RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.

Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

Predição de desfechos clínicos em câncer de mama luminal B por meio de modelos de machine learning baseado em variantes somáticas

Machine learning-based prediction of clinical outcomes in luminal b breast cancer using somatic variant signatures

Amanda Razera– Universidade Estadual do Centro-Oeste

Maiara Luiza Biava Miri – Centro Universitário Campo Real

Eduardo de Almeida Ravarena – Centro Universitário Campo Real

Gabryela Paulista Mateucci – Centro Universitário Campo Real

Camila Padilha Duda – Centro Universitário Campo Real

Resumo

A aplicação de técnicas de machine learning na oncologia tem possibilitado a integração de dados genômicos complexos para a predição de desfechos clínicos. No câncer de mama subtipo Luminal B, a elevada heterogeneidade tumoral representa um desafio para a estratificação prognóstica e definição terapêutica. Este estudo teve como objetivo desenvolver modelos preditivos baseados em variantes somáticas para avaliação de agressividade tumoral. Foi implementado um pipeline de machine learning utilizando algoritmos supervisionados, incluindo XGBoost, Support Vector Machine e Redes Neurais Artificiais. A seleção de variáveis foi realizada com base em importância preditiva, priorizando genes com relevância biológica. O desempenho dos modelos foi avaliado por meio de métricas como área sob a curva ROC, sensibilidade, especificidade e escore F1. Os resultados demonstraram elevada capacidade preditiva, com destaque para o modelo XGBoost (AUC = 0,88), seguido por Redes Neurais (AUC = 0,87) e SVM (AUC = 0,85). A análise de interpretabilidade indicou que genes como PIK3CA, TP53 e ERBB2 foram os principais determinantes das predições. Os achados reforçam o potencial do uso de machine learning na identificação de padrões genômicos associados à agressividade tumoral, contribuindo para estratégias de medicina de precisão.

Palavras-chave: Machine learning. Câncer de mama. Variantes Somáticas. Prognóstico. Genômica.

Abstract

The application of machine learning techniques in oncology has enabled the integration of complex genomic data for predicting clinical outcomes. In Luminal B breast cancer, high tumor heterogeneity poses a major challenge for prognostic stratification and therapeutic decision-making. This study aimed to develop predictive models based on somatic variants to assess tumor aggressiveness. A machine learning pipeline was implemented using supervised algorithms, including XGBoost, Support Vector Machine, and Artificial Neural Networks. Feature selection was based on predictive importance, prioritizing biologically relevant genes. Model performance was evaluated using metrics such as area under the ROC curve, sensitivity, specificity, and F1-score. The results demonstrated high predictive performance, with XGBoost achieving the best results (AUC = 0.88), followed by Neural Networks (AUC = 0.87) and SVM (AUC = 0.85). Interpretability analysis revealed that genes such as PIK3CA, TP53, and ERBB2 were the main contributors to model predictions. These findings highlight the potential of machine learning approaches in identifying genomic patterns associated with tumor aggressiveness, supporting precision medicine strategies.

Keywords: machine learning. breast cancer. somatic variants. prognosis. Genomics

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

1. Introdução

O câncer de mama permanece como uma das principais causas de morbimortalidade entre mulheres em nível global, sendo caracterizado por elevada heterogeneidade biológica, clínica e molecular (SUNG et al., 2021; BRASIL, 2024). A evolução do conhecimento em biologia molecular permitiu a classificação dos tumores em subtipos intrínsecos, como luminal A, Luminal B, HER2-enriquecido e triplo-negativo, contribuindo significativamente para o refinamento prognóstico e terapêutico (PEROU et al., 2000; SØRLIE et al., 2001).

Dentre esses, o subtipo Luminal B apresenta maior agressividade clínica, maior índice proliferativo e maior risco de recorrência quando comparado ao subtipo Luminal A, refletindo sua complexidade molecular e comportamento clínico heterogêneo (PRAT et al., 2015; BURSTEIN et al., 2014). Essa heterogeneidade representa um desafio relevante na prática clínica, especialmente no que se refere à estratificação prognóstica e à definição de estratégias terapêuticas personalizadas.

Com o avanço das tecnologias de sequenciamento de nova geração, tornou-se possível identificar variantes somáticas relevantes na carcinogênese mamária, especialmente em genes envolvidos em vias de proliferação celular, reparo de DNA e sinalização intracelular (CANCER GENOME ATLAS NETWORK, 2012). Genes como PIK3CA, TP53 e ERBB2 têm sido amplamente descritos como determinantes da progressão tumoral e resposta terapêutica, sendo considerados alvos centrais na oncologia de precisão (ANDRÉ et al., 2019; SILWAL-PANDIT et al., 2017).

Entretanto, a crescente disponibilidade de dados genômicos de alta dimensão impõe limitações aos métodos estatísticos tradicionais, dificultando a identificação de padrões complexos associados a desfechos clínicos. Nesse cenário, técnicas de machine learning têm emergido como ferramentas promissoras, permitindo a análise integrada de múltiplas variáveis e a construção de modelos preditivos mais robustos (KOURI et al., 2020; ESTEVA et al., 2019).

Estudos recentes demonstram que algoritmos de aprendizado de máquina são capazes de integrar dados genômicos e clínicos para predizer prognóstico, resposta terapêutica e risco de progressão tumoral, com desempenho superior aos modelos convencionais em diferentes contextos oncológicos (TOPOL, 2019; LIBBRECHT; NOBLE, 2015). Além disso, abordagens interpretáveis, como os métodos baseados em valores SHAP, têm possibilitado maior transparência na compreensão da contribuição individual das variáveis, ampliando a aplicabilidade clínica desses modelos (LUNDBERG; LEE, 2017).

Diante desse contexto, o desenvolvimento de modelos preditivos baseados em variantes somáticas e técnicas de machine learning representa uma estratégia promissora para aprimorar a estratificação prognóstica no câncer de mama Luminal B. Assim, o presente estudo propõe a construção e avaliação de modelos de aprendizado de máquina capazes de identificar padrões genômicos associados à agressividade tumoral, contribuindo para o avanço da medicina de precisão.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

2. Material e método

2.1 Delineamento do estudo

Trata-se de um estudo observacional com abordagem analítica, no qual foi desenvolvido um modelo preditivo baseado em técnicas de machine learning para avaliação de desfechos clínicos a partir de assinaturas genômicas. A estratégia metodológica foi estruturada conforme diretrizes para estudos envolvendo modelos preditivos em saúde (COLLINS et al., 2015).

2.2 Pipeline de machine learning

Foi implementado um pipeline de aprendizado de máquina composto por etapas de pré-processamento, seleção de variáveis, treinamento de modelos e avaliação de desempenho.

O pré-processamento incluiu padronização das variáveis e tratamento de dados ausentes por métodos de imputação apropriados, baseados em proximidade estatística. A seleção de variáveis foi realizada por meio de importância preditiva, combinando regularização L1 e análise de importância permutacional, priorizando atributos com relevância biológica.

2.3 Modelos preditivos

Foram avaliados três algoritmos supervisionados amplamente utilizados em bioinformática - Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM) e Redes Neurais Artificiais (RNA).

A otimização de hiperparâmetros foi realizada por busca Bayesiana com validação cruzada aninhada (nested cross-validation), visando reduzir o risco de overfitting e aumentar a generalização dos modelos (BERGSTRA et al., 2011).

2.4 Avaliação de desempenho

O desempenho dos modelos foi avaliado em conjunto independente por meio das métricas área sob a curva ROC (AUC), sensibilidade, especificidade, valor preditivo positivo (PPV), valor preditivo negativo (NPV) e escore F1.

Considerando o possível desbalanceamento das classes, foram utilizadas métricas adicionais baseadas em precisão-revocação.

2.5 Interpretabilidade dos modelos

A interpretabilidade dos modelos foi avaliada por meio da metodologia SHAP (Shapley Additive Explanations), permitindo quantificar a contribuição individual das variáveis para as predições (LUNDBERG; LEE, 2017). Essa abordagem possibilitou identificar os principais determinantes genômicos associados aos desfechos analisados.

2.6 Análise estatística e software

As análises foram realizadas em ambiente computacional utilizando a linguagem R (versão

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

4.2.1) e Python (versão 3.9), com utilização de bibliotecas específicas para modelagem estatística e machine learning.

Os resultados foram considerados estatisticamente significativos quando $p < 0,05$, com aplicação de correção para múltiplas comparações quando apropriado.

3. Resultados e discussão

A aplicação de modelos de machine learning para análise de assinaturas genômicas demonstrou elevada capacidade preditiva na identificação de padrões associados à agressividade tumoral no câncer de mama subtipo Luminal B. Entre os algoritmos avaliados, o modelo baseado em Extreme Gradient Boosting (XGBoost) apresentou o melhor desempenho global, com AUC-ROC de 0,88, seguido por Redes Neurais Artificiais (0,87) e Support Vector Machine (0,85), conforme apresentado na Tabela 1.

Esse desempenho superior do XGBoost pode ser atribuído à sua capacidade de modelar interações não lineares e capturar relações complexas entre variáveis de alta dimensionalidade, característica intrínseca aos dados genômicos. Além disso, algoritmos baseados em ensemble tendem a apresentar maior robustez frente a ruído e variabilidade biológica, aspectos frequentemente observados em estudos de câncer (TOPOL, 2019).

Tabela 1 – Desempenho dos modelos de machine learning.

Modelo	AUC-ROC	Sensibilidade	Especificidade	PPV	NPV	F1-score
XGBoost	0,88	0,82	0,83	0,81	0,84	0,82
RNA	0,87	0,81	0,82	0,80	0,83	0,81
SVM	0,85	0,79	0,80	0,78	0,81	0,79

Fonte: Autores (2026).

A manutenção de valores equilibrados de sensibilidade e especificidade entre os modelos sugere boa capacidade discriminativa e estabilidade preditiva, mesmo em cenários de potencial desbalanceamento de classes. Esse achado é particularmente relevante em oncologia, onde eventos clínicos adversos frequentemente apresentam distribuição desigual.

A análise de importância das variáveis evidenciou que genes como PIK3CA, TP53 e ERBB2 foram os principais determinantes das predições, conforme demonstrado na Tabela 2.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

Tabela 2 – Importância das variáveis (análise SHAP).

Variável	Importância relativa (SHAP)	Função biológica principal
PIK3CA	0,32	Via PI3K/AKT/mTOR – proliferação celular
TP53	0,28	Controle do ciclo celular e apoptose
ERBB2	0,25	Sinalização de crescimento tumoral
Ki-67	0,20	Índice proliferativo
TFRC	0,15	Metabolismo do ferro e crescimento tumoral

Fonte: Autores (2026).

A relevância dessas variáveis reforça a coerência biológica do modelo, uma vez que esses genes são amplamente reconhecidos como centrais na carcinogênese mamária e na resistência terapêutica (CANCER GENOME ATLAS NETWORK, 2012; ANDRÉ et al., 2019).

Além disso, a contribuição do índice proliferativo (Ki-67) destaca a importância da integração entre dados genômicos e parâmetros fenotípicos, ampliando a capacidade preditiva dos modelos. Esse achado está alinhado com evidências que apontam o Ki-67 como marcador prognóstico relevante no subtipo Luminal B (PRAT et al., 2015).

A análise integrada também permitiu identificar padrões genômicos associados a diferentes níveis de risco, evidenciando que a interação entre variantes é determinante para o comportamento tumoral (Tabela 3).

Tabela 3 – Padrões genômicos associados ao risco predito.

Padrão genético	Interpretação biológica	Associação com agressividade
PIK3CA mutado isolado	Ativação de via proliferativa	Moderada
ERBB2 alterado	Sinalização de crescimento aumentada	Alta
TP53 mutado	Instabilidade genômica	Alta
PIK3CA + ERBB2	Sinergia proliferativa	Muito alta
PIK3CA + TP53	Proliferação + falha de controle	Muito alta

Fonte: Autores (2026).

Esses achados reforçam o conceito de que a progressão tumoral resulta de redes complexas de interação molecular, e não de alterações isoladas. Nesse sentido, modelos de machine learning apresentam vantagem significativa ao capturar essas interações de forma integrada, superando limitações dos modelos estatísticos tradicionais (LIBBRECHT; NOBLE, 2015).

Outro ponto relevante refere-se à utilização de métodos de interpretabilidade, como SHAP, que permitiram compreender a contribuição individual das variáveis para as predições. Essa abordagem representa um avanço importante, uma vez que reduz a opacidade dos modelos e favorece sua aplicabilidade clínica, especialmente em contextos que demandam transparência na tomada de decisão (LUNDBERG; LEE, 2017).

Além disso, os resultados obtidos sugerem que a aplicação de inteligência artificial pode

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

atuar como ferramenta complementar às estratégias tradicionais de estratificação prognóstica, contribuindo para a personalização terapêutica. Essa abordagem pode ser particularmente útil em cenários com acesso limitado a painéis multigênicos comerciais, ampliando a equidade no cuidado oncológico.

Apesar dos resultados promissores, algumas limitações devem ser consideradas. Primeiramente, modelos de machine learning aplicados a dados genômicos estão sujeitos ao risco de overfitting, especialmente quando o número de variáveis excede o número de observações. Embora estratégias como validação cruzada tenham sido empregadas, a validação externa em coortes independentes é essencial para confirmar a robustez dos achados.

Outra limitação refere-se à ausência de integração de dados multiômicos, como transcriptômica, epigenômica e proteômica, que poderiam ampliar a capacidade preditiva dos modelos. A inclusão desses dados pode capturar camadas adicionais de regulação biológica relevantes para a progressão tumoral.

Adicionalmente, fatores clínicos e ambientais, que também influenciam o comportamento tumoral, não foram explorados de forma aprofundada neste recorte analítico. A integração desses elementos em modelos futuros pode contribuir para uma abordagem mais abrangente e translacional.

Por fim, embora técnicas de interpretabilidade tenham sido utilizadas, a aplicação clínica desses modelos ainda depende de validação prospectiva e de sua incorporação em fluxos de decisão médica, o que representa um desafio atual na implementação de inteligência artificial em saúde.

Como perspectivas, Estudos futuros devem focar na validação externa dos modelos em diferentes populações e na integração de dados multiômicos e clínicos, visando aumentar a acurácia e generalização das predições. Além disso, o desenvolvimento de modelos híbridos, combinando machine learning e conhecimento biológico prévio, pode representar uma estratégia promissora para aprimorar a interpretabilidade e aplicabilidade clínica.

Considerações finais

Os resultados deste estudo demonstram que a aplicação de modelos de machine learning baseados em assinaturas genômicas apresenta elevado potencial para a predição de desfechos clínicos no câncer de mama subtipo Luminal B. A capacidade desses modelos em integrar múltiplas variáveis e capturar interações complexas entre vias moleculares permite uma abordagem mais refinada da heterogeneidade tumoral, superando limitações de métodos tradicionais baseados em variáveis isoladas.

A identificação consistente de genes centrais na carcinogênese mamária, como PIK3CA, TP53 e ERBB2, reforça a validade biológica do modelo proposto e evidencia a relevância da incorporação de dados genômicos na construção de ferramentas preditivas. Além disso, a utilização

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

de técnicas de interpretabilidade contribui para a transparência dos modelos, favorecendo seu potencial de aplicação em contextos clínicos.

Os achados também destacam o papel emergente da inteligência artificial como ferramenta complementar na oncologia, com potencial para contribuir na estratificação prognóstica e no suporte à tomada de decisão terapêutica, especialmente em cenários com acesso limitado a testes moleculares avançados.

Entretanto, a consolidação dessas abordagens na prática clínica depende de validação em coortes independentes, bem como da integração com dados clínicos e multiômicos. Dessa forma, estudos futuros são fundamentais para ampliar a robustez e a aplicabilidade dos modelos desenvolvidos.

Em síntese, a integração entre genômica tumoral e técnicas de machine learning representa uma estratégia promissora no avanço da medicina de precisão, com potencial para impactar significativamente o manejo clínico do câncer de mama.

Referências

ANDRÉ, F. et al. *Alpelisib for PIK3CA-mutated, hormone receptor–positive advanced breast cancer*. New England Journal of Medicine, v. 380, n. 20, p. 1929–1940, 2019.

BURSTEIN, H. J. et al. *Estimating the benefits of therapy for early-stage breast cancer: the St. Gallen International Consensus Guidelines*. Annals of Oncology, v. 25, n. 10, p. 1871–1888, 2014.

CANCER GENOME ATLAS NETWORK. *Comprehensive molecular portraits of human breast tumours*. Nature, v. 490, n. 7418, p. 61–70, 2012.

COLLINS, G. S. et al. *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement*. Annals of Internal Medicine, v. 162, n. 1, p. 55–63, 2015.

ESTEVA, A. et al. *A guide to deep learning in healthcare*. Nature Medicine, v. 25, p. 24–29, 2019.

KOURI, A. et al. *Artificial intelligence in oncology: current applications and future directions*. CA: A Cancer Journal for Clinicians, v. 70, n. 4, p. 268–287, 2020.

LIBBRECHT, M. W.; NOBLE, W. S. *Machine learning applications in genetics and genomics*. Nature Reviews Genetics, v. 16, n. 6, p. 321–332, 2015.

LUNDBERG, S. M.; LEE, S.-I. *A unified approach to interpreting model predictions*. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS. [S. l.]: NIPS, 2017. p. 4765–4774.

PEROU, C. M. et al. *Molecular portraits of human breast tumours*. Nature, v. 406, p. 747–752, 2000.

PRAT, A. et al. *Prognostic significance of Ki67 in breast cancer*. Journal of Clinical Oncology, v. 33, n. 36, p. 4234–4242, 2015.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Ano VI, v.1 2026 | submissão: 15/03/2026 | aceito: 17/03/2026 | publicação: 19/03/2026

SILWAL-PANDIT, L. et al. *TP53 mutation spectrum in breast cancer*. Breast Cancer Research, v. 19, n. 1, p. 1–14, 2017.

SUNG, H. et al. *Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide*. CA: A Cancer Journal for Clinicians, v. 71, n. 3, p. 209–249, 2021.

TOPOL, E. J. *High-performance medicine: the convergence of human and artificial intelligence*. Nature Medicine, v. 25, p. 44–56, 2019.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença CreativeCommons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.