Comparando Tipos de Representações Baseadas em Vídeo na Classificação de Movimentos Gerais de Bebês Recém-Nascidos

Beatriz Emily Silva Aguiar[1], **Eulanda M. Santos**[1], **Rafael Giusti**[1], **Ayrles Mendonça**[2]

[1]Instituto de Computação (IComp) – Universidade Federal do Amazonas (UFAM)

[2]Faculdade de Educação Física e Fisioterapia (FEFF) – UFAM
Av. Gen. Rodrigo Octávio, 6200, Coroado I
Campus Universitário – 69080-900 – Manaus – AM – Brasil

{beatriz.aguiar,emsantos,rgiusti}@icomp.ufam.edu.br, ayrles@ufam.edu.br

Abstract. *General Movements Assessment (GMA) is a clinical method for early detection of neuromotor disorders in infants. In this study we compare video-based representations consolidated in the literature applied to classify Writhing Movements (WMs): RGB videos, optical flow, keypoints, and motion histograms. In our experiments conducted using a proprietary dataset, the RGB-based model achieved the highest individual accuracy (0.83), while the model trained using histograms obtained the best F1-Score (0.83) and efficiency. Finally, the fusion of the models via ensemble increased accuracy to 0.87. These results indicate that integrating multiple representations can enhance automated GMA analysis and contribute to the early diagnosis of neuromotor disorders.*

Resumo. *A Avaliação dos Movimentos Gerais (GMA) é um método clínico para detecção precoce de distúrbios neuromotores em bebês. Este estudo compara abordagens de representação de vídeo para a classificação de Writhing Movements (WMs) consolidadas na literatura: vídeos RGB, fluxo óptico, pontos-chave e histogramas de movimento. Nos experimentos executados em uma base de dados própria, o modelo treinado com vídeos RGB obteve maior acurácia individual (0,83), enquanto o modelo treinado com histogramas alcançou melhor F1-Score (0,83) e maior eficiência. Por fim, a fusão dos modelos via comitê elevou a acurácia para 0,87. Os resultados indicam que integrar múltiplas representações pode aprimorar a análise automatizada dos GMs e contribuir para o diagnóstico precoce de distúrbios neuromotores.*

1. Introdução

A paralisia cerebral (PC) é o problema motor mais comum na infância, com prevalência estimada entre 1,5 a 2,5 por 1.000 nascidos vivos em países industrializados [Camargos and Almeida 2019]. A prematuridade é um fator de risco significativo para o desenvolvimento de PC, afetando principalmente bebês nascidos antes das 32 semanas de gestação [Spittle and Doyle 2018]. Os sintomas da PC variam desde uma leve falta de coordenação motora até dificuldades significativas em movimentar membros, podendo incluir paralisia e rigidez articular severa.

A Avaliação dos Movimentos Gerais (*General Movements Assessment* - GMA), desenvolvida por Prechtl, é uma ferramenta não invasiva e de baixo custo que permite a identificação precoce de alterações neurológicas em bebês [Prechtl 1997]. Inicialmente utilizado como preditor de PC, a GMA tem sido utilizada como preditor de diversos distúrbios genéticos e de neurodesenvolvimento [Silva et al. 2021]. Os movimentos gerais (GMs) são espontâneos e

complexos, presentes desde o início da vida fetal até o final do primeiro semestre de vida. Eles podem ser observados em duas fases principais: a fase dos *Writhing Movements* (WMs) e a fase dos *Fidgety Movements* (FMs).

Os WMs ocorrem desde o período fetal até cerca de 6 a 9 semanas após o nascimento. Eles são caracterizados por movimentos amplos, lentos e contínuos, que envolvem todo o corpo, apresentando variações em intensidade, velocidade e amplitude. Esses movimentos têm início e término graduais. A presença de WMs típicos é indicativa de um desenvolvimento neurológico saudável, enquanto padrões atípicos, como movimentos rígidos ou sincrônicos, podem indicar alterações neurológicas [Ferrari and Cioni 2016]. Já os FMs surgem entre 6 e 9 semanas e permanecem até cerca de 20 semanas. Eles consistem em movimentos contínuos, de pequena amplitude e velocidade moderada, que ocorrem em todas as direções, envolvendo o pescoço, tronco e membros. Esses movimentos são mais evidentes quando o bebê está em um estado de alerta calmo [Hadders-Algra 2004]. Durante essa fase, padrões atípicos podem ser observados e estão associados a um risco aumentado de PC.

A avaliação manual dos GMs, embora eficaz, é subjetiva e depende da experiência do avaliador. Portanto, há um crescente interesse no desenvolvimento de métodos automatizados para a classificação de GMs via métodos de aprendizado de máquina. Nesse contexto, abordagens baseadas na classificação de GMs a partir de vídeo são preferíveis devido a pelo menos duas razões: 1) a GMA é originalmente um método baseado em avaliação visual; e 2) é uma estratégia que se alinha bem ao caráter não intrusivo da GMA clássica, pois possibilita a análise sem uso de sensores vestíveis e outros dispositivos [Silva et al. 2021].

Diversas abordagens têm sido exploradas na literatura para representar os dados de vídeos, tais como: os próprios vídeos RGB, pontos-chaves corporais extraídos dos vídeos e características de movimento obtidas via fluxo óptico. Embora seja possível observar uma prevalência de soluções baseadas em pontos-chaves, não há consenso quanto a melhor forma de representar os dados dos vídeos. Além disso, a maioria dos estudos concentra-se na análise de FMs, devido à sua alta preditividade para condições neurológicas como a PC [Warrington et al. 2021]. No entanto, os WMs, que ocorrem nas primeiras semanas de vida, também são cruciais para um diagnóstico precoce, mas têm recebido menos atenção.

Dado esse contexto, este trabalho apresenta uma análise comparativa entre diferentes métodos consolidados na literatura para a GMA, aplicando-os especificamente aos WMs. O objetivo é avaliar como diferentes formas de entrada — vídeos RGB, pontos-chave e fluxo óptico — influenciam o desempenho dos modelos de classificação na detecção de movimentos típicos e atípicos, contribuindo para o desenvolvimento de ferramentas automatizadas de avaliação motora infantil. Além disso, duas diferentes formas de extração de informações dos pontos-chave são investigadas: 1) extração manual de características a partir das coordenadas; e 2) uso das próprias coordenadas dos pontos-chave. Por fim, considerando a diversidade provida pelas diferentes representações dos dados, nós também realizamos a fusão dos modelos treinados com cada representação. O objetivo desta abordagem de comitê é explorar a complementariedade das diferentes representações dos dados.

Os nossos experimentos mostram que a representação dos movimentos tem impacto direto no desempenho da classificação. O modelo treinado diretamente com os vídeos RGB obteve acurácia de 0,83, enquanto que o modelo treinado com histogramas extraídos dos pontos-chave alcançou o melhor F1-score (0,83), além de ser mais eficiente computacionalmente. O uso dos modelos em um comitê simples, criado por meio da fusão das predições individuais elevou a acurácia geral para 0,87. Esses resultados reforçam o potencial da combinação de

múltiplas representações para aprimorar a análise automatizada dos GMs e contribuir para diagnósticos mais precoces e confiáveis. O restante do artigo está organizado da seguinte forma. Na Seção 2, são descritos alguns trabalhos relacionados com nossa pesquisa. Na Seção 3, detalhamos a metodologia utilizada e descrevemos a base de dados, o pré-processamento e o processo de extração de características, assim como detalhamos cada abordagem investigada. Na Seção 4, apresentamos os resultados e as discussões. Por fim, na Seção 5, as conclusões do estudo e trabalhos futuros são destacados.

2. Trabalhos Relacionados

Nesta seção, descrevemos alguns trabalhos voltados para a classificação de GMs de bebês a partir de vídeos. Conforme mencionado na introdução, neste artigo nós investigamos três abordagens principais de representação: vídeos RGB, pontos-chave extraídos dos vídeos e características derivadas do fluxo óptico. Dessa forma, os trabalhos descritos nesta seção estão divididos por tipo de representação.

2.1. Vídeos RGB

A utilização de vídeos RGB como entrada para modelos de aprendizado de máquina tem sido explorada na literatura para a análise de GMs de bebês. Os modelos utilizados são majoritariamente modelos profundos, os quais requerem grandes volumes de dados rotulados e alto poder computacional para processar sequências temporais.

Dentre os trabalhos que exploram vídeos RGB diretamente, destaca-se o estudo de [Hesse et al. 2021]. Os autores seguiram uma abordagem que foca na reconstrução da forma corporal tridimensional dos bebês a partir de vídeos RGB-D. Porém, esse estudo não realiza a classificação de GMs em típicos e atípicos, pois seu objetivo é apresentar um modelo baseado em aprendizado profundo para rastrear e estimar a pose corporal dos bebês em 3D, o que pode ser útil para futuras aplicações na análise automatizada de GMs.

Já no trabalho de [Hashimoto et al. 2022], os autores propuseram uma rede neural convolucional de dois fluxos (*two-stream CNN*) para classificar GMs. O modelo processa simultaneamente informações de vídeos RGB brutos e fluxo óptico, permitindo capturar tanto características espaciais quanto temporais dos movimentos. Os resultados demonstraram que a fusão dessas duas representações melhorou o desempenho da classificação em comparação ao uso isolado de cada modalidade.

Por fim, em [Palheta et al. 2023], os autores utilizaram a rede *MoViNet*, uma CNN 3D pré-treinada, para processar duas modalidades distintas de entrada: vídeos RGB convencionais e vídeos artificiais gerados a partir de pontos-chave tridimensionais. Esses pontos-chave foram obtidos com o *VideoPose3D* e, posteriormente, convertidos em sequências visuais no formato RGB (*keypoints RGB*), buscando representar a dinâmica dos movimentos corporais dos bebês. Cada tipo de entrada foi processado separadamente, gerando duas predições independentes que foram posteriormente combinadas por meio de uma fusão baseada em soma ponderada. Os resultados indicaram que a MoViNet aplicada exclusivamente aos vídeos RGB originais superou tanto a entrada derivada dos pontos-chave quanto a fusão das duas fontes. Esse comportamento sugere que os pontos-chave tridimensionais, ao serem convertidos para o formato RGB, podem ter perdido informações discriminativas relevantes ou introduzido redundâncias em relação aos dados já contidos nos vídeos originais. Também indicam que, embora a fusão de diferentes modalidades de dados seja frequentemente promissora, sua eficácia depende da qualidade e complementaridade das informações providas por cada modalidade. Considerando esse contexto, neste trabalho nós utilizamos o modelo *MoViNet* para a tarefa de classificação de GMs

usando somente dados de vídeos RGB.

2.2. Fluxo Óptico

O fluxo óptico de grande deslocamento (LDOF) é uma técnica que calcula o deslocamento de pixeis entre quadros consecutivos em um vídeo, permitindo rastrear movimentos de forma detalhada. Essa abordagem tem sido amplamente utilizada na área médica, desde o rastreamento ocular em estudos sobre autismo [Solovyova et al. 2020, Washington et al. 2021] até na GMA, como em [Hashimoto et al. 2022, Orlandi et al. 2018, Raghuram et al. 2019, Raghuram et al. 2022].

No estudo de [Raghuram et al. 2022], os autores utilizaram LDOF para extrair características como a velocidade média na direção vertical, mediana, desvio padrão e quantidade mínima de movimento. Essas variáveis foram incorporadas em um modelo multivariado para prever PC. O modelo apresentou uma sensibilidade de 55%, especificidade de 80%, valor preditivo positivo de 26% e valor preditivo negativo de 93%, com um C-statistic indicando um bom ajuste (C = 0,74).

Já em [Orlandi et al. 2018], os autores propuseram uma metodologia para a detecção de movimentos típicos e atípicos, também utilizando características extraídas a partir do fluxo óptico. Inicialmente, o bebê foi segmentado do fundo do vídeo utilizando o fluxo óptico para isolar as áreas em movimento, garantindo que apenas os movimentos relevantes fossem analisados. Para identificar as Sequências de Movimento (*Motion Sequences* - MSs), os autores definiram um limiar de velocidade: movimentos eram considerados relevantes quando ultrapassavam 0,5 pixel/frame, separando regiões ativas de repouso [Orlandi et al. 2018].

Após identificar as MSs, um total de 643 características numéricas foram extraídas de cada vídeo, que passaram por um processo de seleção a fim de identificar as mais relevantes. Ao final, nove características foram selecionadas: média da quantidade de movimento (*Mean Q*), média da orientação (*Mean O*), índice de mobilidade lateral (*E*), valor mínimo da velocidade no eixo x (*Min of Vx*), valor mínimo da velocidade global (*Min of V*), média da velocidade da silhueta do bebê, mediana da velocidade do centroide nas sequências de movimento e no vídeo completo, e porcentagem da área convexa em relação à área do quadro. Essas características foram utilizadas como entrada para diferentes classificadores, incluindo *Support Vector Machine* (SVM), *Random Forest* e *Linear Discriminant Analysis* (LDA). Os melhores resultados obtidos foram: acurácia média de 81,89% e medida F1 de 79,44%.

Ao comparar os dois estudos, observa-se que ambos utilizaram técnicas de análise de movimento para prever condições neurológicas em bebês. Enquanto [Orlandi et al. 2018] focaram na classificação de movimentos típicos e atípicos utilizando múltiplas características e algoritmos de aprendizado de máquina, [Raghuram et al. 2022] concentraram-se na predição de PC em bebês prematuros, utilizando um conjunto específico de características de movimento em um modelo estatístico. Portanto, este trabalho segue a abordagem metodológica proposta por [Orlandi et al. 2018], utilizando fluxo óptico para extrair características de movimento e explorando modelos de aprendizado de máquina para a classificação de movimentos típicos e atípicos em bebês.

2.3. Pontos-Chave

A utilização de pontos-chave extraídos de vídeos tem se mostrado uma abordagem promissora para a classificação automatizada de GMs. Diferente do fluxo óptico, que rastreia o deslocamento de pixeis entre quadros, os pontos-chave permitem um rastreamento mais estruturado, focado na posição e movimento das articulações do bebê. Essa técnica pode capturar padrões

sutis de movimento, oferecendo uma representação mais abstrata e interpretável dos GMs. O uso de pontos-chave é vantajoso por reduzir significativamente a dimensionalidade dos dados, tornando o processamento mais eficiente do que abordagens baseadas em vídeos RGB brutos. Essa representação pode ser utilizada de duas formas principais: (1) Entrada direta das coordenadas dos pontos-chave no classificador; e (2) Extração de características derivadas dos pontos-chave para posterior classificação.

2.3.1. Uso Direto de Coordenadas

Uma abordagem comum para a classificação automática de GMs é a alimentação direta dos pontos-chave extraídos dos vídeos em modelos de aprendizado profundo. Essa estratégia foi explorada no trabalho de [Chopard et al. 2024], que, além de investigar a eficácia das coordenadas brutas dos pontos-chave para a classificação dos movimentos, também propuseram um novo para a extração e rotulagem automatizada dos pontos-chave. Os autores utilizaram o OpenPose e o *framework* proposto para extrair as coordenadas das articulações dos bebês a partir dos vídeos. Em seguida, testaram diferentes modelos de aprendizado para classificar os GMs, incluindo CNN 1D, redes recorrentes do tipo LSTM e Random Forest. Os resultados indicaram que a CNN 1D e a LSTM foram mais eficazes na detecção de padrões motores. O estudo também mostrou que a entrada direta dos pontos-chave apresentou desempenho competitivo em relação à abordagem que utiliza características derivadas dos pontos-chave.

Outro estudo relevante é o de [Zhu et al. 2021], que propuseram uma abordagem baseada em atenção por canal para identificar quais articulações eram mais relevantes para a classificação de PC. Diferentemente do trabalho de [Chopard et al. 2024], que testaram vários modelos de aprendizado, [Zhu et al. 2021] utilizaram exclusivamente CNN 2D combinada com um mecanismo de atenção para destacar os pontos-chave mais importantes na decisão do modelo. Embora o estudo tenha obtido resultados promissores, sua aplicação é mais restrita por duas razões principais. A primeira é o uso de uma base de dados sintéticos: os vídeos do estudo foram gerados com modelos 3D sintéticos de bebês (*MINI-RGBD*), diferentemente dos vídeos clínicos reais utilizados por [Chopard et al. 2024]. A segunda razão é o foco na predição de PC. Enquanto [Chopard et al. 2024] abordam a detecção de GMs em geral, [Zhu et al. 2021] focam exclusivamente na predição de um distúrbio neuromotor específico. Dessa forma, neste estudo nós utilizaremos um dos modelos investigados por [Chopard et al. 2024], precisamente a CNN 1D que, juntamente com a LSTM, apresentou os melhores resultados na classificação de GMs a partir de coordenadas de pontos-chave.

2.3.2. Extração de Características Derivadas de Pontos-Chave

A extração de características a partir dos pontos-chave permite representar os padrões de movimento de forma mais estruturada, reduzindo ruídos e tornando os dados mais interpretáveis para modelos de aprendizado. Em vez de utilizar diretamente as coordenadas brutas das articulações, essa abordagem transforma os pontos-chave em métricas que capturam aspectos biomecânicos essenciais, como amplitude, fluidez e sincronia dos movimentos.

[Doroniewicz et al. 2020] extraíram três principais características para descrever movimentos espontâneos em recém-nascidos a partir de vídeos, utilizando janelas deslizantes de 15 segundos com sobreposição de 10 segundos. As características extraídas foram: FMA (Fator da Área do Movimento), FMS (Fator da Forma do Movimento) e CMA (Centro da Área do Movimento). Diferentes métodos rasos de aprendizado de máquina foram investigados. O

melhor desempenho foi obtido pelo SVM, alcançando 80% de precisão e AUC de 0,83 para a classificação de bebês prematuros com repertório pobre de movimentos (subconjunto dos movimentos de contorção - WMs).

Em um estudo mais recente, [McCay et al. 2022] focaram na extração de oito tipos de características a partir dos pontos-chave obtidos via OpenPose. Inicialmente, 25 juntas foram identificadas, mas as relacionadas à face e aos pés foram removidas devido à baixa confiabilidade. Após essa filtragem, os pontos-chave restantes foram utilizados para calcular as características, as quais foram divididas em duas categorias: pose e velocidade. A primeira categoria consistiu na concatenação dos Histogramas de Orientação das Juntas (HOJO2D), Histogramas de Orientação Angular das Juntas em 2D (HOAD2D), Histogramas de Orientação Relativa das Juntas (HORJO2D)—que analisam a sincronia entre diferentes articulações—e da Transformada Rápida de Fourier aplicada às Orientações das Juntas (FFT-JO). Já a segunda categoria é composta por Histogramas de Deslocamento das Juntas (HOJD2D), Histogramas de Deslocamento Angular Relativo das Juntas (HORJAD2D) e Histogramas de Deslocamento Angular das Juntas (FFT-JD). Após a extração, as características foram normalizadas via *Z-score* e utilizadas como entrada para diferentes classificadores, incluindo Regressão Logística, SVM, Árvores de Decisão e Redes Neurais. Os resultados indicaram que a fusão de múltiplas características melhorou significativamente o desempenho dos modelo e superou em relação estudos anteriores. Por essa razão, o trabalho de [McCay et al. 2022] foi utilizado neste artigo como abordagem de extração de características a partir de pontos-chave.

3. Metodologia

Nesta seção, apresentamos a metodologia utilizada para alcançar os objetivos desta pesquisa. Iniciamos com uma descrição da base de dados e das técnicas aplicadas para o processamento e análise dos vídeos.

3.1. Base de Dados

Os experimentos deste trabalho foram realizados utilizando uma base de dados própria, composta por vídeos de 26 bebês prematuros ou com baixo peso. Esses vídeos foram obtidos em parceria com pesquisadores da Faculdade de Educação Física e Fisioterapia (FEFF), da Universidade Federal do Amazonas (UFAM), e rotulados por um especialista em GMA. As gravações originais, com durações entre 1 minuto e meio e 3 minutos, foram segmentadas em trechos de 30 segundos, resultando em 900 quadros por segmento a uma taxa de 30 quadros por segundo (fps). Dessa forma, o conjunto de dados final é composto por 54 instâncias de vídeos de bebês, sendo 29 instâncias da classe atípica e 25 da classe típica. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da UFAM, sob o CAAE nº 78265824.5.0000.5020 e parecer nº 6.765.526, em conformidade com a Resolução CNS nº 466/2012.

3.2. Processamento dos Vídeos

As técnicas de processamento aplicadas aos vídeos variaram conforme o tipo de representação utilizada nos diferentes métodos de aprendizado de máquina investigados. A seguir, detalhamos as principais abordagens:

- **Vídeos RGB:** Para o treinamento do modelo com vídeos RGB, os vídeos originais foram utilizados. Porém, como o modelo empregado foi a *MoViNet*, nós aplicamos as seguintes operações de aumento de dados: rotação, *zoom*, deslocamento, espelhamento horizontal, adição de ruído, ajuste de contraste, ajuste de brilho e ruído sal e pimenta. Cada operação foi aplicada com uma probabilidade de 50%, visando remover influências indesejadas e garantir uma análise mais robusta dos movimentos dos bebês.

- **Fluxo Óptico de Grande Deslocamento (LDOF):** Para o treinamento com LDOF, utilizamos os vídeos originais e aplicamos a técnica de segmentação proposta por [Li et al. 2022]. No estudo original, a segmentação dos movimentos é realizada diretamente a partir do próprio fluxo óptico, combinada com máscaras de pele para isolar as regiões de interesse. No entanto, como nossos vídeos não foram capturados com uma câmera fixa, a aplicação direta desse método foi inviável. Para contornar essa limitação, adotamos a abordagem de [Li et al. 2022].
- **Pontos-chave:** Para a extração de juntas corporais, utilizamos o MediaPipe, escolhido após uma análise comparativa entre diferentes extratores de pontos-chave, incluindo MoveNet, OpenPose e EfficientPose. Durante os testes iniciais, o MediaPipe demonstrou maior estabilidade e precisão na detecção das articulações dos bebês. Após a extração dos pontos-chave, aplicamos a normalização e a técnica de aumento de dados *pose format* [Moryossef et al. 2023], que permitiu aumentar a variabilidade dos dados de pose e melhorar a robustez dos modelos treinados, especialmente a CNN 1D.

3.3. Extração de Características e Treinamento dos Modelos

Após o processamento, os vídeos RGB foram utilizados como entrada para a MoViNet, uma CNN 3D pré-treinada para reconhecimento de ações em vídeos. Optamos pelo modelo MoViNetA2 Stream por sua maior eficiência computacional, permitindo um tamanho de lote maior durante o treinamento. Foi realizado o ajuste fino durante o treinamento e a definição de hiperparâmetros, buscando otimizar o desempenho do modelo na classificação dos movimentos.

Já para os vídeos segmentados e utilizados pelo método LDOF, aplicamos o cálculo da velocidade dos pixeis entre quadros consecutivos. Diferentemente da abordagem proposta em [Orlandi et al. 2018], nós utilizamos janelas deslizantes de 2 segundos com sobreposição de 1 segundo. Essa abordagem foi escolhida para capturar padrões motores mais amplos, reduzindo a sensibilidade a ruídos e permitindo uma análise mais robusta da dinâmica do movimento ao longo do tempo. Assim, cada vídeo resultou em 29 janelas.

As características extraídas para cada janela incluíram a velocidade média dos pixeis, a velocidade mediana, o valor mínimo das velocidades, a velocidade mínima no eixo x, a velocidade média calculada no centroide da área em movimento, o índice de mobilidade lateral (relação entre as velocidades médias nos eixos x e y) e a proporção entre a área convexa que envolve a silhueta em movimento e a área total do quadro. Os vetores de características extraídos do LDOF foram então utilizados como entrada para métodos de aprendizado de máquina rasos usados em [Orlandi et al. 2018], incluindo Random Forest, Logistic Regression, AdaBoost e LogitBoost, sendo que o primeiro modelo mostrou melhor desempenho. Por essa razão, os resultados apresentados na próxima seção são somente os resultados do método Random Forest.

Os pontos-chave extraídos e pré-processados foram utilizados com dois grupos de métodos de aprendizado de máquina. No primeiro grupo, as coordenadas foram usadas de forma direta nos modelos de redes neurais propostos em [Chopard et al. 2024], precisamente CNN1D e LSTM. Como a CNN1D obteve os melhores resultados, estes são reportados na próxima seção. O segundo grupo segue a abordagem de [McCay et al. 2022]. Os histogramas com as características foram concatenados e avaliados com os seguintes métodos de aprendizado de máquina: Árvore de Decisão, LDA, Logistic Regression, SVM e um comitê de modelos. O método Árvore de Decisão, utilizando as características HOJO2D, obteve as melhores taxas, as quais são destacadas na Seção 4.

Devido ao número reduzido de segmentos de vídeos na base de dados, nós utilizamos a abordagem de validação *Leave-One-Subject-Out* (LOSO). Esse método consiste em treinar

o modelo utilizando os dados dos segmentos de vídeos de todos os bebês, exceto um, que é reservado para teste. O processo é repetido para cada indivíduo da base de dados, garantindo que os dados de cada bebê sejam utilizados como teste ao menos uma vez. Dessa forma, a abordagem LOSO permite avaliar a capacidade de generalização dos modelos para novos indivíduos, o que é essencial em um contexto clínico.

Por fim, usamos uma abordagem de comitê de classificadores por fusão de predição para combinar as saídas dos modelos e gerar uma decisão mais robusta. Na fusão por soma, as probabilidades das classes (0 ou 1) são somadas e a classe com maior valor é escolhida. Na fusão por média, calcula-se a média, e a classe com maior valor médio é selecionada. Já na fusão por produto, as probabilidades são multiplicadas, enfatizando classes com consenso mais forte entre os modelos. Esses métodos ajudam a explorar a complementaridade das representações dos dados e melhorar a precisão da classificação.

4. Experimentos e Resultados

Nesta seção, serão apresentados os resultados dos experimentos. Os resultados são apresentados em três partes: (1) desempenho geral dos modelos individuais; (2) análise da complexidade computacional; e (3) resultados da fusão de predições para explorar a complementaridade das diferentes abordagens.

4.1. Comparações de Desempenhos Individuais

A Tabela 1 resume o desempenho geral dos modelos individuais, utilizando as métricas de acurácia, precisão, revocação e F1-score. O valor mais elevado está destacado em negrito. É importante mencionar que os valores mostrados foram obtidos por meio do cálculo da média entre os dados de cada indivíduo da base de dados.

Tabela 1. Resultados de classificação dos modelos individuais

Modelo	Acurácia	Precisão	Recall	F1-Score
MoViNet (Vídeos RGB)	**0,83**	**0,92**	0,76	0,83
CNN1D (Pontos-Chaves)	0,59	0,61	0,69	0,65
Árvore de Decisão (Histogramas)	0,81	0,81	**0,86**	**0,83**
Random Forest (Fluxo Óptico)	0,72	0,71	0,83	0,76

Como pode ser observado nessa tabela, o modelo MoViNet utilizando os próprios vídeos RGB obteve as maiores taxas de acurácia e de precisão. Porém, a árvore de decisão utilizando as características obtidas via histogramas alcançou as maiores taxas de revocação e F1-Score. Isso mostra uma diversidade nos erros cometidos pelos modelos nas duas classes do problema. Portanto, para analisar melhor a distribuição do erro entre as classes, a Tabela 2 mostra as matrizes de confusão de cada modelo. Com base nessa tabela, nós podemos destacar os seguintes pontos:

- A árvore de decisão (AD) errou menos a classe atípica, que é a classe mais importante em nossa análise, dado que é mais importante classificar corretamente todos os casos de atipicidade de movimento do que de tipicidade. Dentre as 29 instâncias atípicas, a AD classificou incorretamente apenas 4, enquanto a MoviNet classificou incorretamente 7 instâncias da classe atípica.
- A distribuição dos erros dos classificadores individuais entre as classes apresenta elevada variância. Isso indica que a fusão da decisão dos modelos pode melhorar os resultados individuais, dado que conjuntos de classificadores podem superar classificadores individuais somente quando os membros do conjunto apresentam diversidade de erro.

Tabela 2. Matriz de Confusão de todos os modelos individuais

Modelo	TP	TN	FP	FN
MoViNet (Vídeos RGB)	22	23	2	7
CNN1D (Keypoints)	20	12	13	9
Árvore de Decisão (Histogramas)	25	19	6	4
Random Forest (Optical Flow)	24	15	10	5

TP = Verdadeiros Positivos, TN = Verdadeiros Negativos, FP = Falsos Positivos, FN = Falsos Negativos.

Um outro fator importante a avaliar é a capacidade de generalização dos modelos ao variar os dados de diferentes indivíduos. Nesse contexto, nós mostramos na Figura 1 a distribuição das acurácias de cada modelo ao variarmos o indivíduo representado na base de teste. Essa figura mostra o boxplot das acurácias de cada modelo por indivíduo, permitindo visualizar a mediana, os quartis, os valores atípicos e a dispersão dos dados para cada modelo.

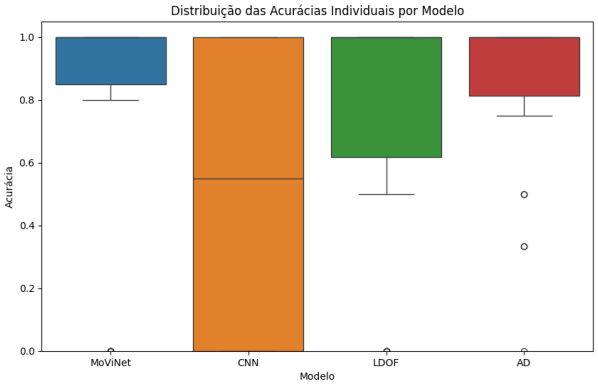

Figura 1. Distribuição das acurácias dos modelo ao variar o indivíduo representado na base de teste.

É possível observar nessa figura que a MoViNet foi o modelo mais estável, dado que a acurácia variou entre 0,8 e 1,0, dependendo do indivíduo representado no teste. Um comportamento oposto foi apresentado pela CNN1D, que exibiu elevada variância na taxa de acurácia (de 0,0 a 1,0), a depender da base de teste. Por fim, a AD também apresentou comportamento estável; no entanto, foi mais sensível aos dados quando comparada à MoViNet, pois mostrou dispersão moderada e alguns *outliers*.

4.2. Avaliação de Custo Computacional

A MoViNet, embora eficiente na captura de padrões visuais dos movimentos, demanda alto poder computacional tanto para treinamento quanto para inferência, o que pode limitar sua aplicação em cenários clínicos com restrição de hardware. O fluxo óptico, apesar de fornecer informações relevantes sobre a dinâmica dos movimentos, exige um tempo de processamento elevado, pois cada pixel contribui para a análise do deslocamento. Isso torna o processo de extração de características mais lento em comparação com métodos baseados em representações compactas.

Quanto aos modelos baseados em pontos-chave, a CNN1D apresenta menor custo computacional, comparada à MoViNet por exemplo, porém, como apresentou desempenho inferior às demais abordagens, pode não ser a melhor opção de equilíbrio entre custo e benefício. Por fim, a abordagem baseada em histogramas demonstrou um bom custo-benefício, sendo a opção

mais acessível para implementação em ambientes com restrição computacional. Além de ser um modelo leve, a AD é também um modelo transparente, fator que ajuda significativamente na explicabilidade das decisões do modelo.

Dessa forma, a escolha do melhor método individual depende dos recursos disponíveis e do contexto de aplicação. Enquanto a MoViNet se destaca em acurácia e baixa variância de desempenho entre dados de diferentes indivíduos, esse modelo requer alto poder computacional. O fluxo óptico oferece informações detalhadas, mas também apresenta um custo elevado. Já os histogramas, combinados com AD, representam uma alternativa eficiente para cenários com restrições de hardware.

4.3. Combinação de Classificadores

Nesta seção nós avaliamos a possibilidade de utilizar a combinação dos classificadores individuais investigados nas subseções anteriores a fim de explorar a complementaridade das diferentes representações de dados.

A combinação é feita por meio da fusão das predições, que consiste em combinar as saídas de múltiplos modelos para obter uma predição final mais robusta e precisa. Neste trabalho, nós testamos três funções de fusão: 1) produto; 2) soma; e 3) média das probabilidades preditas por cada modelo individual. O objetivo foi determinar se a combinação dos modelos treinados com dados representados pelas diferentes abordagens (vídeos RGB, pontos-chave, histogramas e fluxo óptico) pode produzir um desempenho superior em relação aos modelos individuais. Por questão de limitação de espaço, nós comparamos os resultados da fusão com o melhor modelo individual em termos de acurácia, ou seja, a MoViNet.

Os resultados são exibidos na Tabela 3. Esse resultados mostram que a fusão, utilizando tanto o método da soma quanto o da média, resultou em um aumento em todas as medidas, exceto em precisão, cujo desempenho foi similar ao desempenho da MoViNet. É importante destacar o aumento em F1-Score, fato que indica que a fusão foi benéfica para ambas as classes típica e atípica. Esse resultado sugere que a combinação das diferentes representações de dados permite capturar informações complementares sobre a dinâmica dos movimentos gerais, levando a uma classificação mais precisa e acurada.

Tabela 3. Comparando a Fusão de Predições com a MoViNet individual

Método de Fusão	Acurácia	Precisão	Recall	F1-Score
Produto	0,78	0,79	0,79	0,79
Soma	0,87	0,92	0,83	0,87
Média	0,87	0,92	0,83	0,87
MoViNet	0,83	0,92	0,76	0,83

Esses resultados sugerem que a combinação de classificadores treinados com diferentes representações dos dados de vídeos é uma abordagem promissora para melhorar a classificação dos GMs. Essa combinação permite capturar informações complementares sobre a dinâmica dos movimentos, levando a um desempenho superior em relação aos modelos individuais. No entanto, é importante ressaltar que a escolha da melhor combinação de modelos e do método de fusão depende do contexto específico e dos recursos computacionais disponíveis.

5. Conclusão

Este estudo comparou diferentes representações baseadas em vídeo para uso com modelos de aprendizado de máquina na tarefa de classificação de Movimentos Gerais de recém-nascidos,

focando na fase dos Writhing Movements. Foram avaliados vídeos RGB, fluxo óptico (LDOF), pontos-chave e histogramas de movimento calculados a partir dos pontos-chave.

Os resultados mostraram que o modelo que utilizou vídeos RGB apresentou a maior acurácia, mas alto custo computacional. Os histogramas de movimento utilizados por uma árvore de decisão equilibraram melhor desempenho e eficiência computacional, obtendo o melhor F1-score, especialmente porque o modelo errou menos instâncias da classe atípica. O fluxo óptico foi eficaz na extração de padrões motores, porém, exigiu alto tempo de processamento. Já o modelo baseado nas coordenadas dos pontos-chave demonstrou desempenho inferior, sugerindo que a simples extração de juntas pode não capturar informações suficientes para a classificação dos GMs. A fusão de predições entre diferentes modelos melhorou significativamente os resultados, evidenciando que abordagens complementares são mais eficazes.

Para trabalhos futuros, sugerimos: (i) a ampliação da base de dados; (ii) a investigação de modelos mais avançados, como arquiteturas baseadas em Transformers e redes do tipo Graph Neural Networks (GNNs), que têm mostrado bons resultados em tarefas de reconhecimento de ações; (iii) a aplicação das abordagens em cenários clínicos reais; e (iv) a exploração de representações tridimensionais combinadas com pontos-chave. Este estudo reforça o potencial da fusão de predições na análise automatizada dos GMs, contribuindo para o avanço de ferramentas diagnósticas precoces de distúrbios neuromotores em bebês.

6. Agradecimentos

O presente trabalho é resultado do Projeto de Pesquisa e Desenvolvimento (PD) 001/2020, firmado com a UFAM e FAEPI, Brasil, financiado pela Samsung Eletrônica da Amazônia Ltda, nos termos da Lei Federal n°8.387/1991, e sua divulgação está de acordo com o artigo 39 do Decreto nº 10.521/2020. Este trabalho também foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Código de Financiamento 001. Adicionalmente, este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM - por meio do projeto PDPG.

Referências

Camargos, F. O. and Almeida, J. G. A. d. (2019). Paralisia cerebral: revisão e considerações atuais. *Acta Fisiátrica*, 26(3):144–152.

Chopard, D., Laguna, S., Chin-Cheong, K., Dietz, A., Badura, A., Wellmann, S., and Vogt, J. E. (2024). Automatic classification of general movements in newborns. In *Findings of the AHLI Machine Learning for Health (ML4H) Symposium*, Vancouver, Canada.

Doroniewicz, I., Ledwoń, D. J., Affanasowicz, A., Kieszczyńska, C., Latos, D., Matyja, M., Mitas, A. W., and Myśliwiec, A. (2020). Writhing movement detection in newborns on the second and third day of life using pose-based feature machine learning classification. *Sensors*, 20(21):5986.

Ferrari, F. and Cioni, G. (2016). Fidgety movements—tiny in appearance, but huge in impact. *Journal of Pediatrics*, 92:S64–S70.

Hadders-Algra, M. (2004). General movements: A window for early identification of children at high risk for developmental disorders. *The Journal of Pediatrics*, 145(2 Suppl):S12–S18.

Hashimoto, Y., Ishikawa, K., et al. (2022). Automated classification of general movements in infants using two-stream spatiotemporal fusion network. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:456–465.

Hesse, N., Pujades, S., Black, M. J., Arens, M., Hofmann, U. G., and Schroeder, A. S. (2021). Learning and tracking the 3d body shape of freely moving infants from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14703–14713.

Li, P., Xu, Y., Wei, Y., and Yang, Y. (2022). Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271.

McCay, K. D., Hu, P., Shum, H. P. H., Ho, E. S. L., Woo, W. L., Marcroft, C., Embleton, N. D., and Munteanu, A. (2022). A pose-based feature fusion and classification framework for the early prediction of cerebral palsy in infants. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:8–19.

Moryossef, A. et al. (2023). Poseformat: Library for viewing, augmenting, and handling pose files. *arXiv preprint arXiv:2310.09066*.

Orlandi, S., Cinque, L., Ferrante, G., Sgandurra, G., and Cioni, G. (2018). Detection of atypical and typical infant movements using computer-based video analysis. *Journal of Medical Imaging*, 5(2):024001.

Palheta, M., Santos, G., Mendonça, A., Gonçalves, P., Albuquerque, R., Souto, E., and Santos, E. (2023). Fusão de dados de vídeos rgb e pontos-chaves para classificação de movimentos gerais de bebês. In *Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 384–394, Porto Alegre, RS, Brasil. SBC.

Prechtl, H. F. (1997). General movements: A window for early detection of developmental disorders. *Pediatrics*, 99:92–101.

Raghuram, K., Orlandi, S., Church, P., Luther, M., Kiss, A., and Shah, V. (2022). Automated movement analysis to predict cerebral palsy in very preterm infants: An ambispective cohort study. *Children*, 9(6):843–851.

Raghuram, K., Orlandi, S., Shah, V., Chau, T., Luther, M., Banihani, R., and Church, P. (2019). Automated movement analysis to predict motor impairment in preterm infants: A retrospective study. *Journal of Perinatology*, 39(10):1362–1369.

Silva, N., Zhang, D., Kulvicius, T., Gail, A., Barreiros, C., Lindstaedt, S., Kraft, M., Bölte, S., Poustka, L., Nielsen-Saines, K., Wörgötter, F., Einspieler, C., and Marschik, P. (2021). The future of general movement assessment: The role of computer vision and machine learning - a scoping review. *Research in Developmental Disabilities*, 110.

Solovyova, S. et al. (2020). Eye-tracking metrics in children with autism spectrum disorder. *arXiv preprint arXiv:2008.09670*.

Spittle, A. J. and Doyle, L. W. (2018). Identification of neurodevelopmental impairments in preterm infants. *Journal of Pediatrics*, 191:20–28.

Warrington, H. et al. (2021). A systematic review of automated methods for general movements assessment. *Developmental Medicine & Child Neurology*, 63:745–756.

Washington, P. et al. (2021). Computer-based analysis of eye tracking to detect early signs of autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 51(5):1435–1450.

Zhu, M., Men, Q., Ho, E. S. L., Leung, H., and Shum, H. P. H. (2021). Interpreting deep learning based cerebral palsy prediction with channel attention. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.