CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Desbloqueando o hemograma completo como uma ferramenta de estratificação de risco para câncer de mama usando aprendizado de máquina

Unlocking the complete blood count as a risk stratification tool for breast cancer using machine learning

Desbloqueando el hemograma completo como una herramienta de estratificación de riesgo para el cáncer de mama utilizando aprendizaje automático

Daniella Castro Araújo[1], Bruno Aragão Rocha[2], Karina Braga Gomes[3], Daniel Noce da Silva[4], Vinicius Moura Ribeiro[5], Marco Aurelio Kohara[6], Adriano Alonso Veloso[7], Flavia Helena da Silva[8], Pedro Henrique Araújo de Souza[9], Ismael Dale Cotrim Guerreiro da Silva[10]

1 PhD, Founder & CTO, Huna Ltd., São Paulo, Brazil.
2 MD, Coordenador Médico de Inovação, Grupo Fleury, São Paulo, Brazil.
3 Prof. PhD, Departamento de Análises Clínicas e Toxicológicas, Faculdade de Farmácia, Universidade Federal de Minas Gerais/UFMG, Campus Belo Horizonte, Minas Gerais, Brazil
4 MSc, Huna Ltd., São Paulo, Brazil.
5 Founder & CEO, Huna Ltd., São Paulo, Brazil.
6 Founder & COO, Huna Ltd., São Paulo, Brazil.
7 Prof. PhD, Departamento de Ciências da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais/UFMG, Campus Belo Horizonte, Minas Gerais, Brazil
8 PhD, Gerente Sênior Inteligência Analytics, Grupo Fleury, São Paulo, Brazil.
9 MSc, MD, Oncologista, Department of Oncology Clinical Research, Instituto Nacional de Câncer (INCA), Rio de Janeiro, Brazil
10 Prof. PhD, MD, Department of Gynecology, Escola Paulista de Medicina, Federal University of São Paulo, São Paulo, Brazil

Autor correspondente: PhD/Dr. Daniela Castro Araújo
E-mail: dani@huna-ai.com
website do projeto - https://huna-ai.com/
Link do artigo - https://www.nature.com/articles/s41598-024-61215-y

Resumo

Objetivo: Avaliar a eficácia do ML no uso do hemograma para avaliação de risco de câncer de mama. *Método:* Este estudo retrospectivo analisou hemogramas de 396.848 mulheres de 40 a 70 anos. Foram identificados 2861 casos (1882 confirmados por

biópsia e 979 por imagens), enquanto 393.987 foram controles (BI-RADS 1 ou 2). Os dados foram divididos em conjuntos de modelagem (treinamento e validação) e teste com base na certeza diagnóstica. *Resultados:* O modelo de regressão *ridge*, incorporando a razão neutrófilo-linfócito, glóbulos vermelhos e idade, atingiu uma AUC de 0,64. A população do estudo foi estratificada em quatro grupos de risco: alto, moderado, médio e baixo, com razões relativas de 1,99, 1,32, 1,02 e 0,42, respectivamente. *Conclusão:* Este modelo de ML fornece uma ferramenta de baixo custo para triagem personalizada de câncer de mama, potencialmente melhorando a detecção precoce em ambientes com recursos limitados.
Palavras-chave: Câncer de Mama; Aprendizado de Máquina; Contagem de Células Sanguíneas.

Abstract

Objective: To evaluate the efficacy of machine learning (ML) in using complete blood count (CBC) for breast cancer risk assessment. *Method:* This retrospective study analyzed CBCs from 396,848 women aged 40 to 70. A total of 2861 cases were identified (1882 confirmed by biopsy and 979 by imaging), while 393,987 were controls (BI-RADS 1 or 2). Data were divided into modeling (training and validation) and testing sets based on diagnostic certainty. *Results:* The ridge regression model, incorporating the neutrophil-to-lymphocyte ratio, red blood cells, and age, achieved an AUC of 0.64. The study population was stratified into four risk groups: high, moderate, medium, and low, with relative ratios of 1.99, 1.32, 1.02, and 0.42, respectively. *Conclusion:* This ML model provides a cost-effective tool for personalized breast cancer screening, potentially improving early detection in resource-limited settings.
Keywords: Breast Cancer; Machine Learning; Blood Cell Count.

Resumen

Objetivo: Evaluar la eficacia del ML en el uso del hemograma para la evaluación del riesgo de cáncer de mama. *Método:* Estudio retrospectivo analizó hemogramas de 396,848 mujeres de 40 a 70 años. Se identificaron 2861 casos (1882 confirmados por biopsia y 979 por imágenes), mientras que 393,987 fueron controles (BI-RADS 1 o 2). Los datos se dividieron en conjuntos de modelado (entrenamiento y validación) y prueba según la certeza diagnóstica. *Resultados:* El modelo de regresión ridge, que

incorpora la relación neutrófilo-linfocito, los glóbulos rojos y la edad, alcanzó una AUC de 0.64. La población del estudio se estratificó en cuatro grupos de riesgo: alto, moderado, medio y bajo, con razones relativas de 1.99, 1.32, 1.02 y 0.42, respectivamente. *Conclusión:* ML proporciona una herramienta rentable para el cribado personalizado del cáncer de mama, mejorando potencialmente la detección temprana en entornos con recursos limitados.

Palabras clave: Cáncer de Mama; Aprendizaje Automático; Recuento de Células Sanguíneas.

Introduction

Mammography screening is highly recommended for early-stage breast cancer (BC) detection and reducing mortality (1). While the American Cancer Society (ACS) advises age-based screening guidelines, this approach may not consider individual risk factors. Additionally, low- and middle-income countries often struggle to provide serial screening (2). Therefore, adopting a reliable risk stratification tool to prioritize higher-risk women could optimize resource allocation by transitioning from population-based to individual screening (3).

Several BC risk stratification tools have been proposed, with the Tyrer-Cuzick (TC) model (3,4) being the most widely adopted and recommended by the ACS guidelines. This model leverages a composite of family history, demographic details, reproductive health insights, anthropometric data, and breast density to predict the risk of BC. However, despite its comprehensive approach, the TC model's accuracy remains limited. Furthermore, the specific data it requires might not be readily available in standard medical records, thereby necessitating an intricate and specialized data collection process. Recently, a deep learning model based on convolutional neural networks called MIRAI was introduced as a mammography-based risk assessment tool that predicts BC within five years, achieving a mean AUC of 0.78 in a multicenter evaluation (5). Despite its effectiveness, the applicability of the MIRAI model can be limited in low-income countries due to a scarcity of mammograms, frequently resulting in the unavailability of necessary mammographic data for comprehensive risk assessment.

Liquid biopsy, a promising technology for early cancer detection, detects

circulating tumor-associated cells using blood tests but remains expensive and less accessible. The prospect of cancer detection through a simple blood draw has always been appealing, especially for communities with limited medical access; yet, cost is a key factor for a genuine impact on public health. Consequently, affordable and globally accessible routine blood tests emerge as a timely solution, despite their general nonspecificity for diagnosis. Numerous studies have investigated the relationship between routine blood markers and BC, finding promising associations with complete blood count (CBC) parameters (6).

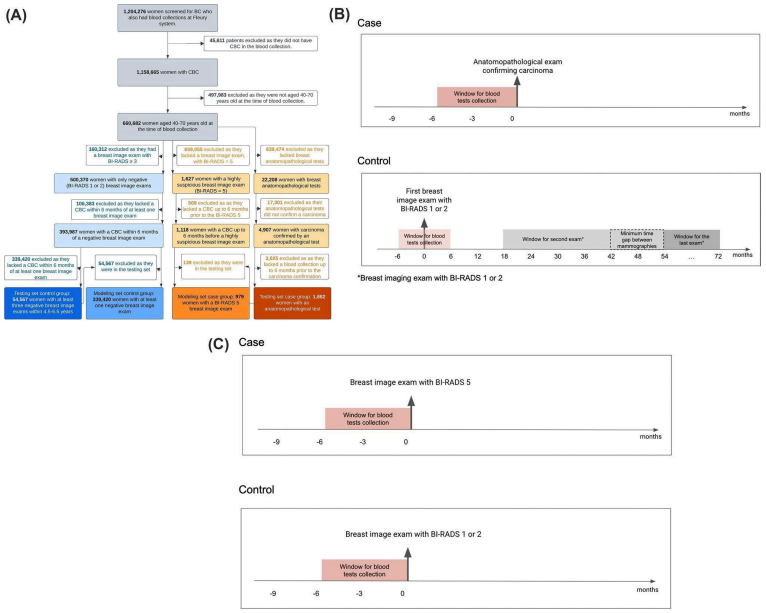
Machine learning (ML) has been widely used to identify complex patterns in blood exams, enabling the combination of various markers to transform nonspecific tests into precise ones for multiple diseases. Here, we propose combining routine blood makers into a powerful AI-based panel to assess the risk of BC within six months of diagnosis. To our knowledge, this is the first study to utilize ML techniques to assess non-linear relationships between routine blood count test CBC markers for BC risk assessment.

Methods

Study design and patients

This retrospective study was approved by the Fleury Group's Research Ethics Committee (CEP-FG) (CAAE: 60984722.4.0000.5474), which is duly qualified by the National Research Ethics Committee (CONEP) of the National Health Council of Brazil. By decision of the CEP-FG, the requirement for informed consent was waived because of the retrospective and anonymized nature of the data. All research was carried out in compliance with the Brazilian legislation, General Data Protection Law (Lei Geral de Proteção dos Dados—LGPD), and the guidelines of the Declaration of Helsinki. We collected CBC test results from 396,848 women aged 40–70, screened for breast cancer (BC) between January 2004 and August 2022 across 309 Fleury clinical laboratory units in eight Brazilian states. Figure 1A shows the schematic diagram of the study population.

Figure 1A - Schematic depiction of study design. (A) Flow diagram of the study population at the Fleury Clinical Labs that met all study inclusion criteria. (B) Time windows for blood exam collection for the testing dataset. (C) Time windows for blood exam collection for the modeling dataset.



The labeling of CBCs was performed by mining carcinoma confirmations from anatomopathological reports or the breast imaging-reporting and data system (BI-RADS) categories from breast imaging reports (mammograms, ultrasounds, or magnetic resonance imaging).

The case group comprised women diagnosed with BC, confirmed by an anatomopathological exam, or highly suspected of it (BI-RADS 5, indicating a cancer likelihood of > 95%), with a blood test collected up to six months before their breast biopsy or imaging exam. The control group included women with negative imaging exams (BI-RADS 1 or 2). Due to the varying levels of diagnostic certainty, we partitioned the database into two main segments: the test set, for the highest certainty, and the modeling set, which combines training and validation sets.

Test Set: This set included 1,882 cases anatomopathologically confirmed with a CBC conducted up to six months before breast biopsy. Controls were women who remained cancer-free for at least 4.5–6.5 years, encompassing women with at least three negative (BI-RADS 1 or 2) breast image exams within this period, spaced apart by at least 18 months. Their CBCs were collected in a 6-month window centered on the baseline image exam, totaling 54,567 CBCs, one per woman. For individuals with multiple tests, only the first CBC test was considered. Dataset construction time windows for the test set are illustrated in Fig. 1B.

Modeling Set: This set included 979 BI-RADS 5 cases, each with one blood

analysis conducted up to six months before the breast image exam. Controls included 339,420 women with one negative (BI-RADS 1 or 2) result, each represented by a CBC conducted up to six months before the breast image exam. The modeling set was further segmented into training (80%) and validation (20%) phases. Dataset construction time windows for the modeling set are detailed in Fig. 1C.Figure 2 provides a visual overview of the case group characteristics.

Materials

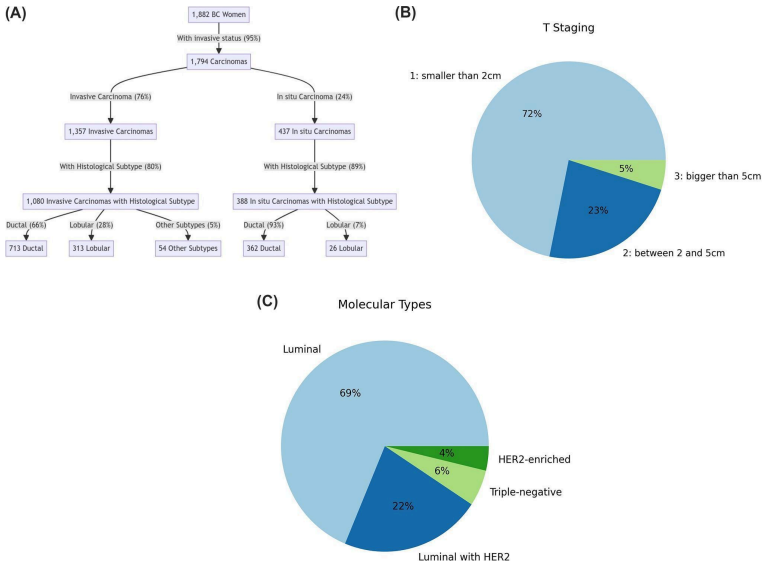
CBC measurements were obtained from EDTA-K3 peripheral blood samples analyzed by the Automated Hematology Analyzer XT or XN series from Sysmex (Sysmex Corporation, Kobe, Japan). Red blood cells (RBC) and platelets were counted and sized through direct current impedance, while hematocrit and hemoglobin levels were determined using pulse height and sodium lauryl sulfate spectrophotometry, respectively.

AI-based models

Our main goal was to use supervised learning to classify blood tests collected from BC patients up to six months before the diagnosis versus those from control patients. Models were developed using two different machine learning models: ridge regression (7,8) and LightGBM (9). Ridge Regression is highly interpretable, enhancing linear regression by incorporating an L2 regularization term. This term penalizes large coefficients to prevent overfitting and address multicollinearity among predictors. On the other hand, LightGBM is a gradient-boosting algorithm known for its efficiency and effectiveness. It builds models sequentially, correcting errors from previous trees, and is particularly suited for handling large datasets, providing high accuracy with reduced training time. Their performance was evaluated through the Area Under the Curve (AUC) metric. For each experiment, we provide the 95% confidence intervals of the AUC, ensuring statistical significance. This was accomplished by conducting 1000 bootstrap resampling with replacement of the training dataset.

Figure 2 - Overview of case group characteristics. (A) Flowchart detailing the status of invasive carcinoma and its histological subtypes. (B) Pie chart illustrating the distribution of molecular subtypes among the invasive carcinomas with immunohistochemical examinations (57% of the cases); (C) Pie

chart depicting the T staging of invasive carcinomas of which the tumor size could be determined (89% of the cases).



Initially, our dataset comprised 13 CBC biomarkers, along with age, resulting in a total of 14 features. Subsequently, we calculated seven CBC-derived ratios, named neutrophils-lymphocyte ratio (NLR), derived neutrophils-lymphocyte ratio (dNLR), lymphocyte-monocyte ratio (LMR), platelet-lymphocyte ratio (PLR), systemic immune-inflammation index (SII), systemic inflammatory response index (SIRI), and aggregate index of systemic inflammation (AISI). Utilizing all 21 features, we proceeded to make an initial prediction for both algorithms.

Given that our dataset was complete with no missing values, no imputation or missing data treatment methods were necessary. Features were normalized based on the training set min–max values to ensure uniformity in scale. To prevent data leakage, hyperparameter tuning and feature selection were conducted solely within the modeling dataset. The testing dataset was reserved strictly for final evaluation.

Following this, we implemented a feature selection method guided by a directed acyclic graph. This approach begins by identifying all potential models with a single feature, selecting the one with the highest AUC value. We then extend our search to models with two features, again choosing the one with the best AUC. This process of enumerating models and selecting the best-performing one based on AUC continues, with one additional feature being incorporated at each step. The procedure halts once

the improvement in AUC no longer exceeds its standard deviation, ensuring a balance between model complexity and performance (10,11).

Given the low prevalence of breast cancer (BC) in the general population, we observed a significant imbalance in the target variable within both the modeling and testing sets, at 0.29% and 3.3%, respectively. To address this imbalance, we applied higher weights to the cases in the training of both models.

For Lightgbm explainability analysis, we used the SHapley Additive exPlanations (SHAP) (11) approach. SHAP evaluates feature importance by measuring the impact of omitting each feature on the model's decision. It produces SHAP values for every input, indicating each feature's significance in predictions.

After applying the best-performing model, we divided the population into four risk categories (high, moderate, average, and low) based on the model's probability outputs, following the methodology suggested by Michaels et al. (12). We used the relative risk (RR) metric to measure the likelihood of breast cancer occurrence in each group relative to the population of this group.

Results

Statistical analysis

This retrospective study included CBCs from 396,848 women aged 40–70 years, of which 2861 (0.72%) were diagnosed or highly suspicious of BC up to 6 months after the blood collection. We did not investigate demographic or clinical data, such as ethnicity, as these fall out of laboratory records. Table 1 shows the linear descriptive statistics for all primary features and the CBC ratios for the full dataset. P-values were calculated using the Student's t-test, and values ≤ 0.002 were considered significant (alpha was corrected from 0.05 to 0.001 using the Bonferroni test). Predictive AUCs were estimated by training a two-level decision tree on the training set and evaluating it on the test set. To estimate the 95% confidence interval, we performed 1000 bootstrap resamplings on the training set.

In addition to age, the biomarkers hemoglobin, hematocrit, mean corpuscular hemoglobin (MCH), and mean corpuscular volume (MCV) were significantly higher in women with BC, whereas lymphocyte levels were significantly lower compared to the control group. All CBC-derived ratios were significant and elevated in BC cases, except

for the LMR, which was lower. Age alone was the sole feature capable of distinguishing between the two groups, yielding a predictive AUC of 0.60 (95% CI 0.58–0.61).

Models' performance and selected features

The selected features for both models were the biomarkers neutrophils-lymphocyte ratio (NLR), red blood cells (RBC), plus age. The AUC performance of the developed models is reported in Table 2 for the validation and testing sets.

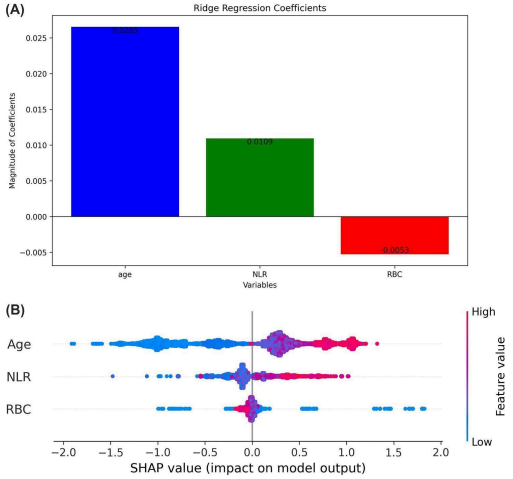
Table 1 - Descriptive statistics of blood markers and age for both groups of the full dataset. P-values ≤ 0.002 are in bold. Abbreviations: MCH = mean corpuscular hemoglobin; MCHC = mean corpuscular hemoglobin concentration; MCV = mean corpuscular volume; MPV = mean platelet volume; RBC = red blood count; RDW = red cell distribution width; AISI = aggregate index of systemic inflammation; dNLR = derived NLR; LMR = lymphocytes-to-monocytes ratio; NLR = neutrophils-to-lymphocytes ratio; PLR = platelets-to-lymphocytes ratio; SII = systemic immune-inflammation index; SIRI = systemic inflammation response index.

Exam	Feature	Control Mean ± STD	Case Mean ± STD	P value	Predictive AUC (95% CI)
	Age (years)	50.04 ± 8.56	53.92 ± 8.50	**0.000**	0.60 (0.58–0.61)
CBC	Eosinophils (/mm^3)	173.62 ± 149.65	170.59 ± 140.29	0.281	0.50 (0.49–0.51)
	Hematocrit (%)	39.84 ± 2.96	40.02 ± 3.18	**0.002**	0.51 (0.50–0.52)
	Hemoglobin (g/dL)	13.23 ± 1.09	13.31 ± 1.15	**0.000**	0.50 (0.49–0.52)
	Leukocytes (/mm^3)	6421.13±1894.44	6344.95 ± 1889.42	0.032	0.50 (0.50–0.51)
	Lymphocytes(/mm^3)	2079.67 ±670.97	2000.46 ± 648.53	**0.000**	0.51 (0.50–0.51)
	MCH (pg)	29.47 ± 2.08	29.66 ± 2.05	**0.000**	0.50 (0.49–0.51)
	MCHC (%)	33.19 ± 1.04	33.25 ± 0.98	0.005	0.50 (0.49–0.50)
	MCV (fL)	88.75 ± 5.19	89.19 ± 5.21	**0.000**	0.50 (0.50–0.51)
	Monocytes (/mm^3)	479.78 ± 155.22	480.73 ± 164.89	0.744	0.50 (0.50–0.51)
	Neutrophils (/mm^3)	3650.25±1468.55	3654.07 ± 1483.28	0.890	0.50 (0.49–0.50)
	Platelets (10^3/mm^3)	256.8 ± 57.2	261.9 ± 61.1	0.540	0.50 (0.49–0.52)
	RBC (10^6/mm^3)	4.50 ± 0.37	4.50 ± 0.40	0.831	0.51 (0.50–0.52)
	RDW (%)	13.33 ± 1.14	13.36 ± 1.14	0.067	0.48 (0.47–0.50)
CBC ratios	AISI (× 10^6)	236.8 ± 196.7	252.3.0 ± 219.4	**0.000**	0.51 (0.50–0.52)
	dNLR	1.37 ± 0.57	1.42 ± 0.62	**0.000**	0.51 (0.50–0.52)
	LMR	4.60 ± 2.03	4.45 ± 1.64	**0.000**	0.51 (0.50–0.51)
	NLR	1.87 ± 0.90	1.97 ± 1.01	**0.000**	0.51 (0.50–0.52)
	PLR	136.11 ± 47.12	143.06 ± 53.81	**0.000**	0.52 (0.50–0.53)
	SII (× 10^3)	481.4 ± 262.7	509.2 ± 287.6	**0.000**	0.51 (0.50–0.52)
	SIRI	921.53 ± 725.26	979.91 ± 836.79	**0.000**	0.51 (0.50–0.51)

Table 2 - AUC (95% CI) for both models and both feature scenarios

Model	Features	Validation (95% CI)	Testing (95% CI)
LightGBM	All features	0.64 (0.61–0.66)	0.62 (0.61–0.64)
	NLR, RBC and age	0.65 (0.63–0.67)	0.63 (0.62–0.64)
Ridge regression	All features	0.65 (0.64–0.66)	0.63 (0.62–0.64)
	NLR, RBC and age	0.66 (0.65–0.66)	0.64 (0.64–0.65)

Figure 3. Explainability analysis. (**A**) Bar graph displaying Ridge Regression coefficients. (**B**) SHAP summary plot for the LightGBM model, showing the effect of each feature on BC risk prediction. Features are depicted in the order of importance. Pink dots are associated with women for which the corresponding feature shows a relatively higher value and blue dots with the opposite. Positions further to the right signify a greater predicted BC risk. NLR = neutrophils-to-lymphocytes ratio; RBC = red blood cells.



Discussion

The possibility of early cancer detection through a simple blood draw has always attracted attention, which presents an opportunity for low-resource countries to improve screening programs. Our study showed that routine laboratory biomarkers from CBC could be useful for early BC risk identification (within six months of the diagnosis), enabling further prioritization for screening. Since CBCs are affordable and often performed during routine health check-ups, our approach would not significantly elevate health costs. It simply leverages a widely-used test to provide invaluable insight into BC risk, resulting in a negligible increase in cost relative to the potential benefits of early detection.

Statistical analysis revealed that age, hematocrit, hemoglobin, MCH, MCV, AISI, dNLR, NLR, PLR, SII, and SIRI were significantly elevated in women with BC, whereas

lymphocytes and the LMR were markedly lower. Given the statistical significance of all CBC-derived ratios, they present a valuable opportunity as cost- effective predictive biomarkers in BC. Furthermore, by harnessing AI, we developed a ridge regression model incorporating age, NLR, and RBC, achieving an AUC of 0.64 (95% CI 0.64–0.65), demonstrating superior performance over more complex models like LightGBM, while offering full interpretability.

To emulate the use of a risk stratification tool, we divided our study population into four risk groups: high, moderate, average, and low. This division was motivated by the necessity to effectively manage mammography queues, a challenge particularly relevant in countries like Brazil, where only 20% of women can access mammography via the public health system (SUS), a circumstance exacerbated by the COVID-19 pandemic. Given such resource scarcity or lengthy waiting times—a common situation in Brazil and many other countries—a risk stratification model for screening proves extremely beneficial. This stratification dictates which women should be screened first, starting with the high-risk group, followed by the moderate-risk group, and so forth, contingent on mammography resource availability. These groups were sorted based on the likelihood of BC occurrence in each group relative to the population of this group. Distinct risk profiles emerged, showing that the first to fourth groups had a BC probability of 1.99, 1.32, 1.02, and 0.42 times the typical risk. Hence, this model could serve as a tool for screening prioritization based on individual risk, enhancing the speed and efficiency of BC case identification, particularly in settings with limited resources. Importantly, the four-group categorization is a demonstrative application of the model's output. However, priority class configurations and allocations can be tailored to meet specific needs, such as accounting for demographic variations or adjusting to the screening capacities of various institutions.

While our dataset lacked specific indicators for identifying metastasis and lymph node status required for TNM staging, it is important to note that the BC study population in the testing set primarily consisted of early-stage carcinomas, as reflected by the fact that 24% of the identified carcinomas were in situ and 95% of the recorded invasive T stages were at T1-2. The explanatory analysis of our model reveals that elevated levels of age and NLR and lower values of RBC may be indicative of an increased risk of BC.

Advanced age is linked to breast cancer incidence due to cumulative exposure to risk factors like hormones and environmental carcinogens, genetic mutations, and cellular aging reducing DNA repair efficiency (1,2,3). High neutrophil-to-lymphocyte ratio (NLR) may indicate inflammation, with neutrophils potentially both promoting tumor growth by suppressing immune response and killing tumor cells when activated in tissues (13–15). High neutrophil levels in blood may suggest reduced tissue activation. Lymphocytes play a critical role in anti-tumor immunity, and a decrease in their count could indicate an impaired immune response. Higher NLR levels are associated with increased breast cancer risk and poorer prognosis for breast cancer and other solid tumors (16). While the role of red blood cells (RBC) in breast cancer is not fully understood, chronic inflammation linked to lower RBC count may promote cancer development and progression (17).

Conclusion

In conclusion, by leveraging CBC biomarkers and age, which are routinely obtained from standard blood tests conducted annually, we developed an AI-based tool for BC risk stratification. This tool provides a potential pathway to a risk-based screening strategy, thereby generating a tangible impact on public health. While clinical decision-making often relies on cutoff points, ML models can perform better by detecting non-linear relationships between markers, thereby revealing the power of hidden patterns. Thus, the same routine CBC blood test could offer additional value by feeding data into our ML tool, with no need for new or additional tests. This tool could be integrated into various clinical settings, either through an application where a physician inputs blood marker data or as an API linked directly to a laboratory or health operator conducting a blood test, assigning each test a BC risk score. To support its clinical application, external validation with diverse populations is of paramount importance.

Limitations and new approaches

Our study has strengths and limitations. First, the strengths include the fact that a total of 396,848 women coming from 309 laboratories spread in eight Brazilian states were included, with blood exams conducted using globally world-adopted equipment and largely standardized biochemical measurements in a College of American

Pathologists (CAP) accredited institution. Several limitations, however, should be noted. Information about previous comorbidities and use of drugs was not available, which could bias our study and interfere with the analyzed blood tests. Additionally, the lack of access to comprehensive demographic and clinical information, including ethnicity, could influence our results. Another critical limitation is the absence of external validation to corroborate our findings across different populations and settings. Furthermore, we acknowledge that the thresholds provided for risk stratification, while based on our study population, may require adjustment for application to populations with different age distributions, or to accommodate varying screening capacities and health care protocols of different institutions. This acknowledges the need for a flexible approach to applying our model, underscoring that different thresholds may be more appropriate depending on specific population characteristics or institutional capabilities.

Although additional markers are present in the retrospective blood tests, they are typically less common and thus do not allow us to conclude a large dataset like the CBC. In the future, we plan to enhance our model's performance by incorporating other markers. By doing so, we aim to explore whether cytokines, sex hormones, thyroid hormones, or other blood markers could improve the prediction. We can assess their potential impact more effectively as we collect more data on these additional markers. When selecting the high-risk group, we maximize specificity (approximately 90% specificity for 20% sensitivity). However, it is essential to note that our model is designed to stratify the risk for women within the recommended age for mammograms overdue for their examinations. In this scenario, unnecessary mammograms are not a concern because the model targets those who should be screened based on current guidelines. Our approach ensures that mammograms are administered to those most likely to benefit, optimizing available resources.

Acknowledgements

This project was funded by the São Paulo Research Foundation (FAPESP) [Grant numbers 2022/07614-3, 2022/13782-6, and 2022/16727-6], as well as by Produtos Roche Químicos e Farmacêuticos Data Applied Science Trail (dASTRO) 2022. KBG and AAV would like to thank the National Council for Scientific and Technological Development (CNPq) for the generous research fellowship.

References

1. Coleman C. Early Detection and Screening for Breast Cancer. Semin Oncol Nurs. 2017 May;33(2):141–55.

2. Araujo DC, Rocha BA, Gomes KB, da Silva DN, Ribeiro VM, Kohara MA, et al. Unlocking the complete blood count as a risk stratification tool for breast cancer using machine learning: a large scale retrospective study. Sci Rep. 2024 May 12;14(1):1–10.

3. Zhang K, Bangma CH, Venderbos LDF, Roobol MJ. Individual and Population-Based Screening. Management of Prostate Cancer. 2017;43–55.

4. Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Hered Cancer Clin Pract. 2012;10(Suppl 2):A29.

5. Yala A, Mikhael PG, Strand F, Lin G, Satuluru S, Kim T, et al. Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. J Clin Oncol [Internet]. 2022 Jun 1 [cited 2024 May 28];40(16). Available from: https://pubmed.ncbi.nlm.nih.gov/34767469/

6. Danesh H, Ziamajidi N, Mesbah-Namin SA, Nafisi N, Abbasalipourkabir R. Association between Oxidative Stress Parameters and Hematological Indices in Breast Cancer Patients. Int J Breast Cancer [Internet]. 2022 Oct 3 [cited 2024 May 29];2022. Available from: https://pubmed.ncbi.nlm.nih.gov/36225290/

7. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin. 2017 Mar;67(2):93–9.

8. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics [Internet]. 1970 Feb 1 [cited 2024 May 29]; Available from: https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634

9. Ke G, Meng Q, Finely T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems 30 (NIP 2017) [Internet]. 2017 [cited 2024 May 30]. Available from: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

10. Zuin G, Araujo D, Ribeiro V, Seiler MG, Prieto WH, Pintão MC, et al. Prediction of SARS-CoV-2-positivity from million-scale complete blood counts using machine learning. Communications Medicine. 2022 Jun 15;2(1):1–12.

11. Amador T, Saturnino S, Veloso A, Ziviani N. Early identification of ICU patients at risk of complications: Regularization based on robustness and stability of explanations. Artif Intell Med [Internet]. 2022 Jun [cited 2024 May 30];128. Available from: https://pubmed.ncbi.nlm.nih.gov/35534141/

12. Michaels E, Worthington RO, Rusiecki J. Breast Cancer: Risk Assessment, Screening, and Primary Prevention. Med Clin North Am [Internet]. 2023 Mar [cited 2024 May 30];107(2). Available from: https://pubmed.ncbi.nlm.nih.gov/36759097/

13. Ethier JL, Desautels D, Templeton A, Shah PS, Amir E. Prognostic role of neutrophil-to-lymphocyte ratio in breast cancer: a systematic review and meta-analysis. Breast Cancer Res [Internet]. 2017 [cited 2024 May 30];19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5217326/

14. De Larco JE, Wuertz BR, Furcht LT. The potential role of neutrophils in promoting the metastatic phenotype of tumors releasing interleukin-8. Clin Cancer Res [Internet]. 2004 Aug 1 [cited 2024 May 30];10(15). Available from: https://pubmed.ncbi.nlm.nih.gov/15297389/

15. Katano M, Torisu M. Neutrophil-mediated tumor cell destruction in cancer ascites. Cancer [Internet]. 1982 Jul 1 [cited 2024 May 30];50(1). Available from: https://pubmed.ncbi.nlm.nih.gov/7083126/

16. Gago-Dominguez M, Matabuena M, Redondo CM, Patel SP, Carracedo A, Ponte SM, et al. Neutrophil to lymphocyte ratio and breast cancer risk: analysis by subtype and potential interactions. Sci Rep [Internet]. 2020 Aug 6 [cited 2024 May 30];10(1). Available from: https://pubmed.ncbi.nlm.nih.gov/32764699/

17. Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related inflammation. Nature [Internet]. 2008 Jul 24 [cited 2024 May 30];454(7203). Available from: https://pubmed.ncbi.nlm.nih.gov/18650914/