Year VI, v.1 2026 | Submission: 03/15/2026 | Accepted: 03/17/2026 | Publication: 03/19/2026

Prediction of clinical outcomes in luminal B breast cancer using somatic variant signatures .

Amanda Razera – State University of the Midwest

Maiara Luiza Biava Miri – Campo Real University Center

Eduardo de Almeida Ravarena – Campo Real University Center

Gabryela Paulista Mateucci – Campo Real University Center

Camila Padilha Duda – Campo Real University Center

Summary

The application of machine learning techniques in oncology has enabled the integration of complex genomic data for the prediction of clinical outcomes. In Luminal B subtype breast cancer, the high tumor heterogeneity represents a challenge for prognostic stratification and therapeutic definition. This study aimed to develop predictive models based on somatic variants for the assessment of tumor aggressiveness. A machine learning pipeline was implemented using supervised algorithms, including XGBoost, Support Vector Machine, and Artificial Neural Networks. Variable selection was performed based on predictive importance, prioritizing genes with biological relevance. The performance of the models was evaluated using metrics such as area under the ROC curve, sensitivity, specificity, and F1 score. The results demonstrated high predictive capacity, with the XGBoost model standing out (AUC = 0.88), followed by Neural Networks (AUC = 0.87) and SVM (AUC = 0.85). Interpretability analysis indicated that genes such as PIK3CA, TP53, and ERBB2 were the main determinants of the predictions.

The findings reinforce the potential of using machine learning to identify genomic patterns associated with tumor aggressiveness, contributing to precision medicine strategies.

Keywords: Machine learning. Breast cancer. Somatic variants. Prognosis. Genomics.

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

Year VI, v.1 2026 | Submission: 03/15/2026 | Accepted: 03/17/2026 | Publication: 03/19/2026

1. Introduction

Breast cancer remains one of the leading causes of morbidity and mortality among women. women globally, characterized by high biological, clinical and heterogeneity. molecular (SUNG et al., 2021; BRAZIL, 2024). The evolution of knowledge in molecular biology This allowed the classification of tumors into intrinsic subtypes, such as luminal A, luminal B, HER2-enriched and triple-negative, contributing significantly to prognostic refinement and therapeutic (PEROU et al., 2000; SØRLIE et al., 2001).

Among these, the Luminal B subtype exhibits greater clinical aggressiveness and a higher index. proliferative and with a higher risk of recurrence when compared to the Luminal A subtype, reflecting its molecular complexity and heterogeneous clinical behavior (PRAT et al., 2015; BURSTEIN et al., 2014). This heterogeneity represents a significant challenge in clinical practice, especially in which refers to prognostic stratification and the definition of personalized therapeutic strategies.

With the advancement of next-generation sequencing technologies, it has become possible to identify relevant somatic variants in breast carcinogenesis, especially in genes involved in cell proliferation pathways, DNA repair, and intracellular signaling (CANCER) (GENOME ATLAS NETWORK, 2012). Genes such as PIK3CA, TP53, and ERBB2 have been widely... described as determinants of tumor progression and therapeutic response, and considered central targets in precision oncology (ANDRÉ et al., 2019; SILWAL-PANDIT et al., 2017).

However, the increasing availability of high-dimensional genomic data makes it necessary to... Limitations of traditional statistical methods make it difficult to identify complex patterns. associated with clinical outcomes. In this scenario, machine learning techniques have emerged as promising tools, allowing for the integrated analysis of multiple variables and the construction of more robust predictive models (KOURI et al., 2020; ESTEVA et al., 2019).

Recent studies demonstrate that machine learning algorithms are capable of Integrate genomic and clinical data to predict prognosis, therapeutic response, and risk of tumor progression, with superior performance compared to conventional models in different contexts. oncological (TOPOL, 2019; LIBBRECHT; NOBLE, 2015). In addition, interpretable approaches, How SHAP-based value-based methods have enabled greater transparency in understanding from the individual contribution of the variables, expanding the clinical applicability of these models. (LUNDBERG; LEE, 2017).

Given this context, the development of variant-based predictive models The use of somatics and machine learning techniques represents a promising strategy for improving... Prognostic stratification in Luminal B breast cancer. Thus, the present study proposes the Construction and evaluation of machine learning models capable of identifying patterns. Genomics associated with tumor aggressiveness, contributing to the advancement of precision medicine.

Year VI, v.1 2026 | Submission: 03/15/2026 | Accepted: 03/17/2026 | Publication: 03/19/2026

2. Material and method

2.1 Study design

This is an observational study with an analytical approach, in which a Predictive model based on machine learning techniques for evaluating clinical outcomes. based on genomic signatures. The methodological strategy was structured according to guidelines for studies involving predictive models in health (COLLINS et al., 2015).

2.2 Machine learning pipeline

A machine learning pipeline was implemented, consisting of pre-stages. Processing, variable selection, model training, and performance evaluation.

Preprocessing included standardization of variables and handling of missing data by appropriate imputation methods, based on statistical proximity, were used. Variable selection was... This was achieved through predictive importance analysis, combining L1 regularization and importance analysis. permutational, prioritizing attributes with biological relevance.

2.3 Predictive models

Three widely used supervised algorithms in bioinformatics were evaluated. Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Artificial Neural Networks (RNA).

Hyperparameter optimization was performed using Bayesian search with cross-validation. Nested cross-validation aims to reduce the risk of overfitting and increase generalizability. of the models (BERGSTRA et al., 2011).

2.4 Performance evaluation

The performance of the models was evaluated independently as a group using the metrics. Area under the ROC curve (AUC), sensitivity, specificity, positive predictive value (PPV), value Negative predictive value (NPV) and F1 score.

Considering the potential imbalance between classes, additional metrics were used. based on precision-recall.

2.5 Interpretability of the models

The interpretability of the models was evaluated using the SHAP (Shapley) methodology. Additive Explanations), allowing the quantification of the individual contribution of variables to the predictions (LUNDBERG; LEE, 2017). This approach made it possible to identify the main genomic determinants associated with the outcomes analyzed.

2.6 Statistical analysis and software

The analyses were performed in a computational environment using the R language (version).

4.2.1) and Python (version 3.9), using specific libraries for statistical modeling and machine learning.

The results were considered statistically significant when $p < 0.05$, with Apply correction for multiple comparisons when appropriate.

3. Results and discussion

The application of machine learning models for the analysis of genomic signatures. It demonstrated high predictive capacity in identifying patterns associated with aggressiveness. Tumor in Luminal B subtype breast cancer. Among the algorithms evaluated, the model based in Extreme Gradient Boosting (XGBoost) showed the best overall performance, with AUC-ROC 0.88, followed by Artificial Neural Networks (0.87) and Support Vector Machine (0.85), according to presented in Table 1.

This superior performance of XGBoost can be attributed to its ability to model nonlinear interactions and capturing complex relationships between high-dimensional variables. an intrinsic characteristic of genomic data. Furthermore, ensemble-based algorithms tend... exhibiting greater robustness against noise and biological variability, aspects that are frequently observed in cancer studies (TOPOL, 2019).

Table 1 – Performance of the machine learning models.

Model	AUC-ROC	Sensitivity	Specificity	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

Source: Authors (2026).

Maintaining balanced sensitivity and specificity values between models. suggests good discriminatory capacity and predictive stability, even in scenarios of potential class imbalance. This finding is particularly relevant in oncology, where events Adverse clinical events are often unevenly distributed.

The analysis of the importance of the variables showed that genes such as PIK3CA, TP53 and ERBB2 were the main determinants of the predictions, as shown in Table 2.

Year VI, v.1 2026 | Submission: 03/15/2026 | Accepted: 03/17/2026 | Publication: 03/19/2026

Table 2 – Importance of variables (SHAP analysis).

Variable	Relative Importance (SHAP)	Main Biological Function
PIK3CA	0.32	Via PI3K/AKT/mTOR – cell proliferation
TP53	0.28	Control of the cell cycle and apoptosis
ERBB2	0.25	Tumor growth signaling
Ki-67	0.20	Iron Proliferative index
TFRC	0.15	**Source:** metabolism and tumor growth

Authors (2026).

The relevance of these variables reinforces the biological coherence of the model, since these Genes are widely recognized as central to breast carcinogenesis and resistance. therapeutic (CANCER GENOME ATLAS NETWORK, 2012; ANDRÉ et al., 2019).

Furthermore, the contribution of the proliferative index (Ki-67) highlights the importance of Integration between genomic data and phenotypic parameters, expanding the predictive capacity of models. This finding is aligned with evidence pointing to Ki-67 as a prognostic marker. relevant in the Luminal B subtype (PRAT et al., 2015).

The integrated analysis also made it possible to identify genomic patterns associated with different risk levels, highlighting that the interaction between variants is crucial for behavior. tumoral (Table 3).

Table 3 – Genomic patterns associated with predicted risk.

Genetic pattern	Biological interpretation	with Aggressiveness association
Isolated mutated PIK3CA	activating a proliferative pathway.	Moderate
ERBB2 modified	Increased growth signaling High	
TP53 mutated	Genomic instability	High
PIK3CA + ERBB2	Proliferative synergy	Very high
PIK3CA + TP53	Proliferation + failure to control	Very high

Source: Authors (2026).

These findings reinforce the concept that tumor progression results from complex networks. of molecular interaction, and not of isolated changes. In this sense, machine learning models They offer a significant advantage by capturing these interactions in an integrated way, surpassing limitations of traditional statistical models (LIBBRECHT; NOBLE, 2015).

Another relevant point concerns the use of interpretability methods, such as SHAP. which allowed us to understand the individual contribution of the variables to the predictions. This This approach represents a significant advancement, as it reduces the opacity of the models and promotes... its clinical applicability, especially in contexts that demand transparency in decision-making. decision (LUNDBERG; LEE, 2017).

Furthermore, the results obtained suggest that the application of artificial intelligence can to act as a complementary tool to traditional prognostic stratification strategies, contributing to personalized therapy. This approach can be particularly useful in Scenarios with limited access to commercial multigene panels, expanding equity in care. Oncological.

Despite the promising results, some limitations should be considered. Firstly, machine learning models applied to genomic data are subject to the risk of Overfitting, especially when the number of variables exceeds the number of observations. Although Strategies such as cross-validation have been employed, as well as external validation in cohorts. Independent testing is essential to confirm the robustness of the findings.

Another limitation relates to the lack of integration of multi-omics data, such as transcriptomics, epigenomics, and proteomics, which could expand the predictive capacity of models. The inclusion of this data can capture additional layers of relevant biological regulation. for tumor progression.

Additionally, clinical and environmental factors also influence behavior. The integration of these tumor-related factors was not explored in depth in this analytical framework. Elements in future models can contribute to a more comprehensive and translational approach.

Finally, although interpretability techniques were used, the clinical application The effectiveness of these models still depends on prospective validation and their incorporation into decision-making flows. This presents a current challenge in the implementation of artificial intelligence in healthcare.

Looking ahead, future studies should focus on the external validation of the models in different populations and the integration of multi-omics and clinical data, aiming to increase accuracy. and generalization of predictions. Furthermore, the development of hybrid models, combining Machine learning and prior biological knowledge can represent a promising strategy for To improve interpretability and clinical applicability.

Final considerations

The results of this study demonstrate that the application of machine learning models Genomic signature-based studies show high potential for predicting clinical outcomes. in Luminal B subtype breast cancer. The ability of these models to integrate multiple variables. and capturing complex interactions between molecular pathways allows for a more refined approach to tumor heterogeneity, overcoming limitations of traditional variable-based methods isolated.

Consistent identification of genes central to breast carcinogenesis, such as PIK3CA, TP53 and ERBB2 reinforces the biological validity of the proposed model and highlights the relevance of incorporating genomic data into the construction of predictive tools. Furthermore, the use

Year VI, v.1 2026 | Submission: 03/15/2026 | Accepted: 03/17/2026 | Publication: 03/19/2026

The use of interpretability techniques contributes to the transparency of models, favoring their potential for application in clinical settings.

The findings also highlight the emerging role of artificial intelligence as a tool. complementary in oncology, with the potential to contribute to prognostic stratification and support. to therapeutic decision-making, especially in settings with limited access to molecular testing. advanced.

However, the consolidation of these approaches in clinical practice depends on validation in independent cohorts, as well as integration with clinical and multiomics data. In this way, Future studies are essential to enhance the robustness and applicability of the models. developed.

In summary, the integration between tumor genomics and machine learning techniques represents a promising strategy in advancing precision medicine, with the potential to have an impact significantly impacts the clinical management of breast cancer.

References

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

BURSTEIN, HJ et al. *Estimating the benefits of therapy for early-stage breast cancer: the St. Gallen International Consensus Guidelines.* Annals of Oncology, vol. 25, no. 10, p. 1871–1888, 2014.

CANCER GENOME ATLAS NETWORK. *Comprehensive molecular portraits of human breast tumors.* Nature, vol. 490, no. 7418, p. 61–70, 2012.

COLLINS, GS et al. *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement.* Annals of Internal Medicine, vol. 162, no. 1, p. 55–63, 2015.

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

KOURI, A. et al. *Artificial intelligence in oncology: current applications and future directions.* CA: A Cancer Journal for Clinicians, vol. 70, no. 4, p. 268–287, 2020.

LIBBRECHT, MW; NOBLE, WS *Machine learning applications in genetics and genomics.* Nature Reviews Genetics, vol. 16, no. 6, p. 321–332, 2015.

LUNDBERG, SM; 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, CM et al. *Molecular portraits of human breast tumors.* Nature, vol. 406, p. 747–752, 2000.

PRAT, A. et al. *Prognostic significance of Ki67 in breast cancer.* Journal of Clinical Oncology, vol. 33, no. 36, p. 4234–4242, 2015.

Year VI, v.1 2026 | Submission: 03/15/2026 | Accepted: 03/17/2026 | Publication: 03/19/2026

SILWAL-PANDIT, L. et al. *TP53 mutation spectrum in breast cancer.* Breast Cancer Research, vol. 19, no. 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, vol. 71, no. 3, p. 209–249, 2021.

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