seguido de perto pela

variante "Sem Consenso" (0,780, diferença de apenas 1,1 pontos percentuais) e mais distante pela variante "Com Consenso" (0,758, diferença de 3,4 pontos percentuais). A proximidade entre PCA e "Sem Consenso", ambas dentro da largura típica dos intervalos de confiança deste estudo, indica que, neste alvo, a redução de dimensionalidade por componentes principais e o uso integral do conjunto de variáveis produzem desempenho equivalente. A diferença real está no consenso, que perdeu posição em relação a ambos. Esse padrão sugere que a seleção supervisionada descarta informação relevante para a discriminação dos padrões transversos. Como esses padrões são definidos primariamente por rotações articulares do quadril e tornozelo, com variâncias bem distribuídas pelo espaço de atributos, tanto a captura por componentes principais quanto a manutenção de todas as variáveis preservam o sinal discriminativo, enquanto o filtro do consenso o degrada. O achado é consistente com a observação de Prakash et al. (23) de que o PCA é adequado quando as relações nos dados são predominantemente lineares, condição que parece satisfeita neste alvo e que igualmente favorece a variante "Sem Consenso".

Para o alvo Balanço, o PCA foi superior à seleção em abordagem equivalente: PCA *Flat* = 0,674 vs "Com Consenso" *Flat* = 0,619, diferença de +5,5 pontos percentuais. O melhor resultado absoluto para Balanço (*DTree* / "Com Consenso" / Hierárquico, Acurácia Balanceada = 0,702) superou o PCA, mas, como discutido na Seção 5.2, essa diferença reflete variabilidade da otimização estocástica e não vantagem estrutural da seleção. A natureza do PCA, combinações lineares que sucessivamente maximizam a variância (76), é adequada quando as componentes principais tendem a alinhar-se com as direções de máxima separabilidade entre classes, condição mais provável quando o número de classes é pequeno e a estrutura covariante das variáveis preditoras é dominada por poucos eixos. No presente caso, a fronteira de decisão entre marcha normal e marcha rígida pode ser adequadamente representada por uma combinação linear das variáveis angulares do joelho. Este resultado é consistente com a literatura, que aponta limitações do PCA em dados não lineares e não gaussianos e sugere maior adequação do método a estruturas predominantemente lineares (26). A vantagem do PCA no alvo Balanço é consistente: além da diferença de +5,5 pontos percentuais entre os melhores modelos, observa-se superioridade do PCA também em configurações pareadas por modelo, *pipeline* e SMOTE.

Para o alvo Apoio, a seleção por consenso foi claramente superior. A variante "Com Consenso" atingiu Acurácia Balanceada = 0,717 contra 0,628 do PCA, diferença de +8,9 pontos percentuais. A distinção entre os padrões SS (Sagital Salto) e SA (Sagital Agachado) depende de relações não-lineares sutis entre ângulos articulares do tornozelo e joelho, que o

PCA dilui ao combinar variáveis em componentes lineares. A seleção supervisionada identificou as 7 variáveis angulares mais relevantes para este alvo, concentrando o poder discriminativo do modelo nas características biomecânicas clinicamente mais informativas. Este resultado é consistente com Prakash et al. (23), que apontam que técnicas lineares como o PCA tendem a ser inadequadas quando o problema é não-linear ou complexo.

Em síntese, a escolha entre seleção de variáveis e PCA deve ser guiada empiricamente para cada problema específico. O PCA é adequado quando as direções de maior variância coincidem com as direções discriminativas, condição satisfeita para Transverso e Balanço. A seleção supervisionada é superior quando as relações discriminativas são não-lineares e distribuídas entre múltiplas variáveis, condição presente no alvo Apoio.

Padrão semelhante de dependência do alvo emerge na comparação entre os *pipelines* de variáveis. O *Pipeline* Cinemática Angular venceu nos alvos definidos no plano sagital, Apoio e Balanço, enquanto o *Pipeline* Cinemática Geral mostrou vantagem apenas no alvo Transverso (Tabela 26, Seção 4.8). A distinção entre marcha normal e rígida no Balanço é determinada pela amplitude de flexão do joelho na fase de balanço, variável diretamente presente no *Pipeline* Angular, e a inclusão dos parâmetros espaço-temporais no *Pipeline* Geral não agregou informação discriminativa relevante para esse alvo binário. No Transverso, por outro lado, os parâmetros espaço-temporais parecem capturar informação complementar associada ao padrão rotacional da marcha, o que explica a vantagem do *Pipeline* Geral nesse plano de movimento. Esses achados reforçam que a escolha do conjunto de variáveis, assim como a escolha da estratégia de redução de dimensionalidade, deve ser ajustada ao alvo clínico em questão, sem hipótese de superioridade universal.

5.4 Análise Comparativa de Desempenho por Classe

A análise por classe revelou padrões que as métricas agregadas mascaram e que possuem implicações clínicas diretas. As Tabelas 19, 21 e 23 apresentam a decomposição por classe para os alvos Transverso, Apoio e Balanço, respectivamente.

No alvo Apoio, a análise comparativa por classe (Tabela 21, Seção 4.5) mostra que a classe SS (Sagital Salto) constitui o gargalo do sistema. Nenhuma abordagem ultrapassou 57% de sensibilidade, refletindo a similaridade biomecânica com SN (ambas compartilham extensão do joelho durante o apoio, diferindo primariamente no equino de tornozelo). É a classe SA que diferencia as estratégias, variando 17 pontos percentuais entre o

pior (PCA) e o melhor ("Com Consenso") modelo. A abordagem Hierárquica revela o custo específico da cascata neste alvo: preserva a detecção de SN, mas reduz SA em 20 pontos percentuais em relação ao *Flat*, padrão consistente com a propagação de erros descrita na Seção 5.2. A Etapa 1 favorece o reconhecimento da classe normal e a Etapa 2 introduz confusões entre os subtipos patológicos.

No alvo Balanço, a análise comparativa de desempenho por classe (Tabela 23, Seção 4.6) expõe um achado clinicamente relevante e contraintuitivo. O modelo com a maior Acurácia Balanceada ("Com Consenso" Hierárquico, 0,702) detecta menos casos de marcha rígida (BR = 50,6%) do que o PCA *Flat* com Acurácia Balanceada inferior (BR = 55,3%). Em outras palavras, o "vencedor" em métrica agregada é o pior do ponto de vista clínico, pois sua superioridade vem de melhor detecção da classe normal (BN, +10 pontos percentuais), à custa da classe patológica que importa rastrear. A variante "Sem Consenso" é a mais frágil clinicamente: 62,4% das crianças com marcha rígida passariam despercebidas. Considerando o equilíbrio entre BN e BR, o PCA *Flat* emerge como a opção mais defensável para uso clínico no alvo Balanço, ainda que atrás em Acurácia Balanceada.

No alvo Transverso, a análise comparativa por classe (Tabela 19, Seção 4.4) mostra que a classe TE (Transverso Externo) é detectada quase perfeitamente nas três abordagens *Flat* (98–100%), enquanto TI (Transverso Interno) é o gargalo (61–67%). Vale notar que PCA e "Sem Consenso" atingem exatamente a mesma sensibilidade para TI (66,5%, com igualdade até a quarta casa decimal), padrão coerente com a interpretação da Seção 5.3: a informação discriminativa para esta classe parece estar concentrada em direções principais do espaço de atributos, capturadas tanto pela retenção integral de variáveis quanto pela redução por componentes principais; o ganho marginal do PCA sobre o "Sem Consenso" vem de outras classes. A excepcional sensibilidade observada para TE, próxima ao limite superior nas três abordagens *Flat*, reflete também a separabilidade biomecânica distintiva desse padrão: a rotação externa do quadril e do tornozelo é marcadamente diferente das demais classes do alvo, condição que facilita a discriminação mesmo sob desbalanceamento amostral severo (TE conta com apenas 57 registros no conjunto de teste, frente a 509 de TN e 194 de TI). A abordagem Hierárquica colapsa neste alvo: TE cai para 58% e TI para 55%, ilustrando o efeito mais severo da propagação de erros entre os três alvos avaliados.

A análise por classe explicita três padrões recorrentes: primeiro, em todos os alvos, a classe minoritária patológica (SS, BR, TI) é o gargalo do desempenho, indicando que o limite do sistema está na capacidade de discriminar fenótipos patológicos sutis, não na detecção de normalidade; segundo, métricas agregadas podem ranquear modelos em ordem

inversa à utilidade clínica, como ilustrado em Balanço, fato que reforça a necessidade de inspeção por classe antes de qualquer recomendação de implantação; terceiro, em todos os três alvos, a abordagem Hierárquica privilegia a detecção da classe normal em detrimento das patológicas, por propagação de erros entre as etapas nos alvos multiclasse (Apoio e Transverso) e por efeito da variabilidade estocástica da otimização em Balanço (Seção 5.2). O desfecho é o mesmo apesar de mecanismos distintos: há ganho em sensibilidade à classe majoritária à custa das classes que importam rastrear.

5.5 Limites Biomecânicos e Estratégias de Mitigação Clínica

A análise por classe (Seção 5.4) identificou três classes, BR (Balanço Rígido), SS (Sagital Salto) e TI (Transverso interno), em que a sensibilidade do melhor modelo permaneceu abaixo do limiar de 70% adotado neste trabalho como referência mínima para triagem clínica de marcha. Esse limiar foi estabelecido com base no piso operacional clássico para instrumentos de triagem em saúde (77,78), conforme detalhado na Seção 5.12, abaixo do qual a fração de falsos negativos torna a ferramenta inadequada para encaminhamento clínico autônomo.

Esta seção examina dois aspectos complementares desse limite. O primeiro é o substrato biomecânico que torna essas três classes especialmente difíceis: restrição articular intrínseca em BR (Seção 5.5.1) e similaridade cinemática com classes adjacentes em SS (Seção 5.5.2). Embora distintos por mecanismo, esses limites compartilham a propriedade de não serem superáveis apenas por aprimoramento algorítmico. O segundo é o desempenho das estratégias de otimização clínica (Seção 4.10) na mitigação dos falsos negativos correspondentes, com ênfase no compromisso entre sensibilidade alcançada e especificidade preservada. A classe TI é discutida brevemente ao final da Seção 5.5.2, dada a proximidade conceitual com BR (ambas refletem limites de separabilidade no espaço cinemático).

5.5.1 Marcha em joelho rígido (BR): limite biomecânico e teto da Estratégia E2

A resistência da classe BR à classificação automática encontra respaldo em evidência biomecânica específica de crianças com PC. Goldberg et al. (79), em estudo com 40 crianças com PC submetidas à transferência do reto femoral (das quais 23 classificadas com

marcha em joelho rígido), demonstraram que a redução da flexão do joelho no balanço, sinal cinemático clássico desse padrão, decorre predominantemente de alterações que se manifestam no duplo apoio (fase final do apoio) e não durante o balanço propriamente dito. Nenhuma das crianças classificadas com joelho rígido apresentou momentos articulares anormais no balanço; ao contrário, 19 de 23 apresentaram velocidade de flexão do joelho abaixo do normal no momento da retirada do pé do chão, com correlação forte com momentos excessivos de extensão do joelho durante o duplo apoio. O substrato muscular desse achado foi caracterizado por Goldberg et al. (80) em simulação dinâmica direta da marcha normal: perturbações individuais de força muscular durante o duplo apoio revelaram que o iliopsoas e o gastrocnêmio são os maiores contribuintes para a velocidade de flexão do joelho nessa fase, enquanto o vasto, o reto femoral e o sóleo atuam em sentido oposto. O quadríceps como um todo apresentou o maior potencial de diminuir o pico de velocidade de flexão, indicando que a hiperatividade do reto femoral no apoio tardio, e não apenas no balanço, pode ser fator determinante da marcha em joelho rígido.

Esse achado tem consequência direta para a classificação automática: a assinatura cinemática de BR observada nos ângulos do balanço é consequência de alterações biomecânicas localizadas na fase de apoio, não plenamente capturadas pelas variáveis angulares do próprio balanço. Esse limite explica por que a Estratégia E2 (ajuste de ponto de corte) saturou para BR: mesmo levando o ponto de corte ao mínimo testado de 0,01, classificando como BR todo registro com probabilidade estimada superior a 1%, a sensibilidade máxima alcançada foi 75,3%, ainda abaixo do alvo de sensibilidade $\geq$ 90% definido para a Estratégia E2 (Seção 3.11). Em contraste, para TE (Transverso Externo) o ponto de corte ótimo de 0,18 atingiu 93% de sensibilidade preservando especificidade de 91,5%, caso em que a E2 funcionou plenamente. Para BR, o problema não está no limiar de decisão, mas na separabilidade intrínseca das classes no espaço de variáveis cinemáticas disponíveis. A melhoria da sensibilidade exigirá provavelmente variáveis complementares, particularmente momentos articulares calculados a partir de plataforma de força e atividade eletromiográfica do reto femoral, músculo cuja contribuição mecânica no apoio tardio foi especificamente identificada pela simulação de Goldberg et al. (80).

5.5.2 Marcha em salto (SS): similaridade com padrões adjacentes e ganhos da Estratégia E3

A dificuldade na classificação de SS é consistente com os achados de Darbandi et al. (20), que reportaram menor acurácia na distinção entre marcha em salto e equino aparente em comparação com equino verdadeiro e marcha agachada, atribuindo o resultado à similaridade biomecânica intrínseca entre os grupos. Caracterização biomecânica complementar foi recentemente publicada por Papageorgiou et al. (42): em estudo com 270 crianças com PC espástica, os autores demonstraram que a marcha em salto é um padrão transicional ao longo do espectro de flexão do joelho, situando-se entre os padrões com joelho estendido (como o equino verdadeiro) e os de flexão acentuada (como a marcha agachada), com cinemática sagital mais próxima do primeiro. O estudo confirmou que marcha em salto, equino aparente e marcha agachada são padrões distintos que compartilham aumento da flexão do joelho durante o apoio, sendo diferenciados principalmente pela posição do tornozelo. A sensibilidade limitada para SS observada no presente estudo (56,6%) reflete, portanto, característica biomecânica intrínseca da classe, não deficiência metodológica do *pipeline*.

Em Apoio, ao contrário de BR, a Estratégia E3 (retreinamento com objetivo clínico) conseguiu reduzir o erro mais grave, a classificação de casos patológicos como Sagital Normal. No modelo padrão, 163 casos patológicos haviam sido classificados como SN, somando os erros vindos de SS (82) e de SA (81). Após o retreinamento, esse número caiu para 33, todos provenientes da classe SS; nenhum caso de SA foi mais classificado como normal. Em contrapartida, a sensibilidade global de SA no modelo E3 ficou em 46,1%, o que indica que parte expressiva dos casos de SA passou a ser classificada como SS, outro padrão patológico, que também leva a encaminhamento clínico. Essa troca é aceitável sob a premissa adotada: classificar um caso patológico como normal é o erro de maior risco para a criança; quando o erro ocorre entre duas classes patológicas (SA confundido com SS), a criança ainda é encaminhada para avaliação.

Para a terceira classe abaixo do limiar clínico, TI (Transverso interno, 66,5%), não há na literatura caracterização biomecânica equivalente do limite de detecção em variáveis angulares isoladas, embora a sobreposição com o padrão normal no plano transverso esteja documentada (Seção 5.4). No conjunto, os três casos sustentam o alerta de Sheldrick et al. (77) de que os limiares de desempenho para instrumentos de triagem clínica têm se tornado progressivamente menos exigentes ao longo do tempo: as estratégias E2 e E3 demonstram que parte do desempenho insuficiente pode ser recuperada por ajuste de ponto de corte ou retreinamento orientado, mas, para classes em que o limite é biomecânico (BR), essa

recuperação é estruturalmente bloqueada e o caminho passa necessariamente por enriquecimento das variáveis preditoras.

5.6 Modelos de Fusão e Análise de Desempenho

Os modelos de fusão apresentaram desempenho competitivo, mas não superaram os melhores modelos individuais em nenhum dos três alvos (Seção 4.3). A margem do líder individual sobre a melhor fusão variou de modesta no Transverso a considerável no Apoio e Balanço.

A ausência de ganho é compatível com os princípios de Dietterich (27): para uma fusão superar seus classificadores base, estes precisam ser simultaneamente acurados (erro inferior a 50%) e diversos, errando em subconjuntos diferentes dos dados (erros ortogonais). Embora os três classificadores base utilizados aqui cobrissem algoritmos de naturezas distintas, os erros tenderam a se concentrar nas mesmas classes biomecanicamente sobrepostas (SS, BR e TI), reduzindo a diversidade efetiva e limitando o ganho marginal da combinação. A mesma lógica de complementaridade é apontada por Wolpert (49) como condição para a eficácia do *Stacking*. Quando um modelo individual é claramente superior, como o *LogReg* em Transverso, a votação entre classificadores base inferiores tende a regredir o desempenho em direção à média do conjunto, em vez de elevá-lo ao nível do melhor componente.

A comparação entre os três modelos de fusão testados, *Voting Hard* (regra da maioria), *Voting Soft* (média das probabilidades) e *Stacking* (meta-modelo treinado sobre predições dos modelos-base), revelou padrão variável por alvo. *Voting Hard* liderou em Apoio e Balanço (com diferença marginal nesse último, dentro da faixa de variabilidade estocástica entre *seeds*), enquanto *Voting Soft* liderou em Transverso. O *Stacking* ficou em terceiro nos três alvos.

Dois achados destacam-se nessa comparação. O primeiro é que o *Stacking* apresentou desempenho inferior aos dois métodos de votação em todos os três alvos, contrariando a expectativa teórica de que um meta-modelo aprendido, capaz de ponderar dinamicamente as predições dos modelos-base, superaria a votação por regras fixas. Esse resultado é coerente com o tamanho amostral do estudo (115 pacientes / 2.564 registros empilhados), considerado limitado para arquiteturas de fusão com camada paramétrica adicional, mais suscetíveis a sobreajuste em condições de pequena amostra (15,28,81).

O segundo é que a inversão entre *Voting Hard* e *Voting Soft* entre alvos diverge da previsão de Kittler et al. (48), segundo a qual a média das probabilidades é mais resiliente a erros na estimação de probabilidades. A divergência coincide com má calibração probabilística nos alvos mais desbalanceados (Apoio e Balanço apresentaram ECE acima do limiar de referência de 0,10 para todas as classes), enquanto em Transverso a classe majoritária TN, que domina o cálculo da média de probabilidades, manteve ECE na fronteira da faixa razoável. Essa diferença é coerente com as propriedades dos algoritmos vencedores em cada alvo: Niculescu-Mizil e Caruana (72) documentaram que Naive Bayes (que assume independência condicional entre variáveis) tende a produzir probabilidades extremas e que a Árvore de Decisão (cujas estimativas se baseiam em proporções de nós folha) apresenta calibração instável; em Transverso, a Regressão Logística (algoritmo bem calibrado por construção) venceu, preservando a vantagem teórica da regra da soma. Probabilidades mal calibradas comprometem a regra da soma e justificam o desempenho relativamente melhor da votação por maioria nos alvos com calibração deficiente.

Por fim, um número expressivo de modelos apresentou MCC negativo (desempenho pior que classificação aleatória), com forte concentração no alvo Balanço, coerente com o desbalanceamento extremo desse alvo (razão BN/BR de 7,7:1). Uma fração desses modelos chegou a apresentar G-Mean zero, indicando sensibilidade nula em pelo menos uma classe. O achado reforça que métricas tradicionais como acurácia podem ocultar desempenho próximo ao acaso em dados desbalanceados, conforme demonstrado por Brodersen et al. (29).

5.7 Adequação dos Algoritmos ao Problema

Os três alvos foram vencidos por algoritmos de complexidade baixa a moderada (Regressão Logística em Transverso, *Naive Bayes* em Apoio e Árvore de Decisão em Balanço), enquanto modelos de alta capacidade representacional (MLP, SVM com *kernel* RBF, XGBoost e LightGBM) ficaram sistematicamente para trás. O padrão, contraintuitivo à primeira vista, é coerente com três características do problema: amostra moderada (115 pacientes, aproximadamente 2.500 registros após empilhamento bilateral), espaço de atributos pequeno após pré-processamento (12 a 25 variáveis) e relações predominantemente lineares em dois dos três alvos (Transverso e Balanço, conforme Seção 5.3). Em contextos com essas

características, modelos parcimoniosos tendem a generalizar melhor que modelos cuja capacidade representacional excede a complexidade real do problema, conforme estabelecido pela teoria do *trade-off* viés–variância (*bias–variance trade-off*) (82): modelos de alta capacidade reduzem o viés (erro por simplificações estruturais), mas amplificam a variância (sensibilidade a ruídos do conjunto de treino), efeito particularmente problemático em amostras moderadas, situação em que arquiteturas mais simples tendem a ocupar o ponto ótimo desse compromisso. Esse princípio é sintetizado por Domingos (83) como guia prático do aprendizado de máquina aplicado.

A Regressão Logística é um modelo linear com regularização L2, técnica que evita que o modelo dependa demais de poucas variáveis e generaliza melhor para novos pacientes. Tem baixa capacidade de memorização (*overfitting* médio de 0,196) e produz probabilidades naturalmente calibradas, propriedade demonstrada empiricamente por Niculescu-Mizil & Caruana (72). Sua adequação à fronteira de decisão essencialmente linear do Transverso é o resultado esperado, e o *overfitting* de apenas 0,036 do melhor modelo confirma o bom ajuste entre estrutura algorítmica e estrutura do problema.

O *Naive Bayes* opera sob a premissa de independência condicional entre variáveis. Essa premissa é frequentemente violada na prática, mas pode ser parcialmente atenuada pela seleção por consenso, que filtra variáveis altamente correlacionadas. Além disso, Domingos & Pazzani (84) demonstraram formalmente que o classificador Bayesiano ingênuo pode acertar a classe correta mesmo quando a premissa de independência é violada, ou seja, mesmo gerando probabilidades imprecisas, ele preserva a ordem relativa entre as classes, o que basta para classificação correta. Esse achado justifica seu desempenho sólido em problemas de classificação multiclasse moderadamente desbalanceados. Em amostras moderadas, sua simplicidade estrutural funciona como regularização forte (*overfitting* médio de 0,159, o menor do estudo).

A Árvore de Decisão beneficia-se da natureza binária do problema Balanço (BN vs. BR), que admite solução por poucas regras condicionais. Entretanto, sua instabilidade característica, bem documentada por Breiman (74), faz com que o resultado dependa da configuração específica encontrada pela otimização estocástica, conforme discutido na Seção 5.2. Essa mesma característica explica o *overfitting* de 0,263 do modelo vencedor, o mais alto entre os três alvos (contra 0,036 no Transverso e 0,082 no Apoio). O *gap*, contudo, é propriedade do algoritmo e não do alvo: as posições 2 a 5 do próprio Balanço (Tabela 22) são ocupadas pela Regressão Logística, regularizada por L2, com Acurácia Balanceada quase equivalente (0,674–0,667), o que indica que a magnitude do *overfitting* decorre da capacidade

de memorização da árvore, e não da dificuldade intrínseca do alvo binário. A robustez do ranqueamento neste alvo é, contudo, reforçada pela concordância entre as duas principais métricas de avaliação: o modelo vencedor por Acurácia Balanceada coincide com o líder em AUC, indicando que o desempenho observado não é artefato da escolha da métrica de seleção.

A contrapartida explica por que os modelos de alta capacidade falharam. MLP, SVM com *kernel* RBF, XGBoost e LightGBM apresentaram *overfitting* médio entre 0,45 e 0,49 no alvo Apoio, com valores individuais atingindo 0,64 em configurações específicas. Esses algoritmos têm alta capacidade de representação, característica que exige conjuntos de treino maiores para estabilizar parâmetros sem memorizar ruído amostral, conforme formalizado pela teoria do aprendizado estatístico (82,85). Além disso, possuem espaços de hiperparâmetros amplos, cuja otimização confiável demandaria mais do que os 50 *trials* de Optuna utilizados neste estudo.

O K-NN apresenta limitação adicional: em 25 variáveis sem ponderação, sofre com a maldição da dimensionalidade. Trata-se do fenômeno em que as distâncias entre pontos no espaço de atributos perdem capacidade discriminativa à medida que a dimensão cresce, consequência bem documentada para algoritmos baseados em vizinhança por Beyer et al. (86).

Esses achados convergem com a observação de Slijepcevic et al. (15), que reportaram, em amostra de 302 pacientes com PC, melhor generalização de algoritmos tradicionais comparados a redes neurais profundas. No conjunto, os resultados ilustram um princípio geral do aprendizado de máquina aplicado: a complexidade do modelo deve ser proporcional à complexidade real do problema, e não à expectativa de que "mais complexo = melhor" (83). Em problemas com amostra moderada, dimensionalidade reduzida e relações majoritariamente lineares, modelos parcimoniosos vencem pela coerência entre estrutura algorítmica e estrutura do problema.

5.8 Análise do *Overfitting*

A análise comparativa entre as variantes "Com SMOTE" e "Sem SMOTE" entre os cinco melhores modelos de cada alvo, sintetizada na Tabela 27 (Seção 4.9), revela padrão alvo-dependente. O "Sem SMOTE" venceu em dois dos três alvos: por desempate via AUC no Transverso e de forma dominante no Balanço, alvo em que os cinco melhores foram todos modelos sem balanceamento sintético. Apenas no Apoio o "Com SMOTE" foi vantajoso,

superando o melhor "Sem SMOTE" em 0,053 pontos absolutos de Acurácia Balanceada. Essa heterogeneidade tem implicação direta para a leitura do *overfitting*: o SMOTE não pode ser interpretado como causa universal da diferença observada entre treino e teste, dado que o melhor modelo do estudo (*LogReg* / Geral / PCA / "Sem SMOTE" para Transverso, Acurácia Balanceada = 0,791 e *overfitting* = 0,036) dispensou o balanceamento sintético. O resultado é coerente com a observação de Chawla et al. (64) de que, em atributos com alta variância, as amostras sintéticas podem sobrepor-se ao espaço da classe majoritária e degradar a discriminação, e fundamenta a recomendação metodológica de tratar o SMOTE como hiperparâmetro a ser validado por alvo, e não como etapa obrigatória de pré-processamento.

Três fatores concorrentes explicam a magnitude do *overfitting* observado neste estudo. O primeiro fator, já detalhado na Seção 3.8.3, é o *dataset shift* intencional introduzido pela aplicação do SMOTE exclusivamente ao conjunto de treino: enquanto o treino opera sobre distribuição artificialmente balanceada, o teste preserva a distribuição original desbalanceada, o que infla sistematicamente a lacuna entre treino e teste sem configurar *overfitting* problemático no sentido tradicional. Como evidenciado pela Tabela 27 (Seção 4.9), este fator atua apenas sobre o subconjunto de configurações com SMOTE, deixando os modelos "Sem SMOTE" imunes a essa fonte de inflação. O segundo fator é o tamanho amostral, moderado para o padrão do campo: 115 pacientes, totalizando 2.549 registros para o alvo Apoio e 2.564 registros para Balanço e Transverso após o empilhamento bilateral. Embora superior ao de parte dos estudos anteriores em classificação da marcha na PC (19,20), o conjunto permanece aquém do necessário para estabilizar algoritmos de alta capacidade de representação. SVM com *kernel* RBF, LightGBM e Redes Neurais apresentaram *overfitting* médio entre 0,45 e 0,49 no alvo Apoio, com valores individuais atingindo até 0,64 em configurações específicas. O terceiro fator é o desbalanceamento estrutural das classes patológicas: com razões de 8,9:1 (Transverso) e 7,9:1 (Balanço), o espaço de atributos das classes minoritárias é representado por poucas dezenas de exemplos no treino, favorecendo modelos que memorizam padrões específicos do treino em vez de aprender representações generalizáveis para o conjunto de teste.

A combinação destes três fatores indica que o *overfitting* observado deve ser interpretado como característica estrutural do problema, e não como falha do *pipeline*. Três evidências sustentam essa interpretação. (I) O melhor modelo do estudo apresentou *overfitting* de apenas 0,036, demonstrando que é possível obter alto desempenho e boa generalização simultaneamente quando o algoritmo, a estratégia de pré-processamento e a estrutura do alvo estão alinhados. (II) Os modelos com menor capacidade de memorização (*Naive Bayes* com

média de 0,159 e Regressão Logística com 0,196) foram exatamente os que ocuparam as primeiras posições nos alvos Apoio e Transverso, em convergência com Slijepcevic et al. (15), que documentaram em 302 pacientes com PC melhor generalização de algoritmos tradicionais comparados a redes neurais profundas, com *overfitting* ocorrendo em todos os modelos mesmo após otimização criteriosa de hiperparâmetros. (III) A preservação da distribuição desbalanceada no conjunto de teste é a escolha metodológica clinicamente correta, pois reflete a prevalência real das condições na população-alvo; protocolos que aplicam SMOTE também ao teste produzem estimativas de desempenho artificialmente otimistas que não se sustentam em uso clínico real (69). Portanto, o *overfitting* reportado neste estudo é o preço necessário de uma avaliação clinicamente honesta, não um indicador de falha do modelo. Decorre disso uma recomendação metodológica direta para estudos futuros: aplicar técnicas de balanceamento sintético exclusivamente ao conjunto de treino, preservando a distribuição natural das classes no conjunto de teste.

5.9 Comparação com a Literatura

O trabalho de Winters et al. (40) representa um marco histórico na classificação objetiva da marcha em hemiplegia espástica, demonstrando que quatro padrões homogêneos podiam ser identificados em 46 pacientes pela combinação de cinemática sagital e eletromiografia dinâmica. Conforme discutido nas Seções 2.6.1 e 2.6.2, a limitação estrutural dessas classificações uniplanares foi evidenciada por Morais Filho et al. (43), reforçando a pertinência do GCS-CP como alvo do presente estudo por incorporar componentes multiplanares com critérios cinemáticos objetivos.

Estudos anteriores de ML aplicados à classificação da marcha em PC reportaram acurácias agregadas amplamente variáveis, frequentemente entre 70% e mais de 90%, conforme o tipo de dado, a granularidade do alvo e o tamanho da amostra (15,19,20,47). Tipicamente, esses trabalhos não reportam métricas apropriadas para dados desbalanceados, o que dificulta comparações diretas com os resultados aqui obtidos. Por isso, o limiar de 0,70 de Acurácia Balanceada adotado nesta dissertação ancora-se no piso operacional clássico de sensibilidade para instrumentos de triagem clínica (77,78,87), e não em tais acurácias agregadas, que raramente refletem desempenho clinicamente honesto em cenários desbalanceados. Esse panorama é confirmado por Nahar et al. (88): em revisão sistemática de 20 estudos sobre ML em PC publicados entre 2013 e 2023, identificaram que dados clínicos e

de marcha são os mais utilizados, Floresta Aleatória é o algoritmo mais frequente, e a maioria dos estudos reporta apenas acurácia agregada, sem decomposição consistente por classe, lacuna metodológica que o presente estudo buscou endereçar sistematicamente.

Kamruzzaman e Begg (19) reportaram acurácia de 96,8% com SVM para diferenciação entre crianças com diplegia espástica e controles saudáveis (88 com PC e 68 controles, validação cruzada de 10 dobras), alcançando AUC de 0,9924 com dados normalizados. Os autores demonstraram que a normalização dos parâmetros da marcha por comprimento da perna e idade foi crucial, elevando a acurácia de 83,3% para 96,8%, ganho de 13,5 pontos percentuais. Entretanto, o referido estudo utilizou apenas dois parâmetros espaço-temporais (comprimento da passada e cadência), sem incorporar informações cinemáticas articulares, e restringiu-se a uma tarefa distinta da aqui investigada: a discriminação binária entre marcha patológica e típica (PC versus controles saudáveis). Trata-se de um problema intrinsecamente mais simples, pois contrasta grupos biomecanicamente muito distintos, ao passo que o presente estudo busca diferenciar subtipos sobrepostos de um mesmo quadro (classificação multiclasse de padrões específicos da marcha na PC), cujas fronteiras cinemáticas são consideravelmente mais estreitas. Por essa razão, os elevados índices de acurácia reportados para a detecção da condição não são diretamente transponíveis para a classificação dos padrões da marcha, lacuna que motiva a presente investigação.

Darbandi et al. (20), utilizando sistema fuzzy para classificação dos padrões de Rodda, reportaram desempenho fortemente heterogêneo entre as classes: sensibilidade e especificidade de 100% para equino verdadeiro e agachamento, contra erros sistemáticos entre salto e equino aparente. A heterogeneidade de desempenho por classe observada no presente estudo, com sensibilidade variando de 56,6% (Salto) a 100% (Transverso Externo) no modelo padrão, é consistente com esse padrão geral da literatura: a separabilidade entre classes não é uniforme, e classes biomecanicamente próximas tendem a ser sistematicamente confundidas.

Yoon et al. (47), no maior estudo de classificação da marcha em PC (833 participantes), obtiveram alta acurácia (~92%) para distinguir pacientes com PC de controles, mas desempenho substancialmente menor e mais variável (53–93%) para discriminar níveis GMFCS adjacentes; a tarefa 1 vs. 2 foi identificada pelos próprios autores como a mais difícil, atribuída à subjetividade inerente ao GMFCS. A fonte da dificuldade é, contudo, distinta da observada no presente estudo: em Yoon et al. (47), a incerteza decorre da subjetividade do rótulo; no Apoio Geral, da sobreposição cinemática entre padrões definidos por critérios objetivos. A convergência entre os dois cenários reforça que a classificação multiclasse de marcha em PC enfrenta limites estruturais, de rotulagem ou de separabilidade, que dificilmente

são superados apenas por aprimoramento algorítmico. Isso não implica, contudo, que o desempenho seja inerentemente intransponível, mas sim que os avanços tendem a depender da origem da limitação: quando o gargalo é a rotulagem, da melhoria da qualidade do rótulo (por exemplo, consenso entre múltiplos avaliadores ou rótulos probabilísticos); quando é a separabilidade, do enriquecimento da informação de entrada (como curvas cinemáticas completas, variáveis cinéticas ou dados de múltiplos planos) e da ampliação da amostra, e não da mera substituição de algoritmos.

A opção de não incluir dados cinéticos (força de reação do solo) é respaldada pelos achados de Slijepcevic et al. (15), cujas configurações com dados cinemáticos atingiram 93,4% de acurácia para os quatro padrões de Rodda, contra apenas 47,2% com força de reação do solo isoladamente, diferenças de 32,6 a 51,4 pontos percentuais conforme o algoritmo. Soma-se a isso a indisponibilidade de dados cinéticos para participantes GMFCS III com dispositivos auxiliares de marcha, conforme descrito na Seção 3.2.

Uma contribuição metodológica relevante deste trabalho foi demonstrar que a escolha entre seleção de variáveis por consenso e PCA é alvo-dependente, sem superioridade universal de nenhuma abordagem para dados biomecânicos de marcha, achado que, até onde é de nosso conhecimento, não encontra precedente direto na literatura da marcha em PC. Samadi Kohnehshahri et al. (24), em revisão sistemática de 63 estudos sobre ML aplicado à marcha em PC e AVE, identificaram que a falta de racionalidade para seleção de variáveis e algoritmos, o uso de conjuntos de dados não representativos e a falta de interpretação clínica dos resultados são os principais fatores que limitam a translação de modelos de ML para a prática clínica. Em consonância com essas recomendações, o presente estudo procurou incorporar práticas frequentemente ausentes na literatura analisada: (I) justificou explicitamente a seleção de variáveis, prática presente em apenas cerca de 6% (4 de 63) dos estudos revisados; (II) preveniu vazamento de dados através de divisão rigorosa por paciente, complementando a recomendação geral da revisão sobre *datasets* representativos; e (III) mostrou que as classes preditas correspondem de forma direta aos padrões biomecânicos definidos pelo GCS-CP, lacuna de interpretação clínica identificada em 18 dos 32 estudos não supervisionados da revisão (interpretação clínica adequada presente em apenas 14, ou 44%).

Outro ponto relevante para contextualizar os resultados é o chamado “teto humano” de desempenho. Melanda et al. (14) reportaram confiabilidade inter-avaliadores apenas moderada para o primeiro nível do GCS-CP (κ total = 0,596), indicando que mesmo avaliadores experientes discordam em parcela relevante das classificações. Como o padrão-

ouro usado no treinamento é, ele mesmo, ruidoso, um classificador supervisionado dificilmente excede a concordância entre os avaliadores que geraram seus rótulos.

Para Transverso e Balanço, o modelo permaneceu abaixo desse teto, comportamento esperado. Para Apoio, houve superação aparente (κ = 0,490 vs κ = 0,409, vantagem de 0,081), mas o resultado deve ser interpretado com cautela: o sagital em apoio é justamente o plano com menor confiabilidade interavaliador entre os três (14); o teto baixo, e não a superioridade do modelo, explica boa parte da diferença. Assim, os desempenhos moderados em Transverso e Apoio refletem mais a fronteira imposta pela discordância entre avaliadores humanos do que limitação do *pipeline* de aprendizado de máquina.

Em síntese categórica pela escala de Landis e Koch (45), os três alvos preservam a hierarquia descrita no parágrafo anterior, com nuances específicas. Em Transverso, o modelo (κ = 0,491) permanece na mesma faixa "moderada" (0,41–0,60) do inter-avaliador humano (κ = 0,516), ligeiramente abaixo do teto. Em Apoio, o modelo (κ = 0,490) e o inter-avaliador (κ = 0,409) também ficam ambos na faixa moderada, com vantagem numérica do modelo já discutida e qualificada acima. Para Balanço, o modelo (κ = 0,354) caiu para a categoria "razoável" (0,21–0,40), afastando-se do teto humano (κ = 0,570) e indicando que, nesse alvo específico, existe espaço para melhoria independente da qualidade dos rótulos, provavelmente associado à separabilidade intrínseca insuficiente entre BN e BR no espaço cinemático disponível, conforme discutido na Seção 5.5.2.

5.10 Otimização Clínica e Minimização de Falsos Negativos

A avaliação dos modelos pela Acurácia Balanceada, embora necessária para comparação sistemática entre algoritmos, não responde à pergunta clinicamente mais relevante: quantas crianças com padrão patológico deixariam de ser encaminhadas para avaliação especializada? Esta seção discute os resultados da otimização clínica (Seção 4.10), que abordou diretamente essa questão através de três estratégias complementares aplicadas às cinco classes patológicas do estudo.

A assimetria entre custos de classificação errada, em que um falso negativo (criança patológica classificada como normal) acarreta consequências clínicas mais graves que um falso positivo (encaminhamento desnecessário), é reconhecida como problema estrutural da inteligência artificial aplicada à medicina. Mohammadi-Seif e Baeza-Yates (33) argumentam que funções de perda padrão tratam todos os erros do modelo como equivalentes. Como

consequência, modelos podem acumular taxas inaceitáveis de falsos negativos enquanto ainda exibem boas métricas agregadas, como a acurácia média. Os autores propõem distinguir dois tipos de erro: a "ambiguidade visual", em que o modelo confunde subtipos próximos da mesma categoria (erro relativamente seguro do ponto de vista clínico), e a "incoerência semântica", em que o modelo confunde categorias clinicamente distintas, caso dos falsos negativos críticos, em que um paciente patológico é classificado como normal. Esta distinção é pertinente ao presente estudo, no qual confusões entre padrões patológicos (p.ex., SS ↔ SA) representam ambiguidade visual aceitável, enquanto confusões patológicas → normal (p.ex., SS → SN) configuram falhas estruturais.

O *RandomForest* com ponderação clínica (RF_PesoClinico), avaliado no ponto de corte 0,50, apresentou falsos negativos por classe variando de 33 (TE) a 97 (TI), com SS em 82, SA em 81, BR em 70 e TE em 33 no conjunto de teste. Esses números representam crianças com padrão patológico classificadas incorretamente como normais, casos que, em um cenário de triagem clínica, não receberiam encaminhamento para avaliação especializada. A magnitude desses erros justifica a busca por estratégias de otimização específicas por classe. Foram avaliadas três estratégias: E1 (seleção, entre os modelos já treinados, daquele com maior sensibilidade patológica em vez do modelo com maior Acurácia Balanceada), E2 (ajuste do ponto de corte de decisão sobre Floresta Aleatória com peso clínico) e E3 (retreinamento do Floresta Aleatória via Optuna, com função objetivo orientada à sensibilidade patológica mínima). A Estratégia E1 produziu ganhos em todas as cinco classes patológicas, com destaque para a eliminação completa dos falsos negativos do TE (de 33 para zero), melhor resultado entre todas as estratégias avaliadas para essa classe, alcançado pela simples seleção de um *pipeline* com maior sensibilidade mínima patológica, sem qualquer ajuste de limiar ou retreinamento. Para as demais classes, a E1 reduziu os FN de SA em 55 (de 81 para 26), TI em 33 (de 97 para 64), BR em 32 (de 70 para 38) e SS em 31 (de 82 para 51), sendo, nessas quatro classes, superada pelas Estratégias E2 ou E3; por essa razão, a discussão subsequente concentra-se em E2 e E3 para SS, SA, BR e TI.

O ponto de corte de 0,01 encontrado para TI e BR representa o limite inferior do espaço de busca. O ganho de sensibilidade nesse extremo não decorre de aprendizado discriminativo, mas da saturação das predições na classe patológica: o modelo minimiza sua função de perda classificando quase todos os casos como doentes. O algoritmo de otimização converge para essa solução como compromisso matemático imposto pela sobreposição das distribuições no espaço de variáveis disponíveis.

A escolha por ajustar o ponto de corte de decisão como estratégia de mitigação clínica encontra fundamento na literatura. Hoang e Liang (73) demonstraram que o ponto de corte padrão de 0,50 é adequado apenas quando as classes são perfeitamente balanceadas, os custos de classificação errada são simétricos e o modelo é bem calibrado, condições raramente satisfeitas em conjuntos de dados médicos. Os autores reforçam que a calibração do ponto de corte deve constituir prática padrão em *pipelines* clínicos de aprendizado de máquina.

A Estratégia E2, baseada no ajuste do ponto de corte de decisão, revelou padrão heterogêneo entre as classes patológicas: produziu resultado aceitável para TE, porém colapso de especificidade para TI e BR. Para TE, o ajuste do ponto de corte de 0,50 para 0,18 sobre o Floresta Aleatória com peso clínico reduziu os falsos negativos de 33 para 4 (redução de 88%), com perda de especificidade de 6,5 pontos percentuais (98,0% → 91,5%), resultado clinicamente aceitável, embora inferior ao da Estratégia E1 (0 FN, sem custo de especificidade), que constitui a escolha prioritária para essa classe. Para TI e BR, em contraste, o ponto de corte ótimo encontrado foi 0,01 (limite inferior testado), e mesmo nesse extremo da curva a sensibilidade não alcança a meta de 90% (86,6% em TI e 75,3% em BR), enquanto a especificidade colapsa (97,2% → 12,5% em TI; 94,7% → 27,6% em BR), inviabilizando aplicação clínica direta. Esses resultados reforçam que BR e TI são as classes mais resistentes à otimização. A causa não é limitação do ponto de corte, nem o limiar mínimo testado eleva a sensibilidade ao patamar clínico desejado, mas separabilidade intrínseca insuficiente entre essas classes no espaço de variáveis cinemáticas disponíveis. Superá-la exigirá, provavelmente, variáveis complementares (cinéticas ou eletromiográficas), conforme discutido na Seção 5.13.

Este padrão de eficácia diferenciada entre estratégias é coerente com a análise comparativa de abordagens de aprendizado sensível ao custo reportada na literatura. A revisão sistemática de 173 estudos em dados médicos desbalanceados conduzida por Araf, Idri e Chairi (32) documenta que apenas 2,9% dos trabalhos utilizam pontos de corte como estratégia principal, apesar de evidências de que o ajuste de limiar pode superar abordagens de retreinamento em cenários com modelos adequadamente especificados. Os autores identificam ainda que ortopedia representa apenas 2,3% dos estudos revisados e que aplicações em análise biomecânica da marcha não figuram entre os domínios investigados, posicionando a presente dissertação como aplicação pioneira de estratégias sistemáticas de *cost-sensitive learning* à classificação automatizada da marcha em paralisia cerebral entre os domínios revisados por Araf, Idri e Chairi (32).

A abordagem adotada na Estratégia E3, que pondera as classes patológicas durante a otimização do Floresta Aleatória via o hiperparâmetro *boost_patologica*, alinha-se à tendência recente na literatura de IA médica de incorporar a assimetria de custos diretamente no processo de aprendizado, e não apenas no pós-processamento. Mohammadi-Seif e Baeza-Yates (33) avaliaram uma função de perda calibrada por risco em quatro modalidades de imagem médica (MRI cerebral, dermatoscopia, histopatologia mamária e prostática). Nessa função, erros do Tipo II (patológico classificado como normal) recebem penalidade β substancialmente maior do que erros do Tipo I (normal classificado como patológico), reduzindo expressivamente os falsos negativos críticos frente a abordagens convencionais. Embora os autores tenham trabalhado com classificação binária por superclasse e redes profundas, o princípio é transponível: a ponderação das classes patológicas adotada neste estudo pode ser compreendida como um caso particular da mesma lógica de otimização assimétrica, aplicada a um problema multiclasse com algoritmos clássicos. Este alinhamento metodológico reforça a validade da Estratégia E3 e sinaliza caminho natural para expansão, discutido na Seção 5.13.

A Estratégia E3 retreina o Floresta Aleatória com o Optuna, mas com um objetivo diferente: em vez de buscar a melhor acurácia média, ela busca a maior sensibilidade possível nas classes patológicas. Para o alvo Apoio, foi a estratégia que mais reduziu encaminhamentos perdidos. No SA, os falsos negativos como Normal caíram de 81 para zero, ou seja, nenhuma criança com marcha agachada foi classificada como normal. No SS, a queda foi de 60% (de 82 para 33 casos tidos como Normal, sensibilidade de 80,7%).

Vale separar aqui dois erros bem diferentes. O primeiro é classificar como Normal uma criança que tem alteração: ela não é encaminhada e o diagnóstico se perde. Esse é o erro grave, e foi ele que a E3 zerou no SA. O segundo é confundir SA com SS, ou seja, trocar um subtipo patológico por outro. Nesse caso a criança continua sendo encaminhada, só com o rótulo errado. É essa confusão entre SA e SS que mantém a sensibilidade do SA em 46%, e não a classificação como Normal. Para uma ferramenta de triagem, cujo papel é não deixar passar o caso patológico, trocar o erro grave por esse erro mais leve é uma troca vantajosa.

Esses resultados mostram que mudar o objetivo da otimização, de "maximizar a acurácia" para "maximizar a sensibilidade patológica", traz ganhos clínicos reais. O preço é a queda da Acurácia Balanceada do Apoio, que vai de 0,717 para 0,434. Mas essa queda é esperada, não um defeito: ao empurrar as predições para longe da classe Normal, o modelo protege as classes patológicas e, em troca, passa a errar mais na classe Normal, classificando

como alterados vários casos que eram normais. Como a Acurácia Balanceada trata todas as classes com o mesmo peso, ela enxerga como perda justamente aquilo que, na triagem, é ganho.

A Tabela 30 sintetiza a viabilidade clínica das estratégias de otimização para cada classe patológica, integrando sensibilidade alcançável, estratégia vencedora e custo associado.

Tabela 30 – Viabilidade clínica das estratégias de otimização por classe patológica

Classe	Sensibilidade	Estratégia	Custo principal	Viabilidade clínica
TE	100,0%	E1 (seleção de *pipeline*)	Sem ajuste de ponto de corte; mudança de *pipeline*	☑ **Plenamente viável**
SA	Sens. clínica 100%[1] (recall 46%)	E3 (retreinamento)	Bal.Acc do alvo Apoio: 0,717 → 0,434	⚠ **Com ressalvas**
SS	80,7%[2]	E3 (retreinamento)	Bal.Acc do alvo Apoio: 0,717 → 0,434	⚠ **Com ressalvas**
TI	86,6%	E2 (thr = 0,01)	Especificidade colapsa 97,2% → 12,5%	✗ **Inviável**
BR	75,3%	E2 (thr = 0,01)	Especificidade colapsa 94,7% → 27,6%	✗ **Inviável**

Nota: Linhas ordenadas por viabilidade clínica decrescente. A coluna 'Sensibilidade' refere-se ao valor obtido pela estratégia indicada (E2 ou E3) para cada classe; a estratégia foi selecionada por produzir o menor número absoluto de falsos negativos para a classe Normal, conforme detalhado nos parágrafos anteriores. A Estratégia E1 é a estratégia vencedora para TE (Tabela 28); para SS, SA, BR e TI foi superada por E2 ou E3 e os resultados dessas classes refletem a estratégia de menor FN entre E2 e E3. [1] Para SA, "sensibilidade clínica" = proporção de casos SA não classificados como Normal (0 FN para Normal em 81 casos = 100%). O *recall multiclasse* convencional permanece em 46% porque 54% dos casos SA são classificados como SS, classe igualmente patológica, que também gera encaminhamento. Para triagem, a métrica relevante é a sensibilidade clínica. [2] Para SS, a sensibilidade reportada (80,7%) é a obtida pela E3, estratégia indicada na linha. A E2 alcança sensibilidade nominalmente maior para SS (90,1%), mas com 42 falsos negativos contra 33 da E3; o critério de seleção da estratégia é o menor número absoluto de falsos negativos para Normal, razão pela qual E3 prevalece apesar da sensibilidade ligeiramente menor. ☑ = compromisso clínico plenamente favorável (sensibilidade alta com especificidade preservada); ⚠ = sensibilidade alta atingida ao custo de queda na Acurácia Balanceada global do alvo, exigindo ressalvas para uso como modelo único de triagem; ✗ = sensibilidade alta atingida apenas ao custo do colapso da especificidade, inviabilizando aplicação clínica. thr = *ponto de corte* de decisão; pp = pontos percentuais; Bal.Acc = Acurácia Balanceada. As classes patológicas referidas são SS (Sagital Salto), SA (Sagital Agachado), BR (Balanço Rígido), TI (Transverso interno) e TE (Transverso externo).
Fonte: *Elaborada pelo autor.*

Do ponto de vista clínico, e conforme apresentado na Tabela 30, nenhuma criança com marcha agachada (SA) precisaria ser enviada para casa sem diagnóstico, e o número de crianças com Transverso Externo não detectadas poderia ser eliminado completamente (de 33 para zero) pela simples seleção do *pipeline* de maior sensibilidade para essa classe (Estratégia E1). Os desafios remanescentes são BR e TI, cuja resistência à otimização sugere que a discriminação dessas classes no espaço cinemático disponível atingiu seu limite estrutural. Sua detecção provavelmente exigirá variáveis complementares, possivelmente parâmetros cinéticos, eletromiográficos ou séries temporais completas do ciclo da marcha.

A análise conjunta das estratégias, sintetizada na Tabela 30, revela um padrão claro: para quatro das cinco classes patológicas avaliadas (TE, SA, SS e TI), sensibilidades clinicamente substanciais (entre 80% e 100%) são alcançáveis com as variáveis cinemáticas disponíveis, embora apenas TE permita esse ganho com compromisso clínico plenamente favorável. SA e SS atingem sensibilidades altas ao custo de queda significativa da Acurácia Balanceada global, configurando uso com ressalvas; TI e BR apresentam colapso de especificidade que inviabiliza aplicação direta. Esse mapeamento sugere que um *pipeline* clínico completo, combinando retreinamento com função objetivo orientada à sensibilidade para todos os algoritmos disponíveis, poderia atingir compromisso favorável para SA e SS. TI e BR permaneceriam como desafios remanescentes, a serem endereçados por variáveis complementares. Essa perspectiva exige investigação adicional, já que a E3 testada neste estudo (restrita ao Floresta Aleatória) não se mostrou eficaz para TI: sensibilidade de 59% contra 86,6% obtidos pela E2. A Seção 5.13 aprofunda essas direções.

5.11 Implicações Clínicas

Os resultados deste estudo possuem implicações práticas relevantes para laboratórios de análise do movimento: a capacidade de classificar automaticamente padrões da marcha pode padronizar a avaliação, reduzindo variabilidade interavaliador; acelerar o processo de análise; fornecer uma segunda opinião objetiva para casos duvidosos; e facilitar estudos epidemiológicos de larga escala. Conforme destacado por Lai et al. (16), a adoção de técnicas de inteligência computacional na prática clínica tem sido lenta devido à natureza de "caixa-preta" de muitos algoritmos, questão de interpretabilidade diretamente endereçada por este estudo, conforme retomado ao final desta seção.

A detecção do padrão Transverso com Acurácia Balanceada de 0,791 e a sensibilidade de 100% para Transverso Externo indicam potencial dos modelos como ferramentas de triagem inicial, permitindo direcionar a atenção do especialista para casos que requerem avaliação mais detalhada. Adicionalmente, os resultados da otimização clínica (Seção 4.10) demonstraram que, com estratégias orientadas à sensibilidade, é possível eliminar completamente os falsos negativos para SA e TE, e reduzir os de SS, ampliando o potencial clínico dos modelos desenvolvidos.

A definição de limiares mínimos de acurácia para instrumentos de triagem permanece, contudo, objeto de debate. Não foram identificadas referências que estabeleçam limiares validados especificamente para classificadores automáticos de marcha em paralisia cerebral, lacuna metodológica do campo. Na ausência de critérios específicos, adotaram-se os limiares propostos para triagem clínica em geral. Glascoe (78) propõe 70% de sensibilidade como referência mínima aceitável. Plante e Vance (87) defendem critério mais rigoroso, com dois patamares categóricos: sensibilidade e especificidade ≥ 80% para um instrumento ser considerado aceitável e ≥ 90% para ser considerado bom. Segundo os autores, abaixo de 80% as taxas de classificação incorreta seriam inaceitáveis. Sheldrick et al. (77), por sua vez, argumentam que esses padrões carecem de fundamentação em análises de custo-benefício e refletem julgamentos implícitos de valor mais do que critérios empiricamente validados. A transposição desses limiares para a marcha automatizada requer validação futura.

A adoção do limiar de 70% no presente estudo reflete, portanto, posição alinhada ao uso dos modelos como ferramentas de triagem, e não como diagnóstico definitivo. Padrões mais rigorosos seriam necessários para aplicação clínica autônoma, e os resultados devem ser interpretados de forma probabilística e contextualizada, posição reforçada pela revisão sistemática de Samadi Kohnehshahri et al. (24), que analisou 63 estudos sobre ML em análise da marcha em PC e AVC e concluiu que a relevância clínica dos modelos avaliados permanece comprometida pela ausência de justificativa na seleção de algoritmos e pela falta de interpretação. Essa cautela é especialmente importante em decisões terapêuticas de alto impacto, como indicação de toxina botulínica, transferências musculares ou osteotomias rotacionais.

A responsabilidade metodológica na implementação clínica é igualmente crítica. Slijepcevic et al. (15), em estudo comparativo com 302 pacientes com paralisia cerebral distribuídos em quatro classes (marcha agachada, equino aparente, marcha em salto e equino verdadeiro), demonstraram que o Floresta Aleatória superou redes neurais profundas tanto em acurácia quanto em transparência metodológica. A melhor configuração, combinando ângulos

sagitais de joelho e tornozelo com Floresta Aleatória, alcançou 93,4% de acurácia média, com foco em regiões biomecanicamente relevantes do ciclo da marcha. Os autores reforçam que explicabilidade é pré-requisito para adoção clínica responsável, posição também adotada no presente estudo. A coerência entre as variáveis selecionadas por consenso e as definições biomecânicas do GCS-CP é evidência concreta de que os modelos não operam como caixas pretas: para o alvo Transverso, seis rotações articulares (mínimos e máximos de quadril, joelho e tornozelo) figuram entre as onze variáveis unanimemente selecionadas pelos quatro métodos, correspondendo a 55% do consenso forte. A combinação de transparência metodológica, otimização orientada à sensibilidade patológica e supervisão especializada estabelece o fluxo responsável de implementação destes modelos em contexto clínico real.

5.12 Limitações

Os dados são provenientes de um único centro, o que pode limitar a generalização para outras populações ou equipamentos. O desbalanceamento de classes, particularmente para TE (7,5% no conjunto de teste) e BR (11,2% no conjunto de teste), representa desafio inerente que reflete a distribuição epidemiológica real destas condições. Conforme documentado por Prakash et al. (23), a maioria dos conjuntos de dados de marcha disponíveis apresenta vieses de gênero e faixa etária: o CASIA *Dataset* B (89) tem razão de 3:1 entre homens e mulheres, e o USF *dataset* (90) é enviesado para a faixa etária de 20-30 anos. No contexto da PC pediátrica, onde a distribuição por idade, sexo e nível funcional é heterogênea, esses vieses podem ser ainda mais pronunciados. Os autores destacam ainda que as características da marcha variam com fatores culturais e sociais, sugerindo cautela na extrapolação dos achados deste estudo, realizado exclusivamente com população brasileira, para outras populações.

Apesar da divisão rigorosa por paciente entre treino e teste, fontes residuais de correlação persistem entre registros do mesmo indivíduo: os parâmetros espaço-temporais do *Pipeline* Geral (velocidade, cadência, comprimento de passada) são globais ao *trial*, de modo que, na parcela de pacientes diplégicos com classificação GCS-CP simétrica entre lados (73,8% a 88,8% da amostra, conforme o alvo), os dois registros permanecem parcialmente correlacionados mesmo sob a Regra Ipsilateral. Na parcela complementar, diplégicos assimétricos (11,2% a 26,2% da amostra) e hemiplégicos (até 62,5% dos quais apresentaram classificação ipsilateral divergente no plano sagital de apoio), essa correlação é mitigada pela

própria divergência das classificações entre lados. Em qualquer cenário, a correlação residual não contamina as métricas reportadas, dado que o conjunto de teste foi segregado por paciente desde a divisão inicial.

Entre os modelos avaliados, alguns apresentaram *overfitting* médio elevado para Apoio Geral (0,383) e Balanço Geral (0,409), com diferença substancial entre acurácia de treino e de teste (Seção 5.7). Os modelos selecionados como melhores configurações de cada alvo apresentaram diferença treino-teste menor em Transverso (0,036) e Apoio (0,082). No alvo Balanço, contudo, mesmo o melhor modelo (*DTree* Hierárquico, diferença de 0,263) manteve diferença elevada, embora muito inferior à média do alvo, limitação que reflete a maior dificuldade intrínseca da classificação BN vs. BR no espaço cinemático disponível, conforme já discutido na Seção 5.5.2.

No tocante à incorporação de custos clínicos, a Estratégia E3 os tratou de forma simplificada. O retreinamento penalizou mais fortemente os erros sobre as classes patológicas por meio de um único parâmetro global (*boost_patologica*), sem distinguir, por exemplo, o impacto clínico de confundir SS com SN frente ao de confundir SS com SA. Uma versão mais refinada exigiria uma tabela detalhada com a gravidade de cada tipo específico de erro entre as classes, conhecida na literatura como matriz de custo, cuja construção dependeria de consenso entre especialistas em reabilitação pediátrica. Araf, Idri e Chairi (32) identificam justamente a dificuldade de quantificar essas gravidades de forma precisa como uma das limitações mais comuns nos estudos que aplicam aprendizado de máquina sensível ao custo a dados médicos. A formalização dessa matriz por consenso clínico é refinamento natural para estudos futuros.

A generalização dos resultados também é restringida pela ausência de validação externa em dados de outros laboratórios. Estudos multicêntricos com dados de centros independentes e protocolos heterogêneos de coleta são necessários para confirmar a generalização dos modelos aqui apresentados.

Por fim, a abordagem hierárquica testada utilizou estrutura fixa de duas etapas (triagem Normal vs. Alterado seguida de subtipagem), com pontos de corte de decisão padrão e sem otimização específica para minimizar a propagação de erros entre os níveis. Estratégias alternativas, como pontos de corte de confiança adaptativos por etapa ou modelos especializados para casos limítrofes na primeira camada, poderiam melhorar o desempenho da cascata e merecem investigação em estudos futuros.

5.13 Contribuições Metodológicas

A principal contribuição deste estudo foi demonstrar que métricas agregadas como a Acurácia Balanceada, embora adequadas para seleção e comparação de modelos em dados desbalanceados, são insuficientes para validação clínica de classificadores de marcha. Os modelos alcançaram Acurácia Balanceada de 0,702 a 0,791 no conjunto de teste independente, atendendo ao limiar de 0,70 adotado neste estudo nos três alvos, embora o atendimento marginal observado em Balanço (0,702) revele-se sensível à *seed* de inicialização e mereça interpretação cautelosa, conforme discutido adiante. Contudo, a decomposição sistemática por classe revelou heterogeneidade crítica: o melhor modelo de cada alvo apresentou sensibilidade de 100% para Transverso Externo, contra 56,6% para Sagital Salto e 50,6% para Balanço Rígido. Esta variação evidencia que cada modelo pode apresentar sensibilidade excelente para alguns padrões e clinicamente inadequada para outros, informação invisível em análises baseadas apenas em métricas agregadas. A análise por classe é, portanto, pré-requisito para implementação clínica responsável de classificadores automáticos de marcha (77,88).

A segunda contribuição metodológica central deste trabalho foi demonstrar que, em classificação automática da marcha na PC, nenhuma decisão de *pipeline* apresenta superioridade universal entre os três componentes do GCS-CP. O estudo identificou padrão alvo-dependente sistemático em cinco eixos metodológicos distintos:

(I) Seleção de variáveis (Consenso vs. PCA): PCA superior para Transverso, produzindo o melhor modelo do estudo (Acurácia Balanceada = 0,791), e também para Balanço na análise de estabilidade estocástica (Seção 3.9.4), com PCA-*LogReg* em 0,674 contra 0,619 do melhor "Com Consenso" (+5,5 pontos percentuais); "Com Consenso" claramente superior para Apoio (+8,9 pontos percentuais).

(II) Abordagem (*Flat* vs. Hierárquica em cascata): *Flat* superior para Transverso e Apoio; equivalência matemática em Balanço, por se tratar de problema binário.

(III) Balanceamento ("Com SMOTE" vs. "Sem SMOTE"): "Sem SMOTE" vencedor em Transverso e Balanço; "Com SMOTE" vantajoso apenas em Apoio (+0,053 em Acurácia Balanceada).

(IV) *Pipeline* de variáveis (Cinemática Angular vs. Cinemática Geral): Angular superior para Apoio e Balanço; Geral superior apenas para Transverso.

(V) Algoritmos individuais vs. modelos de fusão: modelos individuais na primeira posição nos três alvos.

Esta heterogeneidade sistemática indica que cada componente do GCS-CP possui estrutura biomecânica e estatística específica, que interage de forma própria com cada escolha de modelagem. O achado, até onde é de nosso conhecimento sem precedente direto na literatura de marcha em PC, contraria a tendência generalizada de aplicar receitas metodológicas únicas a problemas heterogêneos de classificação da marcha e estabelece a validação empírica por alvo como pré-requisito metodológico para o campo.

A terceira contribuição técnica foi o desenvolvimento de um *pipeline* completo com 1.152 modelos treinados, integrando os seguintes elementos:

- Divisão rigorosa por paciente: 86 pacientes em treino e 29 em teste, sem sobreposição.
- Regra Ipsilateral: redução das 50 variáveis originais (cinemáticas e espaço-temporais bilaterais, 36 angulares e 14 espaço-temporais) para 18 (*Pipeline* Cinemática Angular) ou 25 (*Pipeline* Cinemática Geral), eliminando contaminação cruzada entre lados.
- Estratégias de redução dimensional: seleção por consenso entre quatro métodos (limiar de variância, informação mútua, F-score ANOVA e importância por Floresta Aleatória) ou Análise de Componentes Principais (PCA), conforme a variante avaliada.
- Otimização bayesiana de hiperparâmetros: Optuna com 50 tentativas por modelo e poda antecipada via HyperbandPruner.
- Avaliação por múltiplas métricas complementares: Acurácia Balanceada, *G-Mean*, MCC, Kappa e AUC.

A escolha por reportar simultaneamente Acurácia Balanceada, Média Geométrica (G-Mean) e Coeficiente de Correlação de Matthews mostrou-se particularmente informativa em cenários de desbalanceamento severo, expondo casos em que métricas tradicionais como acurácia simples teriam mascarado desempenho próximo ao acaso (Seção 5.6). Essas três métricas integraram um conjunto de avaliação mais amplo, que reuniu ainda AUC-ROC (macro *one-vs-rest*), Kappa de Cohen, Erro de Calibração Esperado (ECE), sensibilidade e especificidade macro por classe, uma métrica explícita de *overfitting* (diferença entre treino e teste) e intervalos de confiança estimados por *bootstrap* (1.000 reamostragens). A adoção sistemática desse conjunto multidimensional, ancorado na tríade Acurácia Balanceada–G-Mean–MCC e complementado pela análise por classe e pela quantificação da incerteza, é recomendada como prática metodológica para futuros estudos de classificação de marcha com dados desbalanceados.

A coerência clínica entre as variáveis selecionadas e as definições biomecânicas do GCS-CP é particularmente notável para o alvo Transverso: a estratégia "Com Consenso", no *Pipeline* Cinemática Geral, selecionou as seis variáveis de rotação articular

disponíveis (mínimos e máximos de quadril, joelho e tornozelo) com votação unânime entre os quatro métodos, correspondendo a 46% das 13 variáveis finais selecionadas para esse alvo. Esse alinhamento espontâneo entre a seleção estatística e a definição biomecânica do componente Transverso, que se refere ao plano horizontal de rotação, reforça a interpretação dos modelos e a validade do processo de seleção.

A quarta contribuição foi a demonstração empírica de que a abordagem hierárquica em cascata, embora conceitualmente alinhada ao raciocínio clínico de triagem seguida de subtipagem, não apresentou superioridade sobre a classificação multiclasse direta para dois dos três alvos avaliados. A propagação de erros entre etapas penalizou sistematicamente as classes patológicas, com perda de 42,0 pontos percentuais na sensibilidade de Transverso Externo no alvo Transverso e de 20,2 pontos percentuais em Sagital Agachado no alvo Apoio (Seção 4.7).

A aparente superioridade da abordagem Hierárquica para o alvo Balanço foi quantitativamente refutada por análise complementar com múltiplas *seeds* (Seção 3.9.4). A reexecução com a *seed* original (42) confirmou reprodutibilidade da Acurácia Balanceada de 0,702; entretanto, em três *seeds* adicionais (1, 7, 123), o desempenho da Árvore de Decisão variou substancialmente (0,580 ± 0,111), com distribuição bimodal (2 *seeds* em 0,702 e 3 *seeds* em 0,499). A Regressão Logística, em contraste, apresentou Acurácia Balanceada de 0,618 ± 0,000 em todas as *seeds*. O resultado de Acurácia Balanceada = 0,702 obtido com a *seed* 42, portanto, não reflete superioridade real da abordagem Hierárquica, mas variabilidade estocástica do Optuna na otimização da Árvore de Decisão. Este achado tem implicação metodológica direta para futuros estudos que considerem estruturas hierárquicas de classificação da marcha.

Por fim, a quinta contribuição deste estudo foi a demonstração de que estratégias de otimização clínica orientadas à minimização de falsos negativos podem transformar substancialmente o potencial de triagem dos modelos, com eficácia heterogênea entre as classes patológicas. A Estratégia E1 eliminou completamente os falsos negativos de TE (de 33 para zero) pela seleção do *pipeline* de maior sensibilidade mínima patológica, sem nenhum ajuste de ponto de corte, configurando o resultado de maior impacto clínico entre todas as estratégias avaliadas e o único caso de compromisso plenamente favorável. A Estratégia E3 eliminou completamente os falsos negativos para SA e reduziu em 60% os de SS, mas com queda substancial da Acurácia Balanceada global do alvo Apoio (de 0,717 para 0,434). Essa queda não reflete falha do método, e sim uma mudança no papel do modelo: ele deixou de operar como classificador multiclasse e passou a funcionar como filtro de triagem ultra-

sensível, marcando quase todos os casos como potencialmente patológicos (apenas 7 de 202 casos de SN foram mantidos como normais). Trata-se, portanto, de um ganho condicionado ao papel atribuído ao modelo: esse perfil é inadequado para uso autônomo em diagnóstico, mas é justamente o comportamento desejável de uma primeira camada de triagem em uma cascata clínica, na qual a etapa subsequente (humana ou algorítmica) confirma a presença e o tipo específico de padrão patológico. Para BR e TI, a redução de falsos negativos foi acompanhada por colapso de especificidade, indicando limite intrínseco do espaço cinemático disponível. A combinação destes resultados reposiciona os modelos desenvolvidos: de "ferramentas com desempenho agregado moderado" para "ferramentas com potencial clínico diferenciado por classe". Esse reposicionamento exige duas condições: que a função objetivo esteja alinhada às prioridades clínicas de triagem, e que o cenário de uso reconheça os custos específicos por classe.

5.14 Direções para Trabalhos Futuros

Os resultados deste estudo apontam direções para o avanço da classificação automática de padrões da marcha em paralisia cerebral. Papageorgiou et al. (42), em estudo recente de validação dos padrões sagitais, recomendaram explicitamente o emprego de técnicas de aprendizado de máquina em amostras extensas como próximo passo para a definição automatizada de limiares, para o desenvolvimento de algoritmos de classificação automática e para a exploração de abordagens flexíveis. As direções elencadas a seguir articulam-se diretamente com essa agenda, posicionando esta dissertação como aplicação concreta dessa proposta.

A primeira direção é a extensão da Estratégia E3 (retreinamento orientado à sensibilidade) a todos os algoritmos avaliados, hoje aplicada apenas ao Floresta Aleatória, combinada com ajuste subsequente de ponto de corte (E2). A proposta encontra respaldo no princípio demonstrado por Mohammadi-Seif e Baeza-Yates (33): a ponderação assimétrica de custos clínicos produziu ganhos consistentes em quatro modalidades de imagem médica e mostrou-se robusta entre arquiteturas distintas de aprendizado profundo (redes convolucionais e Transformers), sugerindo que o benefício independe do domínio ou do tipo de modelo específico. As classes Balanço Rígido e Transverso Interno devem permanecer, no entanto, como desafios estruturais que dificilmente serão superados apenas por refinamento algorítmico sobre variáveis cinemáticas, conforme discutido nas Seções 5.5.1 e 5.10.

A segunda direção é o desenvolvimento de arquiteturas especializadas por classe. Os modelos avaliados aprenderam a distinguir todas as classes de cada alvo simultaneamente; uma alternativa promissora é treinar modelos especializados em uma única classe difícil, com o objetivo de maximizar especificamente seu desempenho. Duas estratégias se destacam: a abordagem um-contra-todos, em que um modelo é treinado para distinguir apenas Sagital Salto (ou Balanço Rígido) das demais classes; e a votação por especialistas, em que múltiplos modelos especializados, cada um responsável por uma classe, são combinados na decisão final. A abordagem é particularmente promissora para Sagital Salto e Balanço Rígido, classes que apresentaram as menores sensibilidades no presente estudo.

A terceira direção é a incorporação de variáveis cinéticas, especialmente para o alvo Balanço. Conforme discutido na Seção 5.5.1 a partir dos achados de Goldberg et al. (79,80), a sensibilidade limitada para Balanço Rígido reflete um teto intrínseco da classificação baseada apenas em ângulos articulares, pois o padrão rígido associa-se à hiperatividade do reto femoral durante o duplo apoio, mecanismo que se traduz mais diretamente em padrão cinético (momentos articulares e força de reação do solo) do que em padrão cinemático. A integração de dados cinéticos é factível em laboratórios equipados para análise instrumentada da marcha. A eletromiografia, embora capture diretamente o mecanismo gerador, apresenta limitações práticas relevantes em populações pediátricas que tornam sua incorporação rotineira pouco viável a curto prazo.

A quarta direção é o emprego de aprendizado profundo sobre as curvas cinemáticas completas, em vez dos descritores resumidos utilizados neste estudo. Cada curva descreve como o ângulo articular evolui ao longo do ciclo da marcha, e os descritores extraídos aqui (mínimos, máximos, amplitudes) preservam apenas pontos isolados dessa trajetória. Informações como o instante em que o pico ocorre, a largura do platô em que ele se mantém ou a simetria do padrão antes e depois do pico podem ser discriminativas entre classes patológicas e ficam disponíveis quando a curva inteira é utilizada como entrada. Arquiteturas de aprendizado profundo (como redes convolucionais e redes recorrentes) operam diretamente sobre as curvas; alternativamente, técnicas como a PCA funcional podem extrair representações compactas que preservam a forma temporal e, em seguida, alimentar algoritmos clássicos. O presente estudo serve como linha de base para essa agenda: abordagens mais complexas precisarão demonstrar ganho clínico real, e não apenas melhoria em métricas agregadas.

A quinta direção é a parametrização separada por fase do ciclo. Padrões patológicos do GCS-CP não se distribuem uniformemente entre apoio e balanço: Sagital Agachado manifesta-se predominantemente no apoio, enquanto Balanço Rígido, no balanço.

Os atributos calculados sobre o ciclo inteiro mascaram essa especificidade fásica e podem obscurecer alterações sutis em fases biomecanicamente críticas. A proposta é calcular conjuntos de descritores separadamente para cada fase, dobrando o número de variáveis disponíveis e aumentando o foco biomecânico de cada uma. A implementação requer identificação dos eventos de contato inicial e desprendimento dos dedos de cada paciente, etapa viável pelas plataformas de força ou por algoritmos cinemáticos. Permanece o desafio dos participantes que utilizam dispositivos auxiliares de marcha (frequentes no nível GMFCS III), nos quais a coleta cinética é tipicamente inviável, conforme descrito na Seção 3.2.

A sexta direção articula análise estatística e conhecimento clínico por meio da seleção de variáveis guiada por especialistas. Neste estudo, a seleção foi exclusivamente algorítmica, abordagem objetiva e replicável, mas que enxerga apenas relações numéricas, sem filtro de plausibilidade biomecânica. Envolver especialistas em análise da marcha antes da etapa de seleção automática, definindo, em conjunto com a equipe técnica, quais variáveis fazem sentido biomecânico para cada padrão, tenderia a produzir modelos com maior interpretação e menor risco de capturar correlações espúrias presentes nos dados sem base biomecânica.

A sétima direção é a transição da pesquisa para a prática, envolvendo dois componentes complementares. O primeiro é o desenvolvimento de uma interface clínica interpretável, integrada ao fluxo de trabalho do laboratório de marcha, que combine indicador de confiança por classe (probabilidade estimada para cada padrão, e não apenas o rótulo único), visualização dos atributos mais influentes em cada classificação individual e alerta automático para casos limítrofes que mereçam segunda opinião especializada. A interface deve ser concebida como ferramenta de apoio à decisão, oferecendo uma segunda perspectiva objetiva e dados de confiança que hoje não existem na avaliação visual. O segundo é a ampliação da amostra por colaboração interinstitucional, combinando dados de múltiplos laboratórios brasileiros e internacionais, com ênfase na representação das classes hoje minoritárias (Balanço Rígido e Transverso Externo). O ganho é duplo: aumento do número de exemplos das classes hoje sub-representadas e maior heterogeneidade de protocolos, equipamentos e perfis demográficos, condição essencial para que modelos treinados generalizem além do laboratório de desenvolvimento.

6 CONCLUSÃO

Este estudo investigou a aplicação de técnicas de aprendizado de máquina à classificação automática de padrões da marcha em crianças e adolescentes com paralisia cerebral espástica, ancorado no Sistema de Classificação da Marcha para Crianças com Paralisia Cerebral (GCS-CP). Os cinco objetivos específicos foram atingidos, conforme detalhado no Capítulo 4 e analisado nas Seções 5.2 a 5.10. Os achados demonstram que classificadores supervisionados são viáveis para os três componentes do sistema avaliado, embora com desempenho heterogêneo entre alvos e entre classes patológicas, evidência que justifica recomendações metodológicas específicas para implementação clínica.

A principal contribuição deste trabalho é demonstrar empiricamente que métricas agregadas, como Acurácia Balanceada ou *F1-score* médio, são insuficientes para avaliação clínica de classificadores de marcha em PC. A análise por classe revelou padrões de falsos negativos clinicamente críticos que permaneciam invisíveis no ranqueamento agregado, e estratégias de otimização orientadas à sensibilidade patológica mínima reposicionaram os modelos como ferramentas com potencial clínico diferenciado por classe. Confirmou-se, portanto, a necessidade de reportar sensibilidade e especificidade por classe em estudos desta natureza.

As comparações metodológicas conduzidas evidenciaram que não há superioridade universal entre estratégias de pré-processamento, entre algoritmos individuais e modelos de fusão, ou entre classificação multiclasse direta e em cascata: o desempenho relativo é alvo-dependente e exige validação empírica para cada plano biomecânico. Esse achado contradiz recomendações genéricas presentes na literatura e sustenta a necessidade de protocolos comparativos sistemáticos em estudos futuros.

Do ponto de vista da aplicabilidade clínica, apenas a classe Transverso Externo atingiu compromisso plenamente favorável, com eliminação completa dos falsos negativos pela Estratégia E1 (seleção de *pipeline* orientada à sensibilidade mínima patológica), sem custo de especificidade; as demais classes patológicas permanecem como desafios remanescentes. Esta limitação aponta para a necessidade de incorporação de variáveis complementares, particularmente cinéticas, e de métodos de otimização clínica mais sofisticados em trabalhos subsequentes.

O código-fonte do *pipeline* de análise está disponível mediante solicitação ao autor. A base de dados cinemática, por conter informações potencialmente identificáveis de

pacientes pediátricos, não é disponibilizada publicamente, em conformidade com o Termo de Consentimento Livre e Esclarecido aprovado pelo Comitê de Ética em Pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Graham HK, Rosenbaum P, Paneth N, Dan B, Lin JP, Damiano DL, et al. Cerebral palsy. Nat Rev Dis Primers. 2016;2:15082. doi: 10.1038/nrdp.2015.82

2. McIntyre S, Goldsmith S, Webb A, Ehlinger V, Hollung SJ, McConnell K, et al. Global prevalence of cerebral palsy: a systematic analysis. Dev Med Child Neurol. 2022;64(12):1494-1506. doi: 10.1111/dmcn.15346

3. Oskoui M, Coutinho F, Dykeman J, Jetté N, Pringsheim T. An update on the prevalence of cerebral palsy: a systematic review and meta-analysis. Dev Med Child Neurol. 2013;55(6):509-519. doi: 10.1111/dmcn.12080

4. Sellier E, Platt MJ, Andersen GL, Krägeloh-Mann I, De La Cruz J, Cans C, et al. Decreasing prevalence in cerebral palsy: a multi-site European population-based study, 1980 to 2003. Dev Med Child Neurol. 2016;58(1):85-92. doi: 10.1111/dmcn.12865

5. Galea C, McIntyre S, Smithers-Sheedy H, Reid SM, Gibson C, Delacy M, et al. Cerebral palsy trends in Australia (1995-2009): a population-based observational study. Dev Med Child Neurol. 2019;61(2):186-193. doi: 10.1111/dmcn.14011

6. Dan B, Rosenbaum P, Carr L, Gough M, Coughlan J, Nweke N. Updated description of cerebral palsy. Dev Med Child Neurol. 2026;68(4):465-476. doi: 10.1111/dmcn.70149

7. Zhou J, Butler EE, Rose J. Neurologic correlates of gait abnormalities in cerebral palsy: implications for treatment. Front Hum Neurosci. 2017;11:103. doi: 10.3389/fnhum.2017.00103

8. Papageorgiou E, De Beukelaer N, Simon-Martinez C, Mailleux L, Van Campenhout A, Desloovere K, et al. Structural brain lesions and gait pathology in children with spastic cerebral palsy. Front Hum Neurosci. 2020;14:275. doi: 10.3389/fnhum.2020.00275

9. Davids JR, Bagley AM. Identification of common gait disruption patterns in children with cerebral palsy. J Am Acad Orthop Surg. 2014;22(12):782-790. doi: 10.5435/JAAOS-22-12-782

10. Rodda J, Graham HK. Classification of gait patterns in spastic hemiplegia and spastic diplegia: a basis for a management algorithm. Eur J Neurol. 2001;8(Suppl 5):98-108. doi: 10.1046/j.1468-1331.2001.00042.x

11. Wren TAL, Tucker CA, Rethlefsen SA, Gorton GE, Õunpuu S. Clinical efficacy of instrumented gait analysis: systematic review 2020 update. Gait Posture. 2020;80:274-279. doi: 10.1016/j.gaitpost.2020.05.031

12. Chang FM, Rhodes JT, Flynn KM, Carollo JJ. The role of gait analysis in treating gait abnormalities in cerebral palsy. Orthop Clin North Am. 2010;41(4):489-506. doi: 10.1016/j.ocl.2010.06.009

13. Schwartz MH, Ries AJ, Georgiadis AG, Kainz H. Demonstrating the utility of instrumented gait analysis in the treatment of children with cerebral palsy. PLoS One. 2024;19(4):e0301230. doi: 10.1371/journal.pone.0301230

14. Melanda AG, Davids JR, Pauleto AC, Pelegrinelli ARM, Ferreira AEK, Knaut LA, et al. Reliability and validity of the gait classification system in children with cerebral palsy (GCS-CP). Gait Posture. 2022;98:355-361. doi: 10.1016/j.gaitpost.2022.09.083

15. Slijepcevic D, Zeppelzauer M, Unglaube F, Kranzl A, Breiteneder C, Horsak B. Explainable machine learning in human gait analysis: a study on children with cerebral palsy. IEEE Access. 2023;11:65906-65923. doi: 10.1109/ACCESS.2023.3289986

16. Lai DTH, Begg RK, Palaniswami M. Computational intelligence in gait research: a perspective on current applications and future challenges. IEEE Trans Inf Technol Biomed. 2009;13(5):687-702. doi: 10.1109/TITB.2009.2022913

17. Zhang Y, Ma Y. Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia. Comput Biol Med. 2019;106:33-39. doi: 10.1016/j.compbiomed.2019.01.009

18. Franco A, Russo M, Amboni M, Ponsiglione AM, Di Filippo F, Romano M, et al. The role of deep learning and gait analysis in Parkinson's disease: a systematic review. Sensors (Basel). 2024;24(18):5957. doi: 10.3390/s24185957

19. Kamruzzaman J, Begg RK. Support vector machines and other pattern recognition approaches to the diagnosis of cerebral palsy gait. IEEE Trans Biomed Eng. 2006;53(12):2479-2490. doi: 10.1109/TBME.2006.883697

20. Darbandi H, Baniasad M, Baghdadi S, Khandan A, Vafaee A, Farahmand F. Automatic classification of gait patterns in children with cerebral palsy using fuzzy clustering method. Clin Biomech (Bristol, Avon). 2020;73:189-194. doi: 10.1016/j.clinbiomech.2019.12.031

21. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. p. 3146-3154.

22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco (CA): ACM; 2016. p. 785-794. doi: 10.1145/2939672.2939785

23. Prakash C, Kumar R, Mittal N. Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. Artif Intell Rev. 2018;49(1):1-40. doi: 10.1007/s10462-016-9514-6

24. Samadi Kohnehshahri F, Merlo A, Mazzoli D, Bò MC, Stagni R. Machine learning applied to gait analysis data in cerebral palsy and stroke: a systematic review. Gait Posture. 2024;111:105-121. doi: 10.1016/j.gaitpost.2024.04.007

25. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Daelemans W, Goethals B, Morik K, editors. Machine Learning and Knowledge Discovery in Databases. Berlin: Springer; 2008. p. 313-325. doi: 10.1007/978-3-540-87481-2_21

26. Phinyomark A, Petri G, Ibáñez-Marcelo E, Osis ST, Ferber R. Analysis of big data in gait biomechanics: current trends and future directions. J Med Biol Eng. 2018;38(2):244-260. doi: 10.1007/s40846-017-0297-2

27. Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier Systems. Berlin: Springer; 2000. p. 1-15. doi: 10.1007/3-540-45014-9_1

28. Zhou ZH. Ensemble methods: foundations and algorithms. Boca Raton (FL): CRC Press; 2012.

29. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. Istanbul: IEEE; 2010. p. 3121-3124. doi: 10.1109/ICPR.2010.764

30. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv:2008.05756 [Preprint]. 2020. doi: 10.48550/arXiv.2008.05756

31. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427-437. doi: 10.1016/j.ipm.2009.03.002

32. Araf I, Idri A, Chairi I. Cost-sensitive learning for imbalanced medical data: a review. Artif Intell Rev. 2024;57(4):80. doi: 10.1007/s10462-023-10652-8

33. Mohammadi-Seif A, Baeza-Yates R. Risk-calibrated learning: minimizing fatal errors in medical AI. arXiv:2604.12693 [Preprint]. 2026. doi: 10.48550/arXiv.2604.12693

34. Silla CN, Freitas AA. A survey of hierarchical classification across different application domains. Data Min Knowl Discov. 2011;22(1-2):31-72. doi: 10.1007/s10618-010-0175-9

35. McIntyre S, Taitz D, Keogh J, Goldsmith S, Badawi N, Blair E. A systematic review of risk factors for cerebral palsy in children born at term in developed countries. Dev Med Child Neurol. 2013;55(6):499-508. doi: 10.1111/dmcn.12017

36. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol. 1997;39(4):214-223. doi: 10.1111/j.1469-8749.1997.tb07414.x

37. Palisano RJ, Rosenbaum P, Bartlett D, Livingston MH. Content validity of the expanded and revised Gross Motor Function Classification System. Dev Med Child Neurol. 2008;50(10):744-750. doi: 10.1111/j.1469-8749.2008.03089.x

38. Fowler EG, Goldberg EJ. The effect of lower extremity selective voluntary motor control on interjoint coordination during gait in children with spastic diplegic cerebral palsy. Gait Posture. 2009;29(1):102-107. doi: 10.1016/j.gaitpost.2008.07.007

39. Chruscikowski E, Fry NRD, Noble JJ, Gough M, Shortland AP. Selective motor control correlates with gait abnormality in children with cerebral palsy. Gait Posture. 2017;52:107-109. doi: 10.1016/j.gaitpost.2016.11.031

40. Winters TF Jr, Gage JR, Hicks R. Gait patterns in spastic hemiplegia in children and young adults. J Bone Joint Surg Am. 1987;69(3):437-441.

41. Rodda JM, Graham HK, Carson L, Galea MP, Wolfe R. Sagittal gait patterns in spastic diplegia. J Bone Joint Surg Br. 2004;86(2):251-258. doi: 10.1302/0301-620X.86B2.13878

42. Papageorgiou E, Everaert L, Vandekerckhove I, Hanssen B, Van Campenhout A, Ortibus E, et al. The Gait Pattern Classification System for Children with Spastic Cerebral Palsy (GaP-CP): a validity study. Children (Basel). 2025;12(3):269. doi: 10.3390/children12030269

43. Morais Filho MC, Kawamura CM, Fujino MH, Przysiada L, Antunes Silva ME, Lopes JAF. Gait patterns in hemiplegic cerebral palsy: is it time for a new classification? SSRN [Preprint]. 2024. doi: 10.2139/ssrn.4984616

44. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46. doi: 10.1177/001316446002000104

45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. doi: 10.2307/2529310

46. Dobson F, Morris ME, Baker R, Graham HK. Gait classification in children with cerebral palsy: a systematic review. Gait Posture. 2007;25(1):140-152. doi: 10.1016/j.gaitpost.2006.01.003

47. Yoon C, Jeon Y, Choi H, Kwon SS, Ahn J. Interpretable classification for multivariate gait analysis of cerebral palsy. Biomed Eng Online. 2023;22(1):109. doi: 10.1186/s12938-023-01168-x

48. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell. 1998;20(3):226-239. doi: 10.1109/34.667881

49. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241-259. doi: 10.1016/S0893-6080(05)80023-1

50. Choisne J, Fourrier N, Handsfield G, Signal N, Taylor D, Wilson N, et al. An unsupervised data-driven model to classify gait patterns in children with cerebral palsy. J Clin Med. 2020;9(5):1432. doi: 10.3390/jcm9051432

51. Roaas A, Andersson GBJ. Normal range of motion of the hip, knee and ankle joints in male subjects, 30-40 years of age. Acta Orthop Scand. 1982;53(2):205-208. doi: 10.3109/17453678208992202

52. Hemmerich A, Brown H, Smith S, Marthandam SSK, Wyss UP. Hip, knee, and ankle kinematics of high range of motion activities of daily living. J Orthop Res. 2006;24(4):770-781. doi: 10.1002/jor.20114

53. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.

54. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken (NJ): Wiley-Interscience; 2006.

55. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi: 10.1023/A:1010933404324

56. Abdi H, Williams LJ. Principal component analysis. WIREs Comput Stat. 2010;2(4):433-459. doi: 10.1002/wics.101

57. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. arXiv:1907.10902 [Preprint]. 2019. doi: 10.48550/arXiv.1907.10902

58. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: John Wiley & Sons; 2000.

59. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297. doi: 10.1007/BF00994018

60. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont (CA): Wadsworth; 1984.

61. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21-27. doi: 10.1109/TIT.1967.1053964

62. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: John Wiley & Sons; 2001.

63. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-536. doi: 10.1038/323533a0

64. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. doi: 10.1613/jair.953

65. Krogh A, Hertz JA. A simple weight decay can improve generalization. In: Advances in Neural Information Processing Systems 4 (NIPS 1991). 1992. p. 950-957.

66. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929-1958.

67. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit. 2012;45(1):521-530. doi: 10.1016/j.patcog.2011.06.019

68. Fernández A, García S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863-905. doi: 10.1613/jair.1.11192

69. Kelleher JD, Mac Namee B, D'Arcy A. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Cambridge (MA): MIT Press; 2015.

70. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall; 1993. doi: 10.1201/9780429246593

71. Chong SF, Choo R. Introduction to bootstrap. Proc Singap Healthc. 2011;20(3):236-240.

72. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning (ICML '05). Bonn: ACM; 2005. p. 625-632. doi: 10.1145/1102351.1102430

73. Hoang N, Liang Z. Decision threshold optimization in machine learning models for questionnaire-based OSA screening. In: Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies. Marbella: SCITEPRESS; 2026. p. 318-329. doi: 10.5220/0014466500004070

74. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123-140. doi: 10.1007/BF00058655

75. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30.

76. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202. doi: 10.1098/rsta.2015.0202

77. Sheldrick RC, Benneyan JC, Kiss IG, Briggs-Gowan MJ, Copeland W, Carter AS. Thresholds and accuracy in screening tools for early detection of psychopathology. J Child Psychol Psychiatry. 2015;56(9):936-948. doi: 10.1111/jcpp.12442

78. Glascoe FP. Screening for developmental and behavioral problems. Ment Retard Dev Disabil Res Rev. 2005;11(3):173-179. doi: 10.1002/mrdd.20068

79. Goldberg SR, Õunpuu S, Arnold AS, Gage JR, Delp SL. Kinematic and kinetic factors that correlate with improved knee flexion following treatment for stiff-knee gait. J Biomech. 2006;39(4):689-698. doi: 10.1016/j.jbiomech.2005.01.015

80. Goldberg SR, Anderson FC, Pandy MG, Delp SL. Muscles that influence knee flexion velocity in double support: implications for stiff-knee gait. J Biomech. 2004;37(8):1189-1196. doi: 10.1016/j.jbiomech.2003.12.005

81. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.

82. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012;55(10):78-87. doi: 10.1145/2347736.2347755

83. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn. 1997;29(2-3):103-130. doi: 10.1023/A:1007413511361

84. Heaton J. Review of: Goodfellow I, Bengio Y, Courville A. Deep learning. Genet Program Evolvable Mach. 2018;19(1-2):305-307. doi: 10.1007/s10710-017-9314-z

85. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is “nearest neighbor” meaningful? In: Beeri C, Buneman P, editors. Database Theory - ICDT'99. Berlin: Springer; 1999. p. 217-235. doi: 10.1007/3-540-49257-7_15

86. Plante E, Vance R. Selection of preschool language tests: a data-based approach. Lang Speech Hear Serv Sch. 1994;25(1):15-24. doi: 10.1044/0161-1461.2501.15

87. Nahar A, Paul S, Saikia MJ. A systematic review on machine learning approaches in cerebral palsy research. PeerJ. 2024;12:e18270. doi: 10.7717/peerj.18270

88. Institute of Automation, Chinese Academy of Sciences. CASIA gait dataset [Internet]. [cited 2026 Feb 6]. Available from: http://www.cbsr.ia.ac.cn

89. Sarkar S, Phillips PJ, Liu Z, Vega IR, Grother P, Bowyer KW. The humanID gait challenge problem: data sets, performance, and analysis. IEEE Trans Pattern Anal Mach Intell. 2005;27(2):162-177. doi: 10.1109/TPAMI.2005.39

APÊNDICE A – VARIÁVEIS REGRA IPSILATERAL

Cada registro utiliza exclusivamente as variáveis do membro avaliado naquele ciclo (ipsilateral), sem cruzamento com o membro contralateral.

PIPELINE 1 — CINEMÁTICA ANGULAR (18 *features*)

Quadril ipsilateral (HIP) — n=6

1. Min_HIP_Rot — rotação mínima do quadril
2. Max_HIP_Rot — rotação máxima do quadril
3. Min_HIP_Abd — abdução mínima do quadril
4. Max_HIP_Abd — abdução máxima do quadril
5. Min_HIP_Flex — flexão mínima do quadril
6. Max_HIP_Flex — flexão máxima do quadril

Joelho ipsilateral (KNEE) — n=6

7. Min_KNEE_Rot — rotação mínima do joelho
8. Max_KNEE_Rot — rotação máxima do joelho
9. Min_KNEE_Abd — abdução mínima do joelho
10. Max_KNEE_Abd — abdução máxima do joelho
11. Min_KNEE_Flex — flexão mínima do joelho
12. Max_KNEE_Flex — flexão máxima do joelho

Tornozelo ipsilateral (ANK) — n=6

13. Min_ANK_Rot — rotação mínima do tornozelo
14. Max_ANK_Rot — rotação máxima do tornozelo
15. Min_ANK_Abd — abdução mínima do tornozelo
16. Max_ANK_Abd — abdução máxima do tornozelo
17. Min_ANK_Flex — flexão mínima do tornozelo
18. Max_ANK_Flex — flexão máxima do tornozelo

TOTAL *PIPELINE* ANGULAR: 18 variáveis

PIPELINE 2 — CINEMÁTICA GERAL (25 *features*)

Inclui as 18 do *Pipeline* Angular mais 7 parâmetros espaço-temporais ipsilaterais:

Vel — velocidade de marcha

Stride — comprimento da passada

Cad — cadência

Support_Time — tempo de apoio

Non_Support — tempo sem apoio (balanço)

Step_Len — comprimento do passo

Dbl_Support — tempo de duplo suporte

TOTAL *PIPELINE* GERAL: 25 variáveis (18 angulares + 7 espaço-temporais)

Nota: sobre a Regra Ipsilateral: Do conjunto de 70 variáveis disponíveis na base de dados original, foram selecionadas para o *pipeline* de modelagem 36 variáveis cinemáticas angulares bilaterais (18 por lado, distribuídas entre quadril, joelho e tornozelo nos planos sagital, coronal e transverso) e 14 parâmetros espaço-temporais bilaterais (7 por lado), totalizando 50 variáveis. As demais variáveis da base original não foram utilizadas neste estudo. Após a aplicação da Regra Ipsilateral, cada registro retém apenas as variáveis do membro correspondente ao ciclo avaliado, direito para ciclos do lado direito, esquerdo para ciclos do lado esquerdo, resultando em dois *pipelines* alternativos: Cinemática Angular, com 18 variáveis angulares ipsilaterais, e Cinemática Geral, com 25 variáveis (as 18 angulares acrescidas dos 7 parâmetros espaço-temporais ipsilaterais). Esta abordagem elimina a contaminação cruzada entre lados e é metodologicamente consistente com a classificação individual por membro do GCS-CP.
Fonte: Elaborado pelo autor com base nos dados extraídos do sistema Cortex e OrthoTrack 6.5.1, após aplicação da Regra Ipsilateral.