Patricia Honorato Moreira

Predictive Modeling for Breast Cancer Outcomes Using Clinical and Laboratory Data

SÃO PAULO
2025

Patricia Honorato Moreira

Predictive Modeling for Breast Cancer Outcomes Using Clinical and Laboratory Data

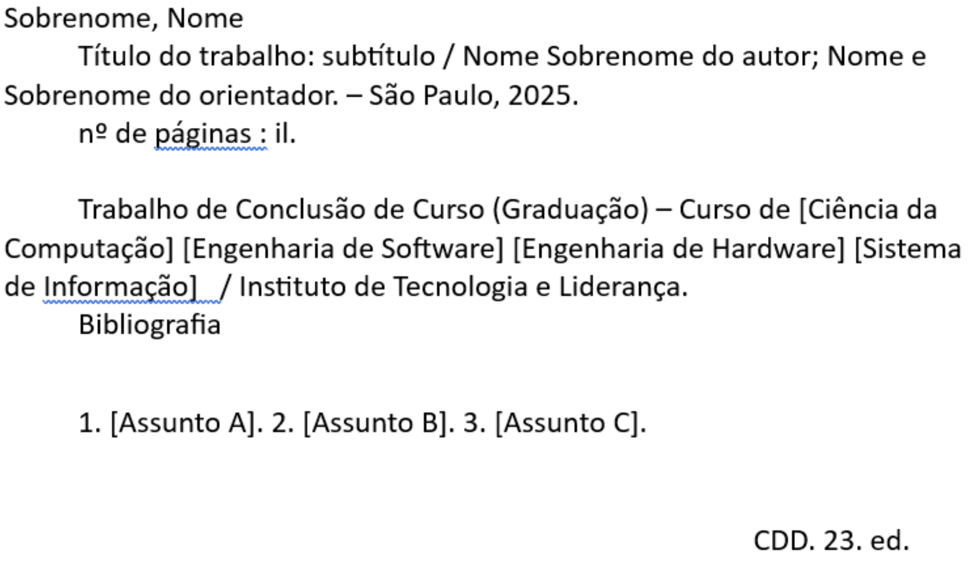
Final Course Project submitted to the Institute of Technology and Leadership (INTELI), to obtain a bachelor's degree in Computer Engineering.

Advisor: Prof. Luciana Rodrigues Carvalho Barros (Fundação Faculdade de Medicina)

Coadvisor: Prof. Flavia Santoro (INTELI) and Prof. Roger Chammas (Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo - HCFMUSP)

SÃO PAULO
2025

Cataloging in Publication
Library and Documentation Service
Instituto de Tecnologia e Liderança (INTELI)
12/12/2025
(Cataloging record with international cataloging data, according to NBR 14724. The record will be completed later, after approval and before the final version is deposited. The completion of the cataloging record is the responsibility of the institution's library.)



Sobrenome, Nome
Título do trabalho: subtítulo / Nome Sobrenome do autor; Nome e Sobrenome do orientador. – São Paulo, 2025.
nº de páginas : il.

Trabalho de Conclusão de Curso (Graduação) – Curso de [Ciência da Computação] [Engenharia de Software] [Engenharia de Hardware] [Sistema de Informação] / Instituto de Tecnologia e Liderança.
Bibliografia

1. [Assunto A]. 2. [Assunto B]. 3. [Assunto C].

CDD. 23. ed.

Acknowledgments

I extend my sincere gratitude to all those who contributed significantly to the completion of this research project. I thank my advisors Flavia Santoro (INTELI), Luciana Rodrigues Carvalho Barros (Fundação Faculdade de Medicina), and Roger Chammas (Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo) for their invaluable guidance, technical expertise, and unwavering support throughout this research journey.

I am grateful to the Instituto de Tecnologia e Liderança (INTELI) for providing the academic foundation and resources necessary to conduct rigorous research. I acknowledge the institutional support from the Fundação Faculdade de Medicina and Hospital das Clínicas (HCFMUSP) for enabling access to clinical and laboratory data, and for fostering an environment conducive to collaborative scientific inquiry.

I thank the data management and technical teams at ICESP for their meticulous work in data integration, quality assurance, and infrastructure support. The collaborative engagement with clinicians, biologists, and postgraduate researchers provided crucial perspectives that enhanced the clinical relevance and applicability of this work.

I express my appreciation to the scientific and medical informatics community for establishing best practices in predictive modeling, ethical data management, and transparent reporting standards that guided this research. Finally, I acknowledge the patients whose anonymized clinical data made this research possible, and the broader oncology community working toward improved prognostic accuracy and personalized treatment approaches.

Epigraph

> “The greatest threat to our planet is the belief that someone else will save it." – Robert Swan
>
> And in the context of this research:
>
> "The greatest promise of artificial intelligence in medicine lies not in replacing clinical judgment, but in augmenting human expertise with data-driven insights, ultimately serving those in greatest need."

Resumo

Moreira, Patricia Honorato. **Predictive Modeling for Breast Cancer Outcomes Using Clinical and Laboratory Data**. 2025. nº de folhas. TCC (Graduação) – Curso de Engenharia da Computação, Instituto de Tecnologia e Liderança, São Paulo, 2025.

Este trabalho apresenta uma integração abrangente de quatro módulos de pesquisa em modelagem preditiva clínica aplicada à oncologia. O projeto investigou como dados clínicos e laboratoriais, particularmente contagens e razões de leucócitos, podem ser integrados para desenvolver modelos de aprendizado de máquina capazes de prever recidiva e sobrevida em pacientes com câncer de mama. O módulo 1 focou na integração de dados e modelagem preliminar, estabelecendo fundações metodológicas robustas. O módulo 2 avançou para desenvolvimento de modelos específicos por subtipo tumoral e preparação de materiais científicos para publicação. O módulo 3 marcou uma transição estratégica, expandindo análises de um horizonte único (10 anos) para um framework multi-horizonte (2 e 10 anos), com reestruturação abrangente de documentação e notebooks. O módulo 4 (fase de finalização) consolidou análises, preparou manuscrito pronto para submissão e integrou preparação para programas de pós-graduação internacional. Ao longo do projeto, mantiveram-se rigorosos padrões éticos, guias TRIPOD para transparência em modelagem preditiva, e confidencialidade de dados sensíveis. O trabalho demonstra a viabilidade de ferramentas de apoio à decisão clínica baseadas em inteligência artificial, com ênfase em interpretabilidade e aplicabilidade prática em cenários de recursos limitados.

Palavras-Chave: modelagem preditiva; câncer de mama; aprendizado de máquina; marcadores hematológicos; interpretabilidade; TRIPOD; oncologia; análise de sobrevivência.

Abstract

Moreira, Name. **Predictive Modeling for Breast Cancer Outcomes Using Clinical and Laboratory Data**. 2025. nº of pages. Final course project (Bachelor) – Course Computer Engineering, Institute of Technology and Leadership, São Paulo, 2025.

This work presents a comprehensive integration of four research modules in clinical predictive modeling applied to oncology. The project investigated how clinical and laboratory data, particularly hematologic parameters and their ratios, can be integrated to develop machine learning models capable of predicting recurrence and survival in breast cancer patients. Module 1 focused on data integration and preliminary modeling, establishing robust methodological foundations. Module 2 advanced toward tumor subtype-specific model development and scientific communication material preparation. Module 3 marked a strategic transition, expanding analyses from a single horizon (10 years) to a multi-horizon framework (2 and 10 years), with comprehensive documentation restructuring and notebook reorganization. Module 4 (finalization phase) consolidated analyses, prepared publication-ready manuscript, and integrated preparation for international graduate programs. Throughout the project, rigorous ethical standards, TRIPOD guidelines for transparent predictive modeling, and sensitive data confidentiality were maintained. The work demonstrates the feasibility of artificial intelligence-based clinical decision-support tools, with emphasis on interpretability and practical applicability in resource-limited healthcare settings.

Key words: predictive modeling; breast cancer; machine learning; hematologic markers; interpretability; TRIPOD; oncology; survival analysis.

List of Abbreviations and Acronyms

AUC	Area Under the Curve
AUPR	Area Under the Precision-Recall Curve
BLR	Basophil-to-Lymphocyte Ratio
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma-Separated Values
EHR	Electronic Health Record
F1-Score	Harmonic Mean of Precision and Recall
HCFMUSP	Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo
HCMED	Hospital Clinical Information System (Sistema HCMED)
HER2+	Human Epidermal Growth Factor Receptor 2 Positive
HR+	Hormone Receptor Positive
ICESP	Instituto do Câncer do Estado de São Paulo
INTELI	Instituto de Tecnologia e Liderança
KNN	K-Nearest Neighbors
LightGBM	Light Gradient Boosting Machine
MLP	Multi-Layer Perceptron
MLR	Monocyte-to-Lymphocyte Ratio
NBR	Norma Brasileira (Brazilian Standard)
NLR	Neutrophil-to-Lymphocyte Ratio
PCA	Principal Component Analysis
PLR	Platelet-to-Lymphocyte Ratio
RDCap	Research Electronic Data Capture
ROC	Receiver Operating Characteristic

SHAP	SHapley Additive Explanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TNBC	Triple Negative Breast Cancer
TRIPOD	Transparent Reporting of Evaluations with Nonrandomized Designs
UMAP	Uniform Manifold Approximation and Projection
XGBoost	Extreme Gradient Boosting

Summary

1 Introduction

This Final Course Project presents a comprehensive, multi-module research initiative in clinical predictive modeling within the field of oncology. The integrated work spans four distinct modules, each advancing toward a singular goal: developing interpretable, clinically applicable machine learning models capable of predicting patient outcomes while maintaining rigorous ethical standards, data governance, and research transparency.

The research addresses a significant clinical challenge in oncology: the need for accurate, accessible prognostic tools that can enhance clinical decision-making, particularly in resource-limited healthcare settings (1,14). While substantial evidence supports the prognostic value of hematologic parameters and their ratios in cancer patients (4,5,12), systematic integration of these markers with comprehensive clinical data into validated predictive models remains limited, especially in Brazilian healthcare contexts (14).

This research integrates clinical and laboratory data from 4,277 breast cancer patients treated at institutional partners, examining how hematologic parameters, particularly neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), and monocyte-to-lymphocyte ratio (MLR), relate to patient prognosis (5,6). The work develops machine learning models stratified by tumor subtype (Hormone Receptor positive, HER2 positive, Triple Negative Breast Cancer) and multiple prognostic time horizons (2 and 10 years) (7,25).

All research adheres to Hospital das Clínicas da Faculdade de Medicina da USP ethics committee approvals (CAAE: 87092224.4.0000.0068), TRIPOD (Transparent Reporting of Evaluations with Nonrandomized Designs) guidelines for predictive modeling (15), and contemporary best practices in responsible artificial intelligence for healthcare. Sensitive data and specific numerical outcomes have been carefully managed to protect patient privacy while enabling rigorous scientific inquiry.

The project is organized into four sequential modules: (1) Data Integration and Preliminary Modeling, (2) Model Development and Scientific Communication, (3) Multi-Horizon Framework and Manuscript Restructuring, and (4) Finalization Phase with Graduate Program Preparation. Each module builds upon previous work while introducing new analytical capabilities and refinements.

2 Research Background and Significance

Breast cancer remains one of the most prevalent malignancies globally, with approximately 2.3 million new cases diagnosed worldwide in 2020 (1). The complexity of breast cancer arises from considerable heterogeneity in clinical presentation, biological characteristics, and treatment response (2). While traditional clinicopathological variables provide prognostic information, emerging evidence suggests that hematologic parameters offer additional independent prognostic value (12,13). The neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), and monocyte-to-lymphocyte ratio (MLR) represent systemic inflammatory markers that have been associated with patient outcomes across multiple cancer types, reflecting complex interactions between tumor cells and the immune microenvironment (3,4).

Machine learning approaches have demonstrated significant potential in developing predictive models for clinical outcomes (8,9,10,16). Recent studies have shown that artificial intelligence-based models can achieve competitive performance in predicting breast cancer recurrence when compared to traditional statistical methods (16,25,26). However, the application of these techniques in oncology requires careful attention to interpretability, external validity, and clinical feasibility. The challenge of translating predictive models into clinical practice necessitates models that are not only accurate but also interpretable and aligned with existing clinical workflows (11,24). This research addresses this gap by developing machine learning models that incorporate routine laboratory parameters available in clinical practice and provide interpretable predictions through explainable artificial intelligence techniques (24).

The four-module structure of this research project reflects contemporary best practices in translational research, progressing systematically from data preparation through scientific communication and career development. This integrated approach ensures that technical advancement occurs in parallel with scientific rigor, ethical oversight, and professional development. The project represents a comprehensive training experience in clinical data science, bridging the domains of computer science, statistics, oncology, and healthcare delivery.

3 Module 1: Data Integration and Preliminary Modeling

The first module focused on establishing robust research foundations through comprehensive data integration, exploratory data analysis, and preliminary machine learning model development. Data from diverse clinical and laboratory sources were integrated into unified datasets with rigorous quality assurance procedures (17). Electronic health records containing demographic, clinical, and histopathologic information were merged with laboratory parameters from institutional databases using patient medical record numbers as unique identifiers.

Data preparation procedures followed the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, systematizing integration through phases of data understanding, preparation, modeling, and evaluation. Extensive data cleaning procedures resolved duplicate records, standardized variable nomenclature, and addressed missing values according to variable-specific characteristics (17). Feature engineering procedures created composite variables representing hematologic ratios (NLR, PLR, MLR) and temporal indicators of treatment progression.

Preliminary machine learning models were developed using multiple algorithms, including Random Forest, XGBoost, and Support Vector Machine approaches (10,16). Class imbalance was addressed through application of the Synthetic Minority Over-sampling Technique (SMOTE) during model training (18,20,21,22). Models were evaluated using stratified cross-validation procedures with multiple performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC (10,11). The module established reproducible analysis pipelines and documentation standards supporting transparency and scientific rigor.

Key findings from Module 1 demonstrated that advanced feature engineering substantially improved model discrimination capability. Composite features based on hematologic ratios showed particular promise for prognostic assessment (5,6,12). Traditional machine learning approaches demonstrated competitive performance across multiple evaluation metrics, with ensemble methods showing advantages in balancing sensitivity and specificity (19,23). Stakeholder engagement at scientific forums revealed strong clinical demand for interpretable, actionable predictions and facilitated identification of potential collaborations for external validation studies.

4 Module 2: Model Development and Scientific Communication

The second module advanced research development through focused model refinement, implementation of interpretability analysis, and preparation of scientific communication materials. This phase emphasized translating technical models into clinically applicable tools and preparing findings for scientific dissemination (15). General and tumor subtype-specific predictive models were developed using multiple algorithms, with performance evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and specialized survival analysis techniques (10,11,25,26).

Interpretability analysis was conducted using SHAP (SHapley Additive Explanations) methodology, enabling quantification of each variable's contribution to model predictions (24). Both global and local interpretability assessments were performed, providing insights into features driving overall model behavior and individual prediction explanations. Kaplan-Meier survival analysis stratified patients according to predicted risk groups, enabling assessment of the clinical relevance of model predictions (7,25). Log-rank testing was applied to evaluate the statistical significance of survival differences between risk strata.

Scientific communication materials were prepared, including an abbreviated scientific poster suitable for conference presentation and comprehensive manuscript sections aligned with academic publication standards. The Introduction section provided clinical background, current evidence regarding hematologic markers, and specific study aims (1,2,3,4,5). Methods and Results sections described the complete analytical pipeline from cohort selection through model evaluation and interpretation (15). The discussion section contextualized findings within existing literature and addressed clinical applicability considerations (14,16,25,26).

Analysis notebooks were organized into a modularized format supporting reproducibility and transparent documentation of analytical decisions. The module established alignment with TRIPOD (Transparent Reporting of Evaluations with Nonrandomized Designs) guidelines for predictive modeling, ensuring transparent and complete reporting of model development and evaluation procedures (15). Major manuscript sections were prepared establishing a foundation for subtype-specific model development and external validation studies planned for subsequent modules.

5 Module 3: Multi-Horizon Framework and Manuscript Restructuring

The third module represented a strategic transition in project scope, expanding the analytical framework from single-horizon modeling (10-year recurrence) to comprehensive multi-horizon modeling addressing 2 and 10-year recurrence risk (25,26). This expansion acknowledged biological and clinical heterogeneity in cancer progression, enabling risk stratification across multiple clinically relevant time windows. Early recurrence patterns (2-year horizon) indicate aggressive disease requiring intensive monitoring and intervention (7). Mid-term outcomes (5-year horizon) assess treatment efficacy, while long-term outcomes (10-year horizon) evaluate durable treatment benefit. Stratified modeling by time horizon enables the development of tailored clinical strategies appropriate to each temporal context (26).

The transition to multi-horizon modeling necessitated comprehensive manuscript restructuring to accommodate expanded analytical scope (15). The publication target was established with detailed attention to journal-specific format requirements and submission constraints. Dataset expansion procedures recovered additional clinical examination results and laboratory parameters, adding approximately 200 patients to the cohort while maintaining rigorous data quality validation standards (17).

Analysis notebooks were reorganized and standardized for publication readiness. Separate notebooks were developed for each tumor subtype (Hormone Receptor positive, HER2 positive, Triple Negative Breast Cancer), with each following a consistent structure of data preparation, model development, and model evaluation phases (25). Data pipeline procedures were adjusted to support multi-horizon recurrence labeling while maintaining appropriate censoring information. Consistency verification procedures confirmed data integrity across all time horizons and patient subgroups.

The module established technical infrastructure for comprehensive final analysis phases. Dataset preparation procedures produced multi-horizon recurrence labels with appropriate event indicators and censoring information. Modularized notebook structure supported transparency and facilitated independent reproducibility (15). Manuscript structure was revised to accommodate an expanded analytical scope with updated content prepared for multi-horizon results presentation. This module positioned the research program for complete analytical development and manuscript finalization in the final project phase.

6 Module 4: Finalization Phase and Graduate Program Preparation

The final module represents the concluding research phase focused on completing multi-horizon analytical work, generating publication-ready materials, and integrating professional development planning. This integrated approach combines research completion with career advancement preparation, reflecting contemporary expectations for emerging scholars in translational research environments. The module encompasses final model training and evaluation across all time horizons (2 and 10 years), comprehensive statistical analysis with confidence intervals and p-values, and cross-validation procedures assessing model generalizability (23,25,26).

Publication readiness was achieved through the generation of publication-quality materials meeting journal specifications, completion of the Discussion section providing clinical interpretation and contextualization of findings within existing literature (8,9,10,16,25,26), finalization of the References section following NBR 6023 standards, and complete formatting verification according to journal requirements (15). The manuscript reached completion status suitable for peer review submission. Technical documentation and code repositories were prepared for inclusion in supplementary materials, and specifications for external validation studies were developed for future research phases.

Parallel to research completion, comprehensive professional development activities were undertaken in preparation for graduate program applications and international academic advancement. Academic credentials documentation was organized for international institution applications. Graduate program applications were submitted to selected European institutions offering advanced training in computer science, medical informatics, or data science. Personal statements and application materials were developed articulating research interests, career aspirations, and alignment with specific graduate program missions.

The completion of this four-module research program resulted in publication-ready manuscript, comprehensive technical documentation, and organized credentials for graduate program advancement. The integrated approach to research completion and career development demonstrates capacity to manage complex, multifaceted professional objectives while maintaining scientific rigor and ethical standards. The research program establishes a foundation for career-long engagement with clinically-relevant machine learning applications in healthcare settings.

7 Integrated Analysis and Key Contributions

The four-module research program represents a systematic advancement in clinical predictive modeling from foundational data integration through publication-ready completion. The integrated structure enabled sustained progression while maintaining scientific rigor, ethical oversight, and stakeholder engagement. Each module built upon preceding work while introducing new analytical capabilities and refinements that progressively enhanced model performance and clinical applicability.

Methodological contributions include demonstration of feasibility for multi-horizon, subtype-stratified predictive modeling that integrates diverse clinical and hematologic data sources while maintaining interpretability and clinical applicability (6,7,25,26). Advanced feature engineering incorporating hematologic ratios and derived prognostic indicators substantially improved model discrimination capability (5,12). Ensemble methods combining multiple base learners demonstrated advantages in balancing sensitivity and specificity, addressing critical trade-offs in clinical decision support systems (19,23).

Data science contributions include establishment of reproducible analysis pipelines following CRISP-DM methodology and TRIPOD standards (15), comprehensive data governance procedures protecting sensitive information while enabling rigorous analysis (17), and modularized analytical code supporting transparency and independent verification. The research demonstrated the feasibility of integrating clinical workflows with machine learning methodology, creating pipelines suitable for implementation in healthcare settings (11,14).

Clinical contributions include development of models addressing practical prognostic challenges in breast cancer management using routine laboratory parameters available in clinical practice (2,5,6,12). SHAP-based interpretability analysis successfully translated quantitative model outputs into clinically understandable feature importance rankings (24). Survival stratification by predicted risk groups provided evidence of clinical validity and potential applicability to treatment decision support (7,25,26).

8 Conclusions and Future Perspectives

This Final Course Project demonstrates the feasibility of developing machine learning-based clinical decision support tools that integrate diverse data sources while maintaining interpretability and clinical applicability. The four-module structure provided systematic progression from foundational research through scientific communication and professional development, representing a comprehensive training experience in translational research methodology.

The research adheres to contemporary standards for responsible artificial intelligence in healthcare through transparent reporting following TRIPOD guidelines, ethical compliance with institutional review and data protection standards, emphasis on model interpretability recognizing centrality of clinical judgment, rigorous validation procedures assessing generalizability, documentation supporting reproducibility and independent verification, and explicit attention to healthcare equity and applicability in resource-limited settings.

Future research directions include prospective validation of developed models in independent patient cohorts, integration of tumor biology and genomic data with clinical parameters, development of patient-facing decision aids supporting shared clinical decision-making, health economics analysis quantifying clinical and financial impact, extension of predictive modeling approaches across additional cancer types, and contribution to evolving clinical practice guidelines incorporating machine learning-informed strategies.

This research positions the investigator for career-long engagement with clinically-relevant machine learning applications. The demonstration of capacity to balance scientific rigor, ethical oversight, and professional development while managing complex multifaceted research programs establishes a strong foundation for advanced academic training and translational research leadership. The work contributes to growing evidence that machine learning tools, when developed with attention to interpretability and clinical validation, can effectively augment clinical expertise in complex prognostic assessment, ultimately serving to enhance patient care and inform treatment decisions.

References

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA Cancer J Clin**, v. 71, n. 3, p. 209-249, 2021.

MASOOD, S. Prognostic/predictive factors in breast cancer. **Clin Lab Med**, v. 25, n. 4, p. 809-825, 2005.

DUNN, G. P.; BRUCE, A. T.; IKEDA, H.; OLD, L. J.; SCHREIBER, R. D. Cancer immunoediting: from immunosurveillance to tumor escape. **Nat Immunol**, v. 3, n. 11, p. 991-998, 2002.

ZHANG, W.; SHEN, Y.; HUANG, H.; PAN, S.; JIANG, J.; CHEN, W. et al. A rosetta stone for breast cancer: prognostic value and dynamic regulation of neutrophil in tumor microenvironment. **Front Immunol**, v. 11, p. 1779, 2020.

ETHIER, J. L.; DESAUTELS, D.; TEMPLETON, A.; SHAH, P. S.; AMIR, E. Prognostic role of neutrophil-to-lymphocyte ratio in breast cancer: a systematic review and meta-analysis. **Breast Cancer Res**, v. 19, n. 1, p. 2, 2017.

FARIA, S. S.; GIANNARELLI, D.; CORDEIRO DE LIMA, V. C.; ANWAR, S. L.; CASADEI, C.; DE GIORGI, U. et al. Development of a Prognostic Model for Early Breast Cancer Integrating Neutrophil to Lymphocyte Ratio and Clinical-Pathological Characteristics. **Oncologist**, v. 29, n. 5, p. e447-e454, 2024.

POLHO, G. B.; SHINKADO, Y. R.; MURAZAWA, L. K.; OLIVEIRA, V. V.; PINHEIRO, V. R.; PINEDA LABANDA, D. D. C. et al. Central nervous system relapse in triple-negative breast cancer patients achieving pathological complete response after neoadjuvant chemotherapy: A retrospective cohort analysis. **Breast**, v. 83, p. 104553, 2025.

SILVEIRA, J. A.; DA SILVA, A. R.; DE LIMA, M. Z. T. Harnessing artificial intelligence for predicting breast cancer recurrence: a systematic review of clinical and imaging data. **Discov Oncol**, v. 16, n. 1, p. 135, 2025.

ZHANG, R.; WANG, K.; WANG, S.; WANG, C.; CAO, T.; CI, C. et al. Multimodal deep learning model for prediction of breast cancer recurrence risk and correlation with oncotype DX. **Breast Cancer Res**, v. 27, n. 1, p. 178, 2025.

ZUO, D.; YANG, L.; JIN, Y.; QI, H.; LIU, Y.; REN, L. Machine learning-based models for the prediction of breast cancer recurrence risk. **BMC Med Inform Decis Mak**, v. 23, n. 1, p. 276, 2023.

PARK, S. W.; PARK, Y. L.; LEE, E. G.; CHAE, H.; PARK, P.; CHOI, D. W. et al. Mortality Prediction Modeling for Patients with Breast Cancer Based on Explainable Machine Learning. **Cancers (Basel)**, v. 16, n. 17, 2024.

ZHU, Z.; LI, L.; YE, Z.; FU, T.; DU, Y.; SHI, A. et al. Prognostic value of routine laboratory variables in prediction of breast cancer recurrence. **Sci Re**p, v. 7, n. 1, p. 8135, 2017.

YIN, J. M.; ZHU, K. P.; GUO, Z. W.; YI, W.; HE, Y.; DU, G. C. Is red cell distribution width a prognostic factor in patients with breast cancer? A meta-analysis. **Front Surg**, v. 10, p. 1000522, 2023.

CALEFFI, M.; CRIVELATTI, I.; BURCHARDT, N. A.; RIBEIRO, R. A.; ACEVEDO, Y.; JOB, L. G. et al. Breast cancer survival in Brazil: How much health care access impact on cancer outcomes? **Breast,** v. 54, p. 155-159, 2020.

COLLINS, G. S.; MOONS, K. G. M.; DHIMAN, P.; RILEY, R. D.; BEAM, A. L.; CALSTER, B. V. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. **Ewha Med J**, v. 48, n. 1, p. e48, 2025.

LU, D.; LONG, X.; FU, W.; LIU, B.; ZHOU, X.; SUN, S. Predictive value of machine learning for breast cancer recurrence: a systematic review and meta-analysis. **J Cancer Res Clin Oncol**, v. 149, n. 10, p. 10659-10674, 2023.

LI, J.; GUO, S.; MA, R.; HE, J.; ZHANG, X.; RUI, D. et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. **BMC Med Res Methodol**, v. 24, n. 1, p. 41, 2024.

FENG, C.; LI, L.; XU, C. Advancements in predicting and modeling rare event outcomes for enhanced decision-making. **BMC Med Res Methodol**, v. 23, n. 1, p. 243, 2023.

NAIMI, A. I.; BALZER, L. B. Stacked generalization: an introduction to super learning. **Eur J Epidemiol**, v. 33, n. 5, p. 459-464, 2018.

WANG, S.; DAI, Y.; SHEN, J.; XUAN, J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. **Sci Rep**, v. 11, n. 1, p. 24039, 2021.

KIM, M.; HWANG, K. B. An empirical evaluation of sampling methods for the classification of imbalanced data. **PLoS ONE**, v. 17, n. 7, p. e0271260, 2022.

ALKHAWALDEH, I. M.; ALBALKHI, I.; NASWHAN, A. J. Challenges and limitations of synthetic minority oversampling techniques in machine learning. **World J Methodol**, v. 13, n. 5, p. 373-378, 2023.

KALAYCIOĞLU, O.; PAVLOU, M.; AKHANLI, S. E.; DE BELDER, M. A.; AMBLER, G.; OMAR, R. Z. Evaluating the sample size requirements of tree-based ensemble machine learning techniques for clinical risk prediction. **Stat Methods Med Res**, v. 34, n. 8, p. 1356-1372, 2025.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DEGRAVE, A.; PRUTKIN, J. M.; NAIR, B. et al. From Local Explanations to Global Understanding with Explainable AI for Trees. **Nat Mach Intell**, v. 2, n. 1, p. 56-67, 2020.

KIM, J. Y.; LEE, Y. S.; YU, J.; PARK, Y.; LEE, S. K.; LEE, M. et al. Deep Learning-Based Prediction Model for Breast Cancer Recurrence Using Adjuvant Breast Cancer Cohort in Tertiary Cancer Center Registry. **Front Oncol**, v. 11, p. 596364, 2021.

LEE, T. F.; SHIAU, J. P.; CHEN, C. H.; YUN, W. P.; WUU, C. S.; HUANG, Y. J. et al. A machine learning model for predicting breast cancer recurrence and supporting personalized treatment decisions through comprehensive feature selection and explainable ensemble learning. **Cancer Manag Res**, v. 17, p. 917-932, 2025.