

Desvendando os Subgrupos Moleculares do Meduloblastoma: Uma Abordagem baseada em IA Explicável com XGBoost e SHAP

Raylam W. da Silva^{*†}, Matheus Dalmolin^{*†‡§} e Marcelo A. C. Fernandes^{*†‡§¶}

^{*}InovAI Lab, nPITI/IMD, UFRN, Natal, RN, Brasil.

[†]Leading Advanced Technologies Center of Excellence (LANCE), nPITI/IMD, UFRN, Natal 59078-970, RN, Brasil

[‡]Centro Multiusuário de Bioinformática (BioME) - IMD, UFRN, Natal, RN, Brasil.

[§]INCT BioOncoPed, Porto Alegre, RS, Brasil

[¶]Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, RN, Brasil.

Email: raylamwendel@hotmail.com, matheusdalmolins@gmail.com, mfernandes@dca.ufrn.br

Resumo—O meduloblastoma (MB) é o tumor cerebral pediátrico maligno mais comum, classificado em quatro subgrupos moleculares (WNT, SHH, Grupo 3 e Grupo 4) com prognósticos e tratamentos distintos. A classificação baseada em expressão gênica é crucial, mas a complexidade dos dados e a natureza de "caixa-preta" de modelos de aprendizado de máquina (ML) dificultam a interpretação biológica. Este estudo utiliza um conjunto de dados público de expressão gênica (GSE85217) para treinar um modelo XGBoost para a classificação dos subgrupos, após balanceamento das classes com a técnica de *Random Over-Sampling*. A metodologia de Inteligência Artificial Explicável (XAI), SHAP (*SHapley Additive exPlanations*), foi empregada para identificar os genes com maior impacto na predição de cada subgrupo e desvendar os mecanismos de decisão do modelo.

O modelo XGBoost alcançou uma acurácia global de 96,7% no conjunto de teste, demonstrando alta precisão na distinção dos quatro subgrupos. A análise SHAP revelou os genes mais influentes para cada classe, corroborando marcadores biológicos bem estabelecidos, como *PPP2R2B* para o Grupo 3 e *RBM24* para o Grupo 4. Crucialmente, a análise demonstrou que a baixa expressão de *OTX2*, um oncogene em outros subgrupos, foi o preditor mais importante para o subgrupo SHH, destacando a capacidade do modelo de aprender regras de decisão contextuais e não apenas marcadores de alta expressão. Adicionalmente, a análise identificou novos candidatos potenciais, como *PDLIM4* para o subgrupo WNT. Notavelmente, uma análise de redução de atributos mostrou que um modelo com apenas os 20 genes mais importantes manteve uma acurácia robusta de 85,87%.

A combinação do XGBoost com SHAP provou ser uma estratégia poderosa não apenas para a classificação precisa, mas também para extrair insights biologicamente relevantes. A abordagem valida a coerência biológica do modelo, abre caminho para o desenvolvimento de painéis diagnósticos mais enxutos e a descoberta de novos alvos terapêuticos na neuro-oncologia pediátrica, ressaltando a importância da interpretabilidade na pesquisa do câncer guiada por IA.

Palavras-chave—Meduloblastoma, Inteligência Artificial Explicável, XGBoost, SHAP.

I. INTRODUÇÃO

O meduloblastoma (MB) é o tumor cerebral maligno pediátrico mais comum, surgindo no cerebelo e compreendendo cerca de 20% das neoplasias do sistema nervoso central

(SNC) em crianças. Este tumor é heterogêneo, sendo classificado em quatro subgrupos moleculares principais: WNT, SHH, Grupo 3 e Grupo 4, cada um com perfis demográficos, transcriptômicos, genéticos e desfechos clínicos distintos [1], [2]. A distinção destes subgrupos moleculares baseia-se na análise de dados de expressão gênica de alta dimensionalidade. Por exemplo, o subgrupo WNT geralmente tem um prognóstico excelente, enquanto o Grupo 3 é frequentemente mais agressivo [3]. A Classificação da Organização Mundial da Saúde (OMS) de 2021 (WHO CNS5) formalizou a integração destes marcadores moleculares na definição do MB, como "Meduloblastoma, WNT-ativado" e "Meduloblastoma, SHH-ativado", sendo a subgrupagem molecular crucial para a estratificação de risco e desenvolvimento de tratamentos personalizados [4], [5].

Dada a complexidade destes dados de expressão gênica, que envolvem milhares de genes, algoritmos de aprendizado de máquina (ML) como o XGBoost são cada vez mais utilizados para classificar amostras tumorais com base nas suas assinaturas de expressão gênica, alcançando alta acurácia. A aplicação de ML na classificação de subgrupos de MB valida a relevância dos perfis de expressão gênica e abre caminho para ferramentas diagnósticas moleculares mais eficientes, embora desafios como efeitos de lote e necessidade de validação rigorosa persistam [3]. Apesar da eficácia preditiva de modelos de ML como o XGBoost, a sua natureza de "caixa-preta" limita a compreensão dos mecanismos subjacentes às suas decisões, um obstáculo em domínios como a investigação biomédica [6], [7]. A Inteligência Artificial Explicável (XAI) surgiu para tornar estes modelos mais transparentes e interpretáveis [6]. O método *SHapley Additive exPlanations* (SHAP), em particular, quantifica a contribuição de cada gene (valor SHAP) para a classificação de uma amostra num determinado subgrupo de MB, permitindo interpretações globais e locais das previsões do modelo [6], [8]. A aplicação do SHAP transforma modelos de ML em ferramentas de geração de hipóteses, identificando variáveis biologicamente relevantes para validação experimental e fomentando a confiança nos achados computacionais [9].

O objetivo central desta investigação é utilizar o modelo XGBoost em conjunto com a metodologia SHAP para classificar com precisão os subgrupos moleculares do MB (WNT, SHH, Grupo 3 e Grupo 4) e, fundamentalmente, identificar os genes que mais contribuem para a distinção entre eles [10]. Esta abordagem visa desvendar os fundamentos genéticos da biologia de cada subgrupo, identificando biomarcadores conhecidos e potenciais novos candidatos, como demonstrado pela comparação dos genes identificados com a literatura científica [3]. A elucidação destes genes chave pode aprofundar o conhecimento sobre as vias moleculares específicas de cada subgrupo, contribuindo para diagnósticos mais precisos, novos alvos terapêuticos e uma medicina personalizada em neuro-oncologia pediátrica, incluindo o potencial para painéis diagnósticos mais concisos [10].

Este artigo está organizado da seguinte forma: a Seção II apresenta uma revisão de trabalhos correlatos e uma análise comparativa com estudos recentes. A Seção III detalha a metodologia, incluindo o conjunto de dados, pré-processamento, treinamento do modelo e a análise de explicabilidade. A Seção IV apresenta os resultados de desempenho e a identificação de biomarcadores. A Seção V discute as implicações biológicas e translacionais dos achados. Finalmente, a Seção VI conclui o trabalho e aponta direções para pesquisas futuras.

II. TRABALHOS RELACIONADOS

A classificação de subgrupos moleculares do MB com ML é uma área de intensa pesquisa, onde modelos alcançam alta acurácia preditiva. No entanto, muitas abordagens se concentram na otimização do desempenho ou na redução da dimensionalidade para criar painéis diagnósticos concisos, muitas vezes sem aprofundar a interpretabilidade biológica dos modelos [12]. Este trabalho se alinha a uma vertente emergente que utiliza a Inteligência Artificial Explicável (XAI) não apenas para validar, mas como uma ferramenta de investigação para extrair conhecimento biológico dos dados [13], [14].

Para contextualizar nossa contribuição, a Tabela I apresenta uma análise comparativa com estudos recentes. O trabalho de Gershanov et al. (2021) [12], por exemplo, focou na seleção de atributos para derivar um painel diagnóstico mínimo de seis genes. Embora a sobreposição de marcadores como *NPR3* e *RBM24* valide parcialmente nossos achados, o obje-

tivo primário daquele estudo era a eficiência diagnóstica via redução de dimensionalidade. Em contraste, nossa abordagem usa SHAP para mapear a contribuição específica de cada gene na classificação de cada subgrupo, permitindo identificar padrões de influência contextuais que um simples ranking de importância não revelaria. A robustez desta abordagem é reforçada pela identificação de *GABRA5* e *ATOH1* em nossa análise, genes também considerados no conjunto de candidatos do estudo de Gershanov et al.

De forma distinta, o estudo de Sharif-Rahmani et al. (2023) [11] evidencia diferenças metodológicas cruciais. A principal divergência reside no tipo de dado: enquanto nosso trabalho analisa a expressão gênica, o estudo de Sharif-Rahmani et al. parte de dados de metilação de DNA. Consequentemente, a estratégia de identificação de biomarcadores é indireta, mapeando os sítios de metilação (CpGs) mais relevantes para seus 16 genes-chave. Nossa abordagem, ao contrário, é direta, com o SHAP quantificando o impacto da expressão de cada gene na classificação. Essa diferença fundamental no dado de origem (expressão vs. metilação) e na metodologia (direta vs. indireta) explica a ausência de sobreposição nos genes identificados, destacando como diferentes camadas de dados ômicos podem revelar facetas distintas da biologia do tumor.

III. MÉTODOS

A metodologia deste estudo foi estruturada em etapas sequenciais de aquisição de dados, pré-processamento, treinamento de modelo e análise de explicabilidade, conforme ilustrado no fluxograma da Figura 1.

A. Conjunto de Dados e Pré-processamento

Foi utilizado o conjunto de dados público de expressão gênica de MB (GSE85217), proveniente do *Medulloblastoma Advanced Genomics International Consortium* e depositado na base *Gene Expression Omnibus* (GEO) [15]. Este dataset é uma coorte de referência na área, amplamente utilizada em estudos de classificação molecular devido ao seu grande número de amostras (763) e à robustez dos dados de microarranjos (plataforma Affymetrix Human Gene 1.1 ST), sendo também a base para outros trabalhos de classificação de subgrupos [11], [16].

Tabela I: Análise comparativa técnica com trabalhos recentes sobre classificação de subgrupos de meduloblastoma.

Critério	Este Estudo	Sharif-Rahmani et al. (2023) [11]	Gershanov et al. (2021) [12]
Base de dados	GSE85217	GSE54880, GSE75153, GSE85212, GSE85217, GSE90496, GSE109379, GSE130051.	GSE85217, GSE37418 e coorte própria
Tipo de Dados	Expressão gênica (Microarray)	Metilação de DNA e Expressão gênica	Expressão gênica (Microarray e qPCR)
Modelo de ML	XGBoost	Redes Neurais (ANN), SVM, RF	SVM (algoritmo SARC), <i>Decision Tree</i> , <i>Random Forest</i>
Acurácia do Modelo	96,7% (teste)	Até 99,25% (validação)	93,48% (painel de 6 genes)
Abordagem de XAI	SHAP	SHAP	Ranking de atributos
Genes em Comum	Ver Tabela II	0/16 (0%)	<i>NPR3</i> , <i>RBM24</i> (2/6, 33,3%)

O conjunto de dados compreende amostras tumorais classificadas nos quatro subgrupos moleculares: Grupo 3, Grupo 4, SHH e WNT. Uma análise exploratória inicial revelou um desbalanceamento significativo entre as classes (Grupo 3: n=264; Grupo 4: n=172; SHH: n=113; WNT: n=63). Para mitigar o viés do modelo em direção às classes majoritárias, foi aplicada a técnica de *Random Over-Sampling* (ROS) da biblioteca *imblearn*. Esta abordagem foi selecionada após uma avaliação preliminar na qual demonstrou superioridade em métricas de acurácia e estabilidade em comparação com outras técnicas (por exemplo, *SMOTE*). A aplicação de ROS resultou em um conjunto de dados de treinamento balanceado, no qual todas as classes possuíam o mesmo número de amostras da classe majoritária (n=264).

Após o balanceamento, o conjunto de dados completo foi dividido em 80% para treinamento e 20% para teste, utilizando uma abordagem estratificada para garantir que a proporção original das classes fosse preservada no conjunto de teste.

B. Construção e Treinamento do Modelo

O XGBoost (*Extreme Gradient Boosting*) é um algoritmo de ensemble baseado em árvores de decisão que se destaca por sua eficiência e desempenho [17]. Ele opera otimizando iterativamente um conjunto de árvores de decisão fracas através da técnica de *gradient boosting*. A cada passo, uma nova árvore é treinada para corrigir os erros residuais do conjunto de árvores anterior. O XGBoost incorpora regularização L1 (*Lasso*) e L2 (*Ridge*) em sua função de custo para prevenir o overfitting, tornando-o particularmente robusto para dados de alta dimensionalidade como os de expressão gênica.

Para a tarefa de classificação multiclasse, foi implementado o algoritmo XGBoost. O modelo foi configurado com o

objetivo `multi:softmax` e os hiperparâmetros foram mantidos em seus valores padrão, incluindo uma profundidade máxima de árvore (`max_depth=3`) e peso mínimo por nó (`min_child_weight=1`). O treinamento (`model.fit()`) foi realizado utilizando exclusivamente o conjunto de treinamento balanceado. O desempenho final do modelo foi avaliada no conjunto de teste, que permaneceu intocado durante as etapas de treinamento e otimização.

C. Análise de Explicabilidade e Redução de Atributos

Para interpretar as previsões do modelo XGBoost, foi empregada a metodologia SHAP [18]. Utilizando o `shap.TreeExplainer`, foi possível calcular a contribuição individual de cada gene (atributo) para a previsão de cada amostra. Os valores SHAP foram então agregados para gerar gráficos de resumo (`summary plots`), que permitiram hierarquizar os genes com base em seu impacto global na classificação de cada subgrupo molecular.

Adicionalmente, foi conduzida uma análise de redução de atributos para avaliar o desempenho do modelo com um subconjunto de genes. Com base no ranking de importância gerado pelo SHAP, o modelo foi treinado e avaliado iterativamente com diferentes números de genes para determinar a viabilidade de um painel genético mais parcimonioso.

Todas as análises foram conduzidas em ambiente *Google Colab* com a linguagem de programação *Python* (3.11), utilizando um conjunto de bibliotecas especializadas: *pandas* (v2.2.2) para manipulação de dados, *scikit-learn* (v1.5.2) para pré-processamento e avaliação, *imblearn* (v0.13.0) para a técnica de *oversampling*, *xgboost* (v2.1.3) para a construção do modelo e *shap* (v0.47.0) para a análise de explicabilidade [18]–[22].

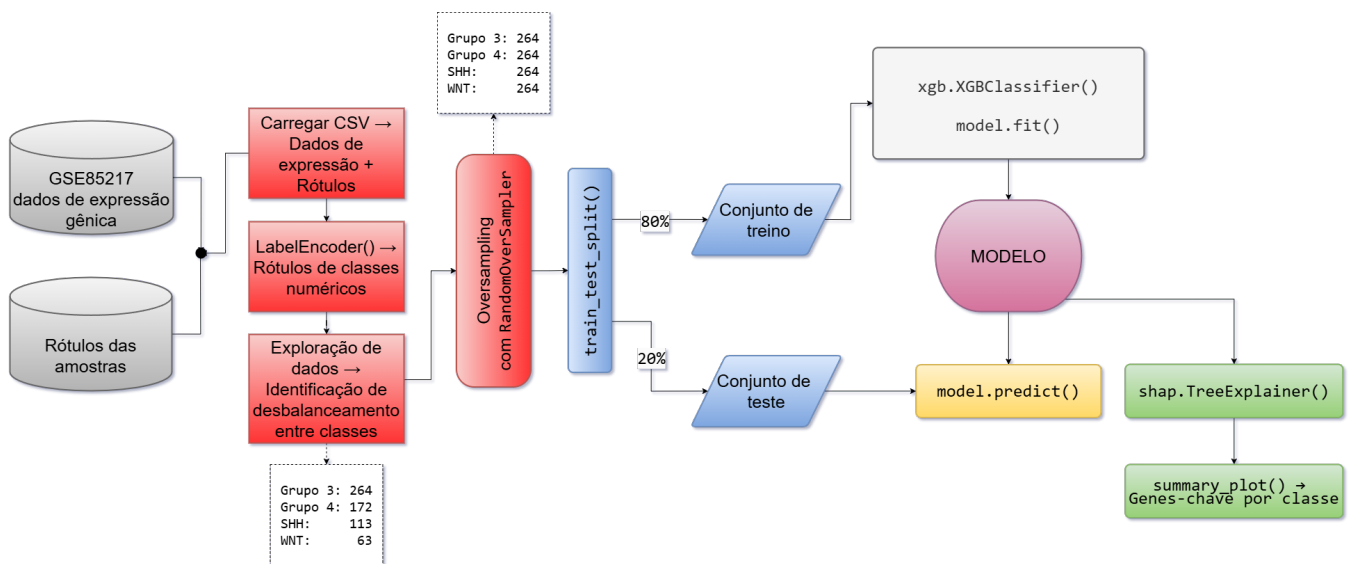
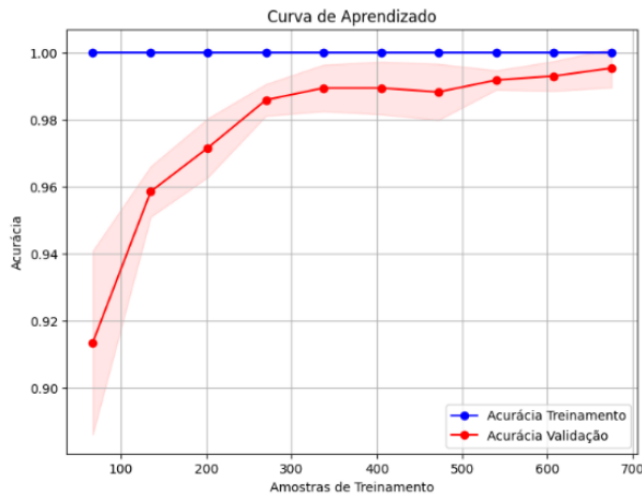
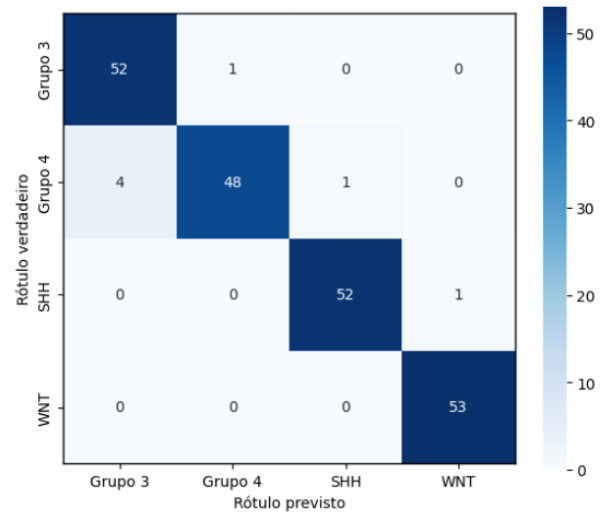


Figura 1: Fluxograma da metodologia adotada, desde a aquisição e pré-processamento dos dados (GSE85217), passando pelo balanceamento de classes com *RandomOverSampler*, treinamento do modelo XGBoost, até a análise de explicabilidade com SHAP para identificação dos genes mais importantes.



(a) Curva de aprendizado mostrando a acurácia de treinamento (azul) e validação (vermelho) em função do número de amostras. A pequena diferença entre as curvas indica boa generalização.



(b) Matriz de confusão para os quatro subgrupos moleculares, demonstrando alta precisão na classificação das amostras de teste.

Figura 2: Desempenho do modelo XGBoost.

IV. RESULTADOS

Nesta seção, apresentamos o desempenho do modelo de classificação, a análise da importância dos genes e a avaliação do modelo com um conjunto reduzido de atributos.

A. Desempenho do Modelo de Classificação

O modelo XGBoost, treinado com dados balanceados pela técnica de *Random OverSampling*, demonstrou alto desempenho na classificação dos quatro subgrupos moleculares de MB. A curva de aprendizado (Figura 2A) ilustra a robustez do modelo, que atingiu uma acurácia de treinamento de 100% e uma acurácia de validação de 99,5%, indicando excelente capacidade de generalização e baixo risco de overfitting.

A matriz de confusão (Figura 2B) corrobora esses resultados, mostrando uma classificação quase perfeita para todas as classes. Dos 212 casos de teste, apenas 7 foram classificados incorretamente, resultando em uma acurácia global de 96,7%. O modelo demonstrou alta precisão para os subgrupos SHH e WNT, classificando corretamente 100% das amostras, e apresentou apenas pequenas confusões entre os subgrupos Grupo 3 e Grupo 4.

B. Identificação de Biomarcadores Genéticos via SHAP

Para interpretar as previsões do modelo e identificar os genes mais influentes para cada subgrupo, foi utilizada a análise SHAP. A Figura 3 resume a importância média dos 20 genes mais relevantes no geral, enquanto a Figura 4 detalha a contribuição desses genes para a classificação de cada subgrupo específico. Os principais genes identificados foram confrontados com a literatura científica, e os achados estão compilados na Tabela II.

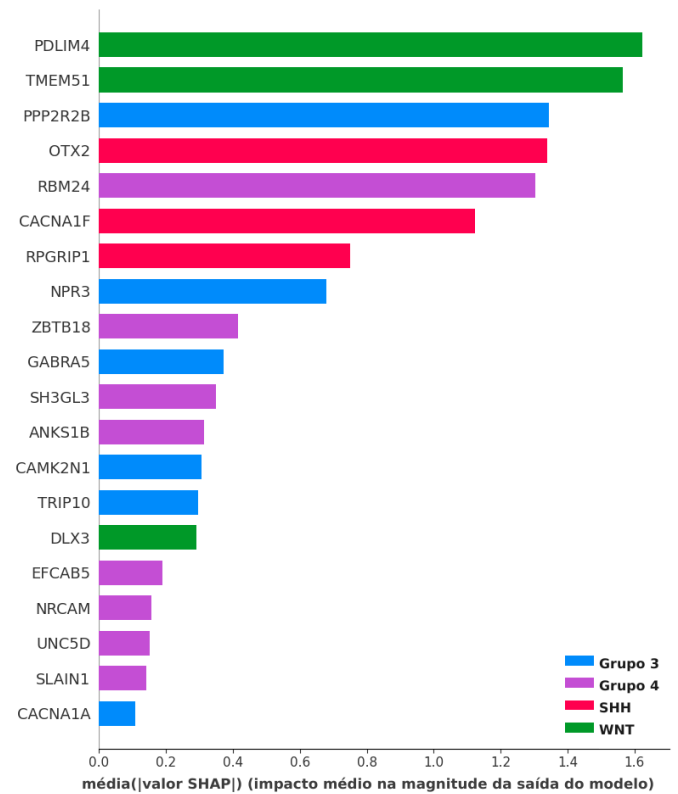


Figura 3: Gráfico de resumo SHAP. Apresenta os 20 genes com maior impacto médio nas previsões do modelo.

C. Análise por Subgrupo

Grupo 3:

O modelo identificou corretamente marcadores bem estabelecidos como *PPP2R2B*, *NPR3* e *GABRA5* como os mais

Tabela II: Resumo dos principais genes identificados pelo modelo SHAP e sua validação na literatura.

Gene	Subgrupo	Função	Específico do Grupo	Tipo de Estudo	Ref.
<i>PPP2R2B</i>	Grupo 3	Marcador para o Grupo 3; subunidade reguladora da PP2A.	Sim	Análise transcriptômica	[23]
<i>NPR3</i>	Grupo 3	Marcador imuno-histoquímico para o Grupo 3.	Sim	IHQ + revisão	[24]
<i>GABRA5</i>	Grupo 3	Altamente expresso no Grupo 3 impulsionado por MYC; receptor (α)5 – GABA _A .	Sim	Estudo funcional e de expressão gênica	[25]
<i>RBM24</i>	Grupo 4	Proteína de ligação a RNA; marcador prognóstico.	Sim	Transcriptômica	[26]
<i>ZBTB18</i>	Grupo 4	Repressor transcricional com expressão elevada no Grupo 4.	Sim	Análise transcriptômica	[27]
<i>SH3GL3</i>	Grupo 4	Forma RNAs circulares através de eventos de fusão no Grupo 4.	Sim	RNA-seq + validação	[28]
<i>OTX2</i>	SHH	Fator oncogênico superexpresso nos Grupos 3/4.	Não	Expressão gênica + número de cópias	[29], [30]
<i>CACNA1F</i>	SHH	Subunidade de canal de cálcio.	Não	Análise de vias em outros cânceres	[31]
<i>RPGRIP1</i>	SHH	Implicado na progressão de outros cânceres.	Não	Análise de expressão em outros tumores	[32]
<i>ATOH1</i>	SHH	Fator de transcrição mestre; condutor oncogênico essencial para o subgrupo SHH.	Sim	Estudo funcional e de expressão gênica	[33]
<i>PDLIM4</i>	WNT	Potencial supressor tumoral em outros cânceres.	Não	Estudo de metilação em outros tumores	[34]
<i>TMEM51</i>	WNT	Estudada no adenocarcinoma pancreático.	Não	Análise transcriptômica e funcional	[35]
<i>DLX3</i>	WNT	Modula a via Wnt em outros tecidos.	Não	Estudo funcional em outros tumores	[36]

importantes para este subgrupo. Esses genes são conhecidos por seu papel como marcadores transcriptômicos e imuno-histoquímicos ou por sua associação com a via de sinalização MYC, característica do Grupo 3 (Tabela II).

Grupo 4:

Para este subgrupo, os genes com maior valor SHAP foram *RBM24*, *ZBTB18* e *SH3GL3*. Esses achados estão alinhados com estudos recentes que propõem *RBM24* como um marcador prognóstico e *ZBTB18* e *SH3GL3* como genes com expressão ou atividade diferencial neste subgrupo (Tabela II).

SHH:

A análise do subgrupo SHH revelou uma lógica de classificação particularmente sofisticada. O modelo destacou genes como *OTX2*, *CACNA1F* e *RPGRIP1* como preditores-chave com base em um padrão contraintuitivo: a sua baixa expressão. No caso do *OTX2*, um oncogene conhecido pela alta expressão nos Grupos 3 e 4, a análise SHAP mostrou que sua relativa ausência no subgrupo SHH se tornou um poderoso marcador discriminativo negativo. De forma semelhante, *CACNA1F* e *RPGRIP1*, genes sem associação direta previamente estabelecida na literatura com o MB SHH, também foram importantes para o modelo por apresentarem baixa expressão. Em forte contraste, o marcador canônico *ATOH1* foi corretamente identificado por sua alta expressão, alinhando-se perfeitamente com seu papel biológico conhecido. Isso demonstra a capacidade do modelo de aprender regras de decisão contextuais, utilizando tanto a ausência de

marcadores de outros grupos quanto a presença de marcadores específicos do próprio subgrupo.

WNT:

Para o subgrupo WNT, os genes mais importantes foram *PDLIM4*, *TMEM51* e *DLX3*. Destes, *PDLIM4* e *TMEM51* carecem de associação direta com MB na literatura atual. No entanto, *DLX3* é conhecido por modular a via Wnt em outros contextos biológicos. Esses resultados sugerem que o modelo pode ter identificado potenciais biomarcadores novos, que merecem investigação experimental futura.

Em suma, a análise de explicabilidade demonstrou que o modelo não apenas atingiu alta acurácia, mas também fundamentou suas decisões em genes biologicamente relevantes, validando a abordagem computacional.

D. Eficiência do Modelo com Atributos Reduzidos

Para avaliar a viabilidade de um painel diagnóstico mais enxuto, foi conduzida uma análise iterativa utilizando um número decrescente de atributos, selecionados com base no ranking de importância SHAP. Os resultados indicam que o modelo mantém um desempenho robusto mesmo com uma redução drástica no número de genes. Notavelmente, utilizando apenas os 20 genes mais importantes, o modelo alcançou uma acurácia de 85,87%. Este achado sublinha a eficácia do SHAP para identificar um conjunto mínimo de biomarcadores de alto impacto e reforça o potencial para o desenvolvimento de modelos mais parcimoniosos e clinicamente aplicáveis no futuro.

V. DISCUSSÃO

Este estudo demonstrou a eficácia da combinação do modelo XGBoost com a análise SHAP para a classificação de subgrupos de MB e para a identificação de biomarcadores genéticos. A abordagem permitiu extrair padrões biologicamente interpretáveis de um conjunto de dados de alta dimensionalidade, alinhando o poder preditivo do aprendizado de máquina com a necessidade de transparência em pesquisas biomédicas [18].

A principal força do modelo reside na sua capacidade de validar marcadores bem documentados na literatura, como *PPP2R2B* para o Grupo 3 e *RBM24* para o Grupo 4, corroborando a validade biológica dos resultados [23], [26]. Contudo, a análise aprofundada da classificação do subgrupo SHH, guiada pela metodologia SHAP, revela a lógica complexa aprendida pelo modelo e a importância crucial da interpretabilidade para evitar conclusões apressadas.

Surpreendentemente, o gene mais importante para a classificação do subgrupo SHH foi o *OTX2*. No entanto, sua relevância não deriva de uma alta expressão, mas sim de sua baixa expressão nesta classe, que gera um alto impacto positivo no valor SHAP (Figura 4c). O *OTX2* é, na verdade, um fator oncogênico conhecido por sua superexpressão e papel na proliferação do MB dos Grupos 3 e 4 [29], [30], [37]. A hipótese mais provável, portanto, é que o modelo aprendeu uma regra de exclusão: embora a alta expressão de *OTX2* seja uma característica compartilhada pelos Grupos 3 e 4, sua ausência relativa torna-se um poderoso marcador negativo, altamente específico para a identificação precisa do subgrupo SHH. Isso sugere que o modelo não está apenas reconhecendo marcadores, mas sim aprendendo regras de decisão contextuais e discriminatórias.

Um padrão semelhante foi observado para os genes *CACNA1F* e *RPGRIP1*, que também foram destacados pelo modelo como importantes preditores do SHH devido à sua baixa expressão. A literatura científica não estabelece uma associação direta entre esses genes e o MB do subgrupo SHH. O *CACNA1F* está ligado a canais de cálcio, cuja desregulação é uma característica de vários tipos de câncer, mas sem vínculo direto comprovado com o MB [31]. Similarmente, a variante *RPGRIP1L* de *RPGRIP1* tem sido implicada em outros tumores, como o adenocarcinoma pancreático, mas sua função no MB permanece desconhecida [32]. A identificação desses genes pelo modelo representa, portanto, uma potencial geração de novas hipóteses, mas reforça que a importância computacional nem sempre se traduz em um papel biológico direto no subgrupo em questão.

Em contraste direto com esses achados, o gene *ATOH1*, o quarto preditor mais importante para SHH, exibe o comportamento biológico esperado: sua alta expressão está associada a um alto valor SHAP, impulsionando a classificação positiva para o subgrupo SHH. Isso está em perfeito acordo com a literatura, que estabelece o *ATOH1* como um fator de transcrição mestre e um condutor oncogênico essencial para a tumorigênese no MB do subgrupo SHH [33], [38]. A capacidade do modelo de identificar tanto os marcadores

contextuais (baixo *OTX2*) quanto os marcadores canônicos (alto *ATOH1*) sublinha a robustez e a profundidade da análise permitida pela XAI.

Além da identificação de marcadores individuais, o estudo demonstrou uma implicação prática com grande potencial translacional: a análise de redução de atributos indicou ser possível alcançar uma acurácia preditiva de 85,87% com apenas 20 genes. Por outro lado, os resultados para o subgrupo WNT, onde os genes identificados pelo SHAP carecem de associação direta com o MB, merecem uma análise cuidadosa, possivelmente refletindo o menor tamanho amostral ou, alternativamente, apontando para novos candidatos como *PDLIM4*, *TMEM51* e *DLX3* [34]–[36].

Finalmente, é fundamental ressaltar as limitações inerentes a este estudo. Como uma análise *in silico*, os resultados representam associações estatísticas e computacionais que requerem validação experimental (por exemplo, por qRT-PCR, imuno-histoquímica ou estudos funcionais) para confirmar o papel biológico e o potencial clínico dos genes identificados. O desempenho do modelo também é dependente da qualidade e da representatividade do conjunto de dados de treinamento.

VI. CONCLUSÃO

Este estudo demonstrou que a aplicação combinada do classificador XGBoost com a metodologia de interpretabilidade SHAP constitui uma estratégia robusta e eficaz para a análise de dados de expressão gênica em MB. A abordagem não só alcançou alta acurácia na classificação dos subgrupos moleculares, mas também, de forma crucial, permitiu a extração de insights biologicamente relevantes, validando a sua coerência através da identificação de biomarcadores já consagrados na literatura e descobrindo novos padrões de decisão.

A principal contribuição deste trabalho reside na demonstração da importância de interpretar criticamente o papel atribuído aos genes pelo modelo. A análise de explicabilidade revelou dois cenários distintos: genes cuja importância para o modelo reflete uma função biológica direta no subgrupo, como a alta expressão de *ATOH1* no subgrupo SHH; e genes cuja importância é contextual e discriminatória. O caso do *OTX2* é exemplar deste segundo cenário, onde sua baixa expressão se tornou um poderoso preditor para o subgrupo SHH, não por um papel biológico direto nesta classe, mas por ser um marcador de exclusão dos Grupos 3 e 4. Este padrão destaca que genes com alta importância computacional devem ser analisados com cautela, distinguindo entre relevância para a biologia do subtipo e relevância para a lógica de classificação do modelo. Por outro lado, a identificação de genes sem associação prévia na literatura, especialmente para o subgrupo WNT, aponta para o potencial da metodologia em gerar novas hipóteses e descobrir candidatos promissores para investigações futuras.

Como direcionamento para trabalhos futuros, a validação experimental dos achados é um passo fundamental. Sugere-se a quantificação da expressão gênica (qRT-PCR ou imuno-histoquímica) para os novos candidatos identificados, como *PDLIM4* e *TMEM51* no subgrupo WNT, e a confirmação

do perfil de baixa expressão de *OTX2* em amostras de SHH em comparação com os Grupos 3 e 4. Adicionalmente, o desenvolvimento de um painel diagnóstico reduzido, baseado nos genes mais informativos, representa uma promissora via de aplicação translacional. A aplicação desta metodologia a outros tumores pediátricos e a comparação com diferentes técnicas de XAI poderão aprofundar ainda mais nossa compreensão sobre a biologia do câncer guiada por dados.

AGRADECIMENTOS

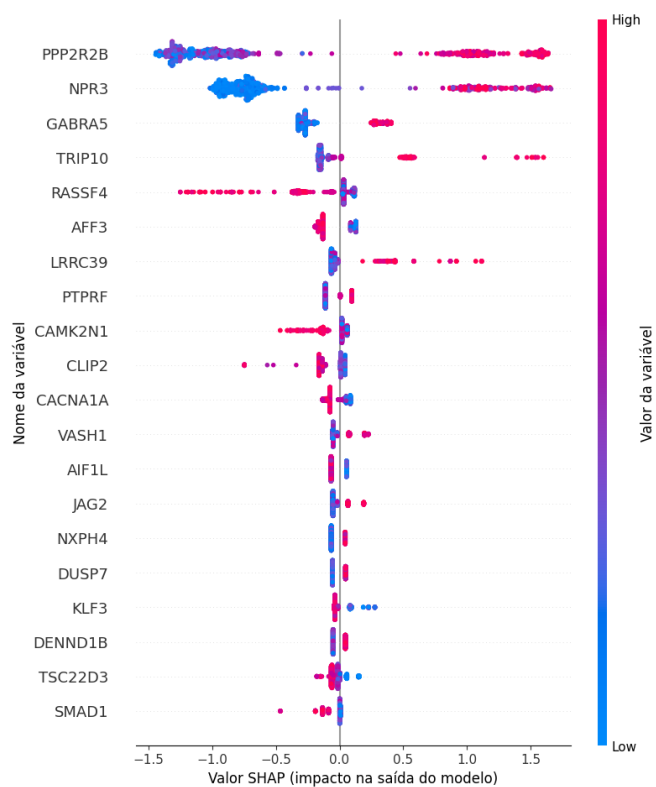
Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

REFERÊNCIAS

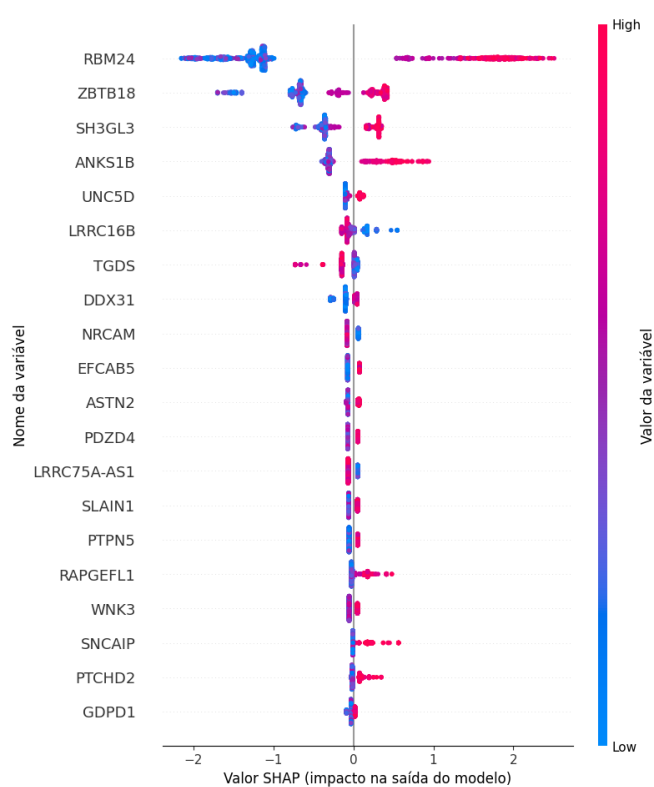
- [1] St. Jude Children's Research Hospital, "Medulloblastoma in children and teens," <https://together.stjude.org/en-us/conditions/cancers/medulloblastoma.html>, 2025, acessado em 6 de junho de 2025.
- [2] M. D. Taylor, P. A. Northcott, A. Korshunov, M. Remke, Y.-J. Cho, S. C. Clifford, C. G. Eberhart, D. W. Parsons, S. Rutkowski, A. Gajjar, R. J. Wechsler-Reya, S. L. Pomeroy, R. J. Gilbertson, and S. M. Pfister, "Molecular subgroups of medulloblastoma: the current consensus," *Acta Neuropathologica*, vol. 123, no. 4, pp. 465–472, 2012, acessado em 6 de junho de 2025.
- [3] A. N. Author, "Accurate identification of medulloblastoma subtypes from diverse data sources with severe batch effects by RaMBat," *PMC*, 2025, acessado em 6 de junho de 2025.
- [4] D. N. Louis, A. Perry, P. Wesseling, and etc., "The 2021 WHO classification of tumors of the central nervous system: a summary," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021, acessado em 6 de junho de 2025.
- [5] V. Vemireddy, J. A. Kalapurakal, A. D'Souza, S. Mohankumar, J. L. Finlay, and B. Rood, "Medulloblastoma: Challenges and advances in treatment and surveillance," *Cancer*, vol. 127, no. 8, pp. 1228–1240, 2021, acessado em 6 de junho de 2025.
- [6] A. N. Author, "Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics," *IU Indianapolis ScholarWorks*, 2023, acessado em 6 de junho de 2025.
- [7] Omics Tutorials, "Explainable AI for omics data," <https://omicstutorials.com/explainable-ai-for-omics-data/>, 2025, acessado em 6 de junho de 2025.
- [8] XGBoosting.com, "Explain XGBoost predictions with SHAP," <https://xgboosting.com/explain-xgboost-predictions-with-shap/>, 2025, acessado em 6 de junho de 2025.
- [9] A. N. Author, "Characterization of immune microenvironment associated with medulloblastoma metastasis based on explainable machine learning," *PubMed Central*, 2024, acessado em 6 de junho de 2025.
- [10] A. J. Gajjar, G. W. Robinson, K. S. Smith, and etc., "Outcomes by clinical and molecular features in children with medulloblastoma treated with risk-adapted therapy: Results of an international phase iii trial (SJMB03)," *Journal of Clinical Oncology*, vol. 39, no. 7, pp. 822–835, 2021, acessado em 6 de junho de 2025.
- [11] E. Sharif Rahmani, A. Lawarde, P. Lingasamy, S. V. Moreno, A. Salumets, and V. Modhukur, "Mbmethpred: a computational framework for the accurate classification of childhood medulloblastoma subgroups using data integration and ai-based approaches," *Frontiers in Genetics*, vol. 14, p. 1233657, 2023.
- [12] S. Gershanov, S. Madiwale, G. Feinberg-Gorenshtein, I. Vainer, T. Nehushtan, S. Michowiz, N. Goldenberg-Cohen, Y. Birger, H. Toledano, and M. Salmon-Divon, "Classifying medulloblastoma subgroups based on small, clinically achievable gene sets," *Frontiers in oncology*, vol. 11, p. 637482, 2021.
- [13] P. Rawal, D. Ahuja, M. Saini, H. Chandel, and R. Raj, "Cancer detection and treatment using explainable ai," *International Journal of Scientific Research in Multidisciplinary Studies*, vol. 3, no. 9, 2024.
- [14] D. Chakraborty, E. Gutierrez-Chakraborty, C. Rodriguez-Aguayo, H. Başağaoğlu, G. Lopez-Berestein, and P. Amero, "Discovering genetic biomarkers for targeted cancer therapeutics with explainable ai," *bioRxiv*, pp. 2023–07, 2023.
- [15] F. M. Cavalli, M. Remke, L. Rampasek, J. Peacock, D. J. Shih, B. Luu, L. Garzia, J. Torchia, C. Nor, A. S. Morrissy *et al.*, "Intertumoral heterogeneity within medulloblastoma subgroups," *Cancer cell*, vol. 31, no. 6, pp. 737–754, 2017.
- [16] B. Lan, "Leveraging random forest machine learning for subgroup classification of medulloblastoma," *Journal of Research in Social Science and Humanities*, vol. 3, no. 12, 2024.
- [17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016, p. 785–794. [Online]. Available: <http://dx.doi.org/10.1145/2939672.2939785>
- [18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
- [19] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 2010, pp. 51–56.
- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [21] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: <http://jmlr.org/papers/v18/16-365.html>
- [22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: <http://doi.acm.org/10.1145/2939672.2939785>
- [23] J. Vriend and R. B. Tate, "Differential expression of genes for ubiquitin ligases in medulloblastoma subtypes," *Cerebellum*, vol. 18, no. 3, pp. 469–488, 2019.
- [24] M. D. Taylor, P. A. Northcott, A. Korshunov, M. Remke, Y.-J. Cho, S. C. Clifford, C. G. Eberhart, D. W. Parsons, S. Rutkowski, A. Gajjar *et al.*, "Molecular subgroups of medulloblastoma: the current consensus," *Acta neuropathologica*, vol. 123, pp. 465–472, 2012.
- [25] S. Sengupta, S. D. Weeraratne, H. Sun, J. Phallen, S. K. Rallapalli, N. Teider, B. Kosaras, V. Amani, J. Pierre-Francois, Y. Tang *et al.*, "α5-GABAA receptors negatively regulate MYC-amplified medulloblastoma growth," *Acta neuropathologica*, vol. 127, pp. 593–603, 2014.
- [26] D.-L. Shi, "Rbm24 in the post-transcriptional regulation of cancer progression: Anti-tumor or pro-tumor activity?" *Cancers*, vol. 14, no. 7, p. 1843, 2022.
- [27] J. Vriend and M. Rastegar, "Ubiquitin ligases and medulloblastoma: genetic markers of the four consensus subgroups identified through transcriptome datasets," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1866, no. 10, p. 165839, 2020.
- [28] A. Azatyan and P. G. Zaphiropoulos, "Circular and fusion rnas in medulloblastoma development," *Cancers*, vol. 14, no. 13, p. 3134, 2022.
- [29] D. C. Adamson, Q. Shi, M. Wortham, P. A. Northcott, C. Di, C. G. Duncan, J. Li, R. E. McLendon, D. D. Bigner, M. D. Taylor, and H. Yan, "Otx2 is critical for the maintenance and progression of shh-independent medulloblastomas," *Cancer research*, vol. 70, no. 1, pp. 181–191, 2010.
- [30] Y. Lu, C. M. Labak, N. Jain, I. J. Purvis, M. R. Guda, S. E. Bach, A. J. Tsung, S. Asuthkar, and K. K. Velpula, "OTX2 expression contributes to proliferation and progression in Myc-amplified medulloblastoma," *American Journal of Cancer Research*, vol. 7, no. 7, p. 1556, 2017.
- [31] M. A. Khan, P. Pramanik, S. Marbut, F. Yunus, J. N. Galeas, S. Singh, A. P. Singh, and S. Dasgupta, "Abstract 3318: Frequent loss of CACNA1C, a calcium voltage-gated channel subunit is associated with lung adenocarcinoma progression and poor prognosis," *Cancer Research*, vol. 83, no. 7_Supplement, pp. 3318–3318, 2023.
- [32] J. Guo, Y. Yang, Z. Ji, M. Yao, X. Xia, X. Sha, and M. Huang, "Case Report: Novel RPGRIP1L Gene Mutations Identified by Whole Exome Sequencing in a Patient With Multiple Primary Tumors," *Frontiers in Genetics*, vol. 12, p. 620472, 2021.
- [33] K. B. Grausam, S. D. R. Dooyema, L. Bihannic, H. Premathilake, A. S. Morrissy, A. Forget, A. Schaefer, J. H. Gundelach, S. Macura, D. M.

Maher *et al.*, "ATOH1 Promotes Leptomeningeal Dissemination and Metastasis of Sonic Hedgehog Subgroup Medulloblastomas," *Cancer Research*, vol. 77, no. 20, pp. 5545–5557, 2017.

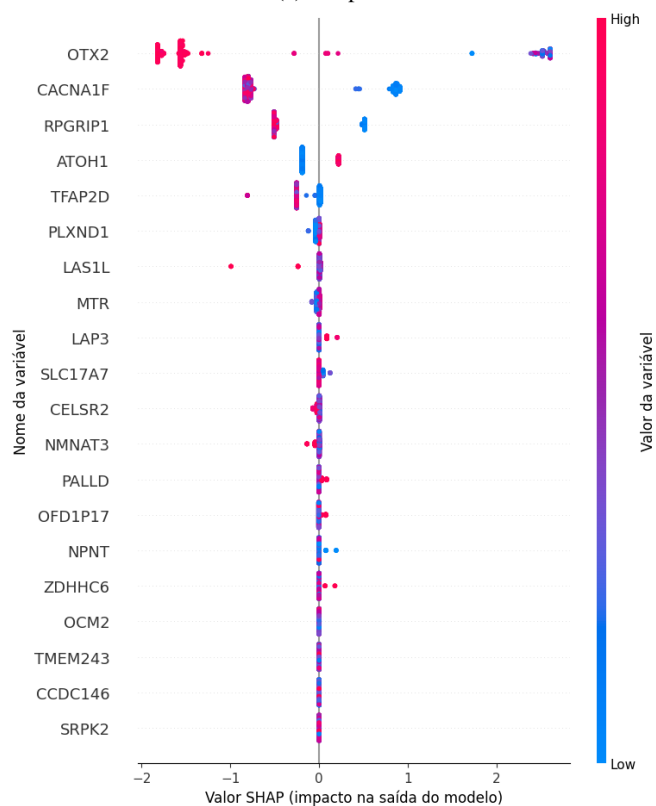
- [34] Y.-H. Hsiao, W.-Y. Lin, S.-H. Chen, C.-L. Chen, Y.-F. Lin, Y.-H. Hsieh, M.-T. Su, H.-C. Chu, H. Lin, P.-Y. Liew *et al.*, "Pdlim4, a novel pdz-lim domain protein, is a potential tumor suppressor in renal cell carcinoma," *BMC cancer*, vol. 10, pp. 1–11, 2010.
- [35] C. Ye, S. Ren, A. Sadula, X. Guo, M. Yuan, M. Meng, G. Li, X. Zhang, and C. Yuan, "The expression characteristics of transmembrane protein genes in pancreatic ductal adenocarcinoma through comprehensive analysis of bulk and single-cell rna sequence," *Frontiers in Oncology*, vol. 13, p. 1047377, 2023.
- [36] L.-L. E, L.-X. Zhang, J.-C. Zhang, Y.-Y. Gao, Y.-L. Wei, J. Li, J.-J. Zhu, X. Liu, and S.-Z. Liu, "Dlx3 regulates bone marrow-derived mesenchymal stem cell differentiation and migration by modulating the wnt/ β -catenin pathway," *Journal of cellular physiology*, vol. 234, no. 10, pp. 17 777–17 792, 2019.
- [37] R. T. LeGates, "Data from OTX2 Is Critical for the Maintenance and Progression of Shh-Independent Medulloblastomas," 2023.
- [38] K. M. Radke, "Data from ATOH1 Promotes Leptomeningeal Dissemination and Metastasis of Sonic Hedgehog Subgroup Medulloblastomas," 2023.



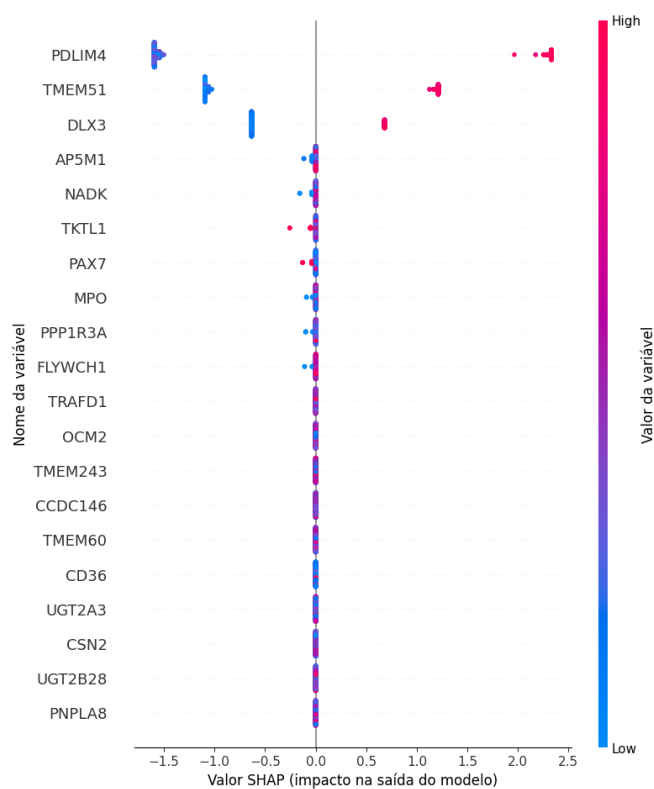
(a) Grupo 3



(b) Grupo 4



(c) SHH



(d) WNT

Figura 4: Gráficos de dependência SHAP para cada subgrupo molecular. Cada ponto representa uma amostra. A posição no eixo X indica o impacto do gene na previsão (positivo ou negativo), e a cor representa o nível de expressão do gene (alto em vermelho, baixo em azul).