

Diagnóstico Diferencial entre Deficiência do Hormônio do Crescimento e Baixa Estatura Idiopática por Aprendizado de Máquina Interpretável

Gabriel A. Abreu¹, Tessa M. S. Sasson², André C. P. L. F. Carvalho¹

¹Instituto de Ciências Matemáticas e de Computação, USP, São Carlos – SP – Brasil

²Faculdade de Medicina, USP, São Paulo – SP – Brasil

`gabriel.abreu@usp.br, tessasasson@hotmail.com, andre@icmc.usp.br`

Abstract. *Growth Hormone Deficiency (GHD) requires diagnosis based on clinical, laboratory, and neuroradiological data, often with invasive stimulation tests. This study evaluates machine learning models, including an interpretable approach, to differentiate GHD from Idiopathic Short Stature (ISS) in 297 anonymized patients from Hospital das Clínicas, University of São Paulo, comprising 146 with GHD and 151 with ISS. Explainable Boosting Machine (EBM), Random Forest, XGBoost, LightGBM, and CatBoost were tested in four attribute scenarios. In the detailed scenario, all models achieved accuracy above 0.90, and EBM provided competitive performance with interpretable feature contributions.*

Resumo. *A Deficiência do Hormônio do Crescimento (DGH) exige diagnóstico baseado em dados clínicos, laboratoriais e neurorradiológicos, frequentemente associado a testes provocativos invasivos. Este estudo avalia modelos de aprendizado de máquina, incluindo uma abordagem interpretável, na diferenciação entre DGH e Baixa Estatura Idiopática (BEI) em 297 pacientes anonimizados do Hospital das Clínicas da Universidade de São Paulo, sendo 146 com DGH e 151 com BEI. Explainable Boosting Machine (EBM), Random Forest, XGBoost, LightGBM e CatBoost foram testados em quatro cenários de atributos. No cenário apresentado em detalhe, todos alcançaram acurácia superior a 0,90, e o EBM apresentou desempenho competitivo com contribuições interpretáveis.*

1. Introdução

A deficiência do hormônio do crescimento (DGH) na infância é uma condição associada à redução da secreção de hormônio do crescimento (GH), com impacto no crescimento linear e no desenvolvimento infantil. Seu diagnóstico depende da integração de informações clínicas, auxológicas, bioquímicas e neurorradiológicas relacionadas ao eixo GH–fator de crescimento semelhante à insulina tipo 1 (IGF-1) [Molitch et al. 2011, Growth Hormone Research Society 2000]. Embora clinicamente relevante, a DGH apresenta baixa prevalência e parte dos casos permanece sem etiologia definida, o que dificulta sua caracterização diagnóstica [Thomas et al. 2004, Ranke and Wit 2018].

A baixa estatura é uma manifestação frequente da DGH, mas também ocorre em outras condições, como a Baixa Estatura Idiopática (BEI), tornando o diagnóstico diferencial desafiador. Testes provocativos de GH, amplamente utilizados nesse contexto, apresentam limitações de sensibilidade, especificidade, reprodutibilidade e potenciais efeitos

adversos [Ghigo et al. 1996, Grimberg et al. 2016]. Além disso, biomarcadores como o IGF-1, embora úteis na triagem inicial, podem ser influenciados por fatores fisiológicos, nutricionais e puberais, reduzindo sua capacidade discriminativa quando analisados isoladamente [Ranke and Wit 2018, Rosenfeld et al. 1995].

Nos últimos anos, técnicas de aprendizado de máquina têm sido investigadas como suporte a problemas diagnósticos complexos em saúde, por permitirem integrar múltiplas fontes de informação e identificar padrões não triviais em bases clínicas [Deo 2015, Clément et al. 2021, Murray et al. 2018, Ly 2025]. Em aplicações médicas, contudo, não basta obter alto desempenho preditivo: também é necessário compreender como as variáveis influenciam a decisão do modelo, favorecendo interpretação clínica e confiança no sistema.

Neste contexto, este trabalho investiga o uso de modelos de aprendizado de máquina, incluindo uma abordagem interpretável, para apoiar o diagnóstico diferencial entre DGH e BEI com base em dados clínicos reais do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFMUSP). Foram avaliados cinco algoritmos em diferentes cenários de atributos, considerando a presença ou ausência de variáveis de ressonância magnética (RNM) e de deficiências hormonais adicionais. Como contribuições, o estudo apresenta: (i) comparação entre modelos preditivos para o diagnóstico diferencial entre DGH e BEI; (ii) avaliação do impacto de diferentes conjuntos de atributos clínicos, auxológicos, laboratoriais, hormonais e neurorradiológicos; e (iii) análise de interpretabilidade baseada em Explainable Boosting Machine (EBM) e SHapley Additive exPlanations (SHAP).

2. Trabalhos Relacionados

Técnicas de aprendizado de máquina têm sido empregadas no apoio ao diagnóstico e manejo de distúrbios endócrinos complexos, em geral integrando dados clínicos, laboratoriais e de imagem para melhorar a capacidade discriminativa dos modelos.

No contexto da DGH, Murray et al. [Murray et al. 2018] investigaram o uso de aprendizado de máquina sobre dados transcriptômicos para prever a presença e a gravidade da doença. Song et al. [Song et al. 2024], por sua vez, combinaram características radiômicas de ressonância magnética com parâmetros clínicos, utilizando XGBoost para melhorar o desempenho diagnóstico.

Além do foco em desempenho, a literatura recente tem valorizado modelos mais transparentes. Ly [Ly 2025] avaliou o uso de modelos interpretáveis, como EBM, em decisões clínicas relacionadas ao tratamento com hormônio do crescimento. No contexto brasileiro, Sasson [Sasson 2022] propôs uma ferramenta baseada em inteligência artificial para auxiliar o diagnóstico de DGH a partir de dados clínicos e laboratoriais.

Apesar desses avanços, ainda são escassos os estudos voltados ao diagnóstico diferencial entre DGH e BEI com dados clínicos reais e ênfase simultânea em desempenho preditivo e interpretabilidade. O presente trabalho busca contribuir para essa lacuna ao comparar diferentes algoritmos de aprendizado de máquina, incluindo um modelo interpretável, em cenários com distintas composições de atributos clínicos, hormonais e neurorradiológicos.

3. Metodologia

Para avaliar o uso de modelos de aprendizado de máquina com análise de interpretabilidade no diagnóstico diferencial entre DGH e BEI, foi conduzido um estudo retrospectivo com dados clínicos reais. A metodologia incluiu a curadoria da base, a definição de cenários experimentais, o treinamento e a avaliação dos modelos, além da análise de interpretabilidade das previsões.

3.1. Base de Dados

Foram utilizados dados retrospectivos anonimizados de 297 pacientes acompanhados no ambulatório de endocrinologia do HCFMUSP, incluindo 146 pacientes com DGH e 151 com BEI. A coorte contemplou variáveis clínicas, auxológicas, laboratoriais e radiológicas, com inclusão de pacientes acompanhados no serviço especializado e exclusão de registros com informações insuficientes para análise.

A organização e curadoria das variáveis foram realizadas em colaboração com a equipe clínica, garantindo a coerência entre o significado médico dos atributos e sua aplicação na modelagem computacional. O estudo foi conduzido em conformidade com os princípios éticos para pesquisa envolvendo seres humanos e aprovado pelo Comitê de Ética em Pesquisa do HCFMUSP em 11/08/2023.

3.2. Cenários Experimentais

Foram definidos quatro cenários para investigar o impacto de diferentes grupos de atributos no desempenho dos modelos: (i) com deficiência hormonal adicional e com variáveis de RNM, totalizando 37 variáveis; (ii) sem deficiência hormonal adicional e sem RNM, com 28 variáveis; (iii) com deficiência hormonal adicional e sem RNM, com 33 variáveis; e (iv) sem deficiência hormonal adicional e com RNM, com 32 variáveis. Neste estudo, a expressão deficiência hormonal adicional refere-se à presença de outras alterações hormonais associadas além da deficiência de GH.

Essa estratégia permitiu avaliar o desempenho global dos algoritmos e a contribuição potencial de informações neurorradiológicas e de comorbidades hormonais no diagnóstico diferencial entre DGH e BEI.

3.3. Treinamento dos Modelos

O pré-processamento incluiu a identificação e o tratamento de valores ausentes, a recodificação de variáveis categóricas e a codificação binária do diagnóstico. A base foi dividida de forma estratificada em treinamento e teste, na proporção 70/30. Após comparação entre estratégias de imputação, foi adotado o KNN com $k = 6$, ajustado apenas no conjunto de treinamento e aplicado posteriormente aos conjuntos de treinamento e teste. Ao final, variáveis categóricas imputadas foram arredondadas para preservar sua representação discreta.

Foram avaliados cinco algoritmos de aprendizado de máquina: EBM, Random Forest, XGBoost, LightGBM e CatBoost. Para cada modelo, foi realizado ajuste de hiperparâmetros por *Grid Search* com validação cruzada estratificada no conjunto de treinamento. A avaliação final foi realizada no conjunto independente de teste, com cálculo de métricas de classificação. Além do desempenho preditivo, foi conduzida análise de interpretabilidade com foco no EBM, complementada por análises baseadas em SHAP.

4. Resultados

Os modelos avaliados apresentaram desempenho elevado na diferenciação entre pacientes com DGH e BEI. Os experimentos foram conduzidos em quatro cenários de atributos, com diferentes combinações de variáveis clínicas, auxológicas, laboratoriais, hormonais e neurorradiológicas. A Tabela 1 apresenta, em detalhe, as métricas obtidas no cenário *Com Deficiência Hormonal Adicional — Com RNM*.

Tabela 1. Métricas de desempenho dos modelos no conjunto de teste para o cenário *Com Deficiência Hormonal Adicional — Com RNM*.

Modelo	AUC	Acurácia	Sensibilidade	Especificidade	VPP	VPN
Random Forest	0.996	0.989	0.977	1.000	1.000	0.979
XGBoost	0.997	0.978	0.977	0.978	0.977	0.978
CatBoost	0.996	0.956	0.955	0.957	0.955	0.957
LightGBM	0.996	0.967	0.977	0.957	0.956	0.978
EBM	0.991	0.978	0.955	1.000	1.000	0.958

AUC: área sob a curva, do inglês *Area Under the Curve*, calculada a partir da curva ROC (*Receiver Operating Characteristic*); VPP: valor preditivo positivo; VPN: valor preditivo negativo.

A análise da matriz de confusão indicou baixa ocorrência de erros de classificação no conjunto de teste. Considerando DGH como classe positiva, os modelos destacados — Random Forest, pela maior acurácia, e EBM, pela interpretabilidade — apresentaram baixa taxa de erro: o Random Forest obteve 43 verdadeiros positivos, 46 verdadeiros negativos, 1 falso negativo e nenhum falso positivo, enquanto o EBM apresentou 42 verdadeiros positivos, 46 verdadeiros negativos, 2 falsos negativos e nenhum falso positivo. Nesse cenário, todos os modelos alcançaram AUC entre 0,991 e 0,997, com equilíbrio entre sensibilidade e especificidade. O Random Forest obteve a maior acurácia, enquanto o EBM apresentou resultado competitivo, mantendo a vantagem de interpretabilidade.

4.1. Interpretabilidade

Além da avaliação preditiva, foi realizada análise de interpretabilidade para compreender como variáveis clínicas, auxológicas, laboratoriais e neurorradiológicas contribuíram para as decisões do modelo. Para essa etapa, foi selecionado o EBM, por combinar desempenho competitivo e transparência estrutural.

No cenário *Com Deficiência Hormonal Adicional — Com RNM*, a análise do EBM permitiu investigar a importância global das variáveis e os principais termos de interação aprendidos. A Figura 1 apresenta, à esquerda, o gráfico de resumo dos valores SHAP e, à direita, o ranqueamento das dez interações mais relevantes identificadas pelo modelo.

Entre as interações mais relevantes destacaram-se *zaltura_inic & haste*, *zdeltaTH & neuro* e *zaltura_inic & neuro*, além de combinações envolvendo *idade_inicial*, *IGF1_interpret*, *tipo_parto*, *zvc_inic* e *adeno*. Esses achados sugerem que o modelo se apoia em relações conjuntas entre atributos auxológicos, hormonais e neurorradiológicos, o que é coerente com a complexidade do diagnóstico diferencial entre DGH e BEI.

5. Discussão

Os resultados indicam que a integração de variáveis clínicas, auxológicas, laboratoriais, hormonais e neurorradiológicas permite elevada discriminação entre DGH e BEI. A

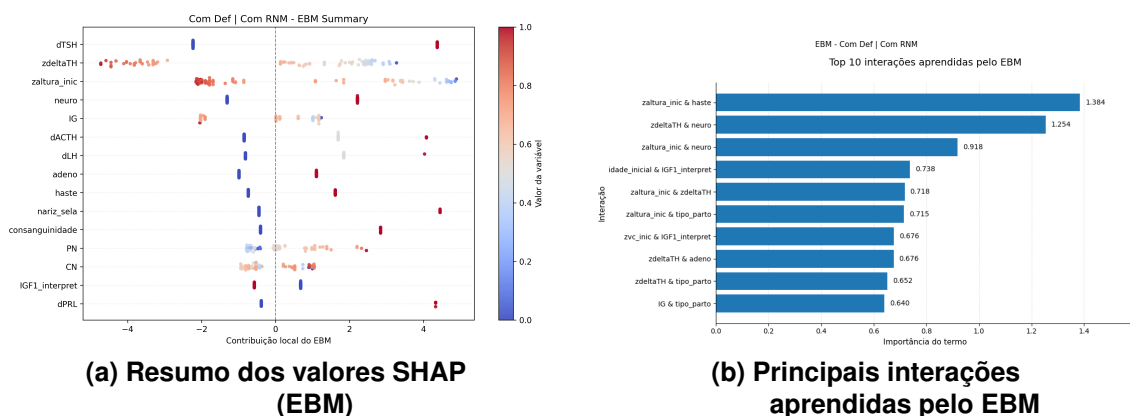


Figura 1. Análise de interpretabilidade do modelo EBM no cenário *Com Deficiência Hormonal Adicional — Com RNM*.

avaliação em diferentes cenários permitiu investigar a contribuição de variáveis de RNM e de deficiências hormonais adicionais, sendo que, no cenário apresentado em detalhe, essa combinação esteve associada a elevado desempenho preditivo.

Os métodos baseados em *ensemble*, especialmente Random Forest e XGBoost, estiveram entre os modelos de melhor desempenho. Ainda assim, o EBM obteve desempenho competitivo, o que é relevante em aplicações clínicas, nas quais desempenho e interpretabilidade devem ser considerados conjuntamente. As análises baseadas em SHAP e nas interações do EBM evidenciaram maior relevância de variáveis auxológicas, hormonais e neurorradiológicas, em concordância com o conhecimento clínico da endocrinologia pediátrica.

Apesar dos resultados promissores, algumas limitações devem ser consideradas. Trata-se de um estudo retrospectivo e unicêntrico, baseado em pacientes de um hospital terciário, o que pode limitar a generalização dos modelos. Além disso, a quantidade de dados disponível ainda é restrita, especialmente diante da divisão em diferentes cenários experimentais, e aspectos relacionados ao caráter unicêntrico da coleta, à completude dos dados, à presença de valores ausentes, à heterogeneidade clínica e à variabilidade de preenchimento dos prontuários podem ter influenciado os resultados. Ainda não foi realizada validação externa em bases independentes, embora essa etapa esteja prevista para ser conduzida em breve, a fim de avaliar a robustez e a generalização dos modelos. Por fim, outros algoritmos ou estratégias de modelagem não considerados neste estudo poderão ser explorados em trabalhos futuros.

6. Conclusão

Este trabalho avaliou modelos de aprendizado de máquina para o diagnóstico diferencial entre DGH e BEI, com base em dados clínicos reais. Os resultados indicaram elevado desempenho preditivo, especialmente no cenário apresentado em detalhe, no qual o Random Forest obteve a maior acurácia e o EBM apresentou desempenho competitivo aliado à interpretabilidade.

As análises de explicabilidade sugeriram papel relevante de variáveis auxológicas, hormonais e neurorradiológicas nas decisões do modelo, reforçando a coerência dos achados com o conhecimento clínico. Como continuidade, pretende-se ampliar o conjunto de dados, realizar validação externa em base independente, aprofundar as análises de in-

terpretabilidade e investigar a viabilidade de um protótipo computacional para apoio ao diagnóstico diferencial entre DGH e BEI.

Referências

- Clément, F., Grinspon, R. P., Yankelevich, D., et al. (2021). Development and validation of a prediction rule for growth hormone deficiency. *Frontiers in Endocrinology*, 11:624684.
- Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20):1920–1930.
- Ghigo, E., Bellone, J., Aimaretti, G., et al. (1996). Reliability of provocative tests to assess growth hormone secretory status. *Journal of Clinical Endocrinology and Metabolism*, 81(9):3323–3327.
- Grimberg, A., DiVall, S. A., Polychronakos, C., et al. (2016). Guidelines for growth hormone and insulin-like growth factor-i treatment in children and adolescents. *Hormone Research in Paediatrics*, 86(6):361–397.
- Growth Hormone Research Society (2000). Consensus guidelines for the diagnosis and treatment of growth hormone deficiency in childhood and adolescence. *Journal of Clinical Endocrinology and Metabolism*, 85(11):3990–3993.
- Ly, H. (2025). Applying prediction models and ai to optimize growth hormone therapy in children with short stature. *University of Gothenburg*.
- Molitch, M. E., Clemmons, D. R., Malozowski, S., Merriam, G. R., Vance, M. L., and Society, E. (2011). Evaluation and treatment of adult growth hormone deficiency: an endocrine society clinical practice guideline. *Journal of Clinical Endocrinology and Metabolism*, 96(6):1587–1609.
- Murray, P. G., Stevens, A., Leonibus, C. D., Koledova, E., Chatelain, P., and Clayton, P. E. (2018). Transcriptomics and machine learning predict diagnosis and severity of growth hormone deficiency. *JCI Insight*, 3(7):e93247.
- Ranke, M. B. and Wit, J. M. (2018). Growth hormone – past, present and future. *Nature Reviews Endocrinology*, 14(5):285–300.
- Rosenfeld, R. G., Albertsson-Wikland, K., Cassorla, F., et al. (1995). Diagnostic controversy: the diagnosis of childhood growth hormone deficiency revisited. *Journal of Clinical Endocrinology and Metabolism*, 80(5):1532–1540.
- Sasson, T. M. S. (2022). Desenvolvimento e aplicação de ferramenta baseada em algoritmos de inteligência artificial para diagnóstico de deficiência do hormônio do crescimento. Projeto de doutorado, Faculdade de Medicina da Universidade de São Paulo.
- Song, K., Ko, T., Chae, H. W., et al. (2024). Development and validation of a prediction model using mri-based radiomics. *Journal of Medical Internet Research*, 26:e54641.
- Thomas, M., Massa, G., Craen, M., et al. (2004). Prevalence and demographic features of childhood growth hormone deficiency in belgium during the period 1986–2001. *European Journal of Endocrinology*, 151(1):67–72.