Journal of Health Informatics

J Health Inform. 2026;18:e1810
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1810

ARTIGO COMPLETO
CBIS 2026

Breast cancer risk detection based on DNA methylation

Detecção de risco de câncer de mama com base em metilação do DNA

Leandro Carneiro de Oliveira[1] ✉, Wellington Pinheiro dos Santos[2], Maíra Araújo de Santana[2], José Barbosa de Araújo Neto[2]

ORCID: 0009-0004-0386-1233 · 0000-0003-2558-6602 · 0000-0002-1796-7862 · 0009-0007-4994-1695

[1] Universidade de Pernambuco (PE), Brazil.

[2] Universidade Federal de Pernambuco (PE), Brazil.

ARTICLE INFORMATION

TYPE
Artigo completo · CBIS 2026

IDENTIFIER
e1810

DOI
10.59681/2175-4411.v18.iEspecial.2026.1810

RECEIVED
17 May 2026

ACCEPTED
20 July 2026

REVIEW PROCESS
Double Blind Review

HANDLING EDITOR
Juliano de Souza Gaspar

ACCESS
Open, CC BY 4.0

CORRESPONDING AUTHOR
Leandro Carneiro de Oliveira
carneiro@outlook.ie

LICENSE

This is an open access article published under the Creative Commons CC BY 4.0 license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

CC BY 4.0

ABSTRACT

Objective: to investigate whether peripheral-blood DNA methylation signatures can anticipate breast cancer risk before clinical diagnosis, distinguishing healthy, at-risk and breast cancer patients. **Methods:** methylation profiles of 426 individuals from three Gene Expression Omnibus datasets were integrated and sample-level stratified into HEALTHY (n=79), RISK (n=282) and BRCA (n=65), with no instance overlap between partitions; SMOTE was applied only on the training set, and five tree-based ensemble models were compared. **Results:** AdaBoost reached 89.5% accuracy, a Cohen Kappa of 0.80 and an F1-score of 0.991 for the RISK class on the test set, which was intentionally left unbalanced to preserve the original class distribution. PCA and t-SNE revealed a distinct RISK cluster. **Conclusion:** these findings suggest a possible pre-tumorigenic epigenetic anomaly, to be confirmed with cross-validation and feature selection in future work.

Keywords: DNA Methylation; Breast Neoplasms; Machine Learning.

RESUMO

Objetivo: investigar se assinaturas de metilação do DNA em sangue periférico permitem distinguir mulheres saudáveis, em risco e com câncer de mama, antecipando a detecção do risco oncológico antes do diagnóstico clínico. **Método:** foram integrados perfis de metilação de 426 indivíduos de três datasets do Gene Expression Omnibus, estratificados por amostra em HEALTHY (n=79), RISK (n=282) e BRCA (n=65), com divisão treino/validação/teste sem sobreposição de instâncias entre partições; SMOTE foi aplicado apenas no treino, e cinco modelos ensemble baseados em árvores foram comparados. **Resultados:** o AdaBoost obteve acurácia de 89,5%, Kappa de 0,80 e F1-score de 0,991 para a classe RISK no teste, mantido propositalmente desbalanceado para preservar a distribuição original dos dados. PCA e t-SNE mostraram agrupamento distinto da classe RISK. **Conclusão:** os resultados sugerem uma possível anomalia epigenética pré-tumorigênica, a ser confirmada com validação cruzada e seleção de características em trabalhos futuros.

Descritores: Metilação de DNA; Neoplasias da Mama; Aprendizado de Máquina.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Paper presented at the 21st Brazilian Congress of Health Informatics (CBIS 2026), held from September 23 to 25, 2026, in Brasília (DF), Brazil, and published in the CBIS 2026 section of the Journal of Health Informatics.

HOW TO CITE THIS ARTICLE: de Oliveira LC, dos Santos WP, de Santana MA, de Araújo Neto JB. Breast cancer risk detection based on DNA methylation. J Health Inform 2026;18:e1810. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1810

1 INTRODUCTION

Breast cancer remains one of the leading public-health challenges worldwide because of its high incidence and impact on female mortality [1]. Although consolidated screening methods such as mammography exist, they detect the disease at stages when tumor alterations are already established, limiting truly early intervention. Inequalities in access to diagnosis make this scenario even more critical in vulnerable populations. Therefore, there is a growing need to identify non-invasive biomarkers capable of detecting biological changes that precede the clinical manifestation of the disease.

Epigenetic modifications, especially DNA methylation at CpG islands, emerge as promising biomarkers for early cancer detection because of their stability in peripheral blood, accessibility through non-invasive methods and the fact that they precede histopathological changes [2,3]. Aberrant methylation patterns can be detected in cytologically normal cells years before neoplastic transformation, suggesting that these signals represent risk markers rather than mere consequences of the tumor [4,5].

Detecting biomarkers in epigenetic profiles, however, creates computational challenges. Platforms such as the Illumina HumanMethylation27 BeadChip interrogate more than 27,000 CpG islands, a high-dimensional space in which an exhaustive search for informative combinations becomes computationally infeasible [6]. Tree-based ensemble methods such as Random Forest, Gradient Boosting, XGBoost, LightGBM and AdaBoost have gained prominence in clinical bioinformatics by combining strong predictive performance, robustness in high-dimensional scenarios and interpretability [7,8].

Previous studies have focused predominantly on binary classification scenarios such as healthy vs. cancer, often with tissue-specific data [9]. Although useful, these models do not adequately capture the temporal heterogeneity associated with pre-diagnostic progression. Distinguishing between healthy states, pre-clinical risk conditions and established malignancies therefore requires a multiclass framework that explicitly models methylation dynamics throughout cancer evolution.

This study investigates whether peripheral-blood DNA methylation signatures can anticipate breast cancer risk before clinical diagnosis. We propose an interpretable machine-learning framework based on tree ensembles to identify predictive biomarkers in a high-dimensional, low-sample setting. The approach integrates methylation profiles of 426 individuals from three Gene Expression Omnibus (GEO) repositories organized into three clinically relevant classes: HEALTHY, RISK and BRCA. The main contribution is to test a multiclass formulation able to distinguish a healthy state, a pre-diagnostic risk state and confirmed cancer instead of restricting analysis to the traditional binary task.

2 METHODS

2.1 Data acquisition and group definition

DNA methylation profiles were obtained from three public datasets deposited in GEO [10]. All samples were processed with the Illumina Infinium HumanMethylation27 BeadChip v1.2, which quantifies methylation levels at approximately 27,000 CpG sites in peripheral-blood cells from women. The first dataset (GSE58119) comprises 282 samples: 148 PRE-BRCA (women diagnosed within five years) and 134 HEALTHY-UNK (women not diagnosed within the same period). The second dataset (GSE57285) contains 84 samples (35 BRCA-MT, 7 HEALTHY-MT and 42 HEALTHY-WT). The third dataset (GSE32396) contains 60 samples (15 BRCA-MT, 15 BRCA-WT, 15 HEALTHY-MT and 15 HEALTHY-WT). Three classes were defined: BRCA (BRCA-MT + BRCA-WT, n=65), HEALTHY (HEALTHY-MT + HEALTHY-WT, n=79) and RISK (PRE-BRCA + HEALTHY-UNK, n=282).

2.2 Pre-processing

A 2-D matrix of Beta methylation values was generated from the raw data. Beta values were computed as $\beta = TI / (U + TI + \alpha)$, where TI is the total intensity of the methylated allele, U is the intensity of the unmethylated allele and $\alpha = 100$ stabilizes the values and prevents division by zero. Missing values were imputed with the mean across all samples for the same CpG island, preserving biological variability while avoiding sample exclusion. After filtering and labeling, the final dataset consisted of 426 samples and 27,578 CpG islands.

2.3 Stratification and class balancing with SMOTE

The dataset was split into independent training, validation and test partitions using a stratified 60/20/20 scheme that respected the original class distribution (Table 1). The minority class BRCA had only 35 training samples, below the 40-sample threshold adopted in this work; the synthetic minority oversampling technique (SMOTE) [11] was applied exclusively to the training partition, generating 5 synthetic BRCA samples and raising the count to 40, while HEALTHY (41) and RISK (179) remained untouched. Test and validation partitions were never resampled, so final evaluation reflects unmodified data.

Table 1. Class distribution across training, test and validation partitions.

Class	Training	Test	Validation
BRCA	35 (40 after SMOTE)	12	18
HEALTHY	41	19	19
RISK	179	55	48

Source: Prepared by the authors (2026).

Partitioning into training, validation and test sets was performed at the sample level, with each individual assigned to exactly one partition prior to any resampling step; this design prevents instance-level leakage between partitions. Validation and test sets preserved the original class distribution and were deliberately left unbalanced: resampling the test partition is not recommended, as it would bias evaluation metrics away from real-world class proportions. The residual imbalance in the test set (12 BRCA, 19 HEALTHY, 55 RISK) should therefore be kept in mind when interpreting weighted metrics, since RISK samples dominate that partition.

2.4 Models and training procedure

Five tree-based ensemble classifiers were selected: eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Gradient Boosting (GB) and AdaBoost. The choice was motivated by the need for interpretability, since the study aims to identify the most relevant CpG islands for classification. All models were configured for the multiclass problem (BRCA, HEALTHY, RISK) with random_state=42 for reproducibility, n_estimators=1000 and learning_rate=0.1 (where applicable). Models were trained exclusively on the SMOTE-balanced training partition; validation and test partitions remained untouched for final evaluation.

No explicit feature-selection step was applied prior to training. This was a deliberate choice: the goal was to observe classification behavior across the full CpG-island space without an a priori assumption about which features were most informative. Feature selection (and the interpretability gains it could bring) is left for future work.

2.5 Evaluation metrics

Model performance was evaluated using accuracy, Cohen Kappa, weighted sensitivity, weighted precision, weighted F1-score and per-class F1-score (BRCA, HEALTHY, RISK). Weighted averaging was used to handle residual class imbalance. Confusion matrices on the test partition were inspected to characterize error patterns of the best model.

2.6 Rationale for a single train/test split

A percentage-based train/validation/test split was adopted instead of k-fold cross-validation. Given the modest dataset size (426 samples, only 65 in the BRCA class), cross-validation with a high fold count would yield small, statistically fragile test folds, risking unstable estimates that swing between overly optimistic and overly pessimistic results. The single-split design instead prioritized a test partition large enough (n=86) for statistically meaningful evaluation. Repeated or stratified k-fold cross-validation is planned for future work, contingent on additional data collection.

3 RESULTS

3.1 Exploratory analysis (PCA and t-SNE)

Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were applied to assess natural separability of the samples. The first two principal components retained 32.4% and 21.3% of the total variance, respectively (53.7% combined). BRCA and HEALTHY samples partially overlap in the projection, while the RISK group occupies a distinct region, suggesting a pre-tumorigenic epigenetic anomaly. In Figure 1, it is possible to notice two sub-clusters further visible inside the RISK group, a finding that warrants future investigation.

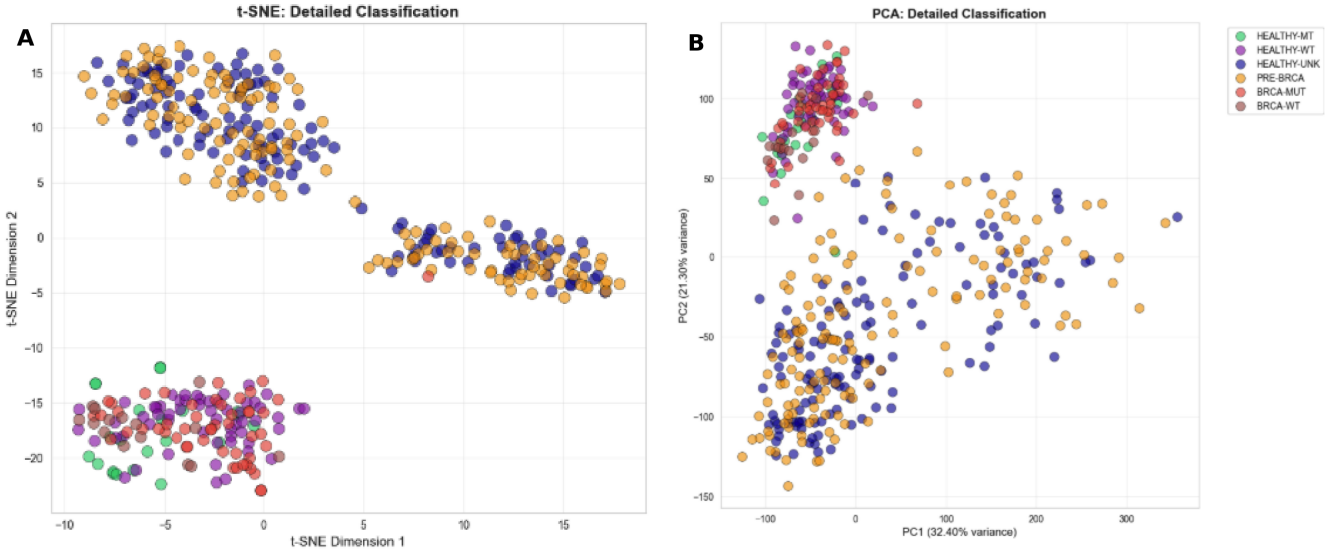
3.2 Model performance and comparison

Tables 2 and 3 summarize the performance of the five ensembles on the test partition. AdaBoost achieved the best overall results: 89.53% accuracy, Cohen Kappa of 0.7979 and weighted F1-score of 0.8940 on the test set. The model classified the RISK class with high accuracy (F1 = 0.9910 on test), while F1 was 0.6667 for BRCA and 0.7568 for HEALTHY, indicating greater confusion between these two classes.

3.3 Statistical validation (Kruskal-Wallis)

A non-parametric Kruskal-Wallis test, complemented by pairwise post-hoc comparisons, was applied to the 100 most important CpG islands. Of these, 95 (95.0%) showed statistically significant differences ($p < 0.05$) across the three groups. Pairwise analyses revealed: HEALTHY vs BRCA only 28% of sites differ; HEALTHY vs RISK 90% differ; BRCA vs RISK 93% differ. The RISK group therefore presents the most distinctive methylation profile, supporting the hypothesis of an epigenetic anomaly detectable before clinical diagnosis.

Figure 1. t-SNE and PCA plot analysis.



Source: Prepared by the authors (2026).

Table 2. Performance metrics of ensemble models on the test set.

Model	Accuracy	Kappa	Sensitivity	Precision	F1-Score
AdaBoost	0.8953	0.7979	0.8953	0.893	0.894
Random Forest	0.8837	0.7694	0.8837	0.8768	0.8762
LightGBM	0.8488	0.707	0.8488	0.844	0.8456
XGBoost	0.8605	0.7254	0.8605	0.8535	0.8554
Gradient Boosting	0.8256	0.6514	0.8256	0.8055	0.8136

Source: Prepared by the authors (2026).

Table 3. Performance metrics of ensemble models on the test set (per class F1-score).

Model	F1-Score (BRCA)	F1-Score (HEALTHY)	F1-Score (RISK)
AdaBoost	0.6667	0.7568	0.9910
Random Forest	0.6000	0.7692	0.9735
LightGBM	0.5455	0.6667	0.9730
XGBoost	0.5833	0.6857	0.9735
Gradient Boosting	0.3810	0.6486	0.9649

Source: Prepared by the authors (2026).

4 DISCUSSION

The high accuracy and Kappa value (0.80) achieved on a relatively large RISK class (n=282) indicate that the model captured a specific, complex pattern rather than the majority class. The pattern of confusion between BRCA and HEALTHY suggests that these two populations share partially stable methylation levels even though they sit at opposite ends of the disease spectrum; one possible explanation is that diagnosed patients may have started clinical treatment that partially restored normal patterns [12,13]. The high accuracy of the RISK classification implies that the trajectory from healthy to cancer is not a linear continuum for these methylation markers, but rather crosses a transitional pre-cancerous state.

In tumorigenesis terms, this detectable anomaly can be interpreted as an epigenetic field defect: a generalized methylation shift in apparently normal tissue years before a tumor becomes clinically detectable. This view is consistent with growing evidence that cancer is not solely a genetic disease and is also driven by epigenetic alterations.

Within the Brazilian public health system (SUS), this blood-based DNA methylation assay could potentially be operationalized as a scalable, low-cost triage test within the primary care network to screen asymptomatic women for elevated epigenetic risk, extending beyond the current reliance on strict family-history criteria. Women with a positive epigenetic signature could then be referred into the same high-

risk surveillance protocols, including annual breast magnetic resonance imaging (MRI) and specialized genetic counseling, currently reserved for BRCA1/2 mutation carriers, potentially expanding equitable access to preventive care across the public system. Prospective, multicenter validation would be required before any such triage pathway could be considered for clinical deployment.

The residual class imbalance in the test partition (RISK = 55 of 86 samples) means weighted metrics are pulled toward RISK performance; the high RISK F1-score (0.991) should be read with this in mind rather than as evidence the model performs equally well across all three classes. Reporting per-class F1 (Table 3) alongside weighted metrics was intended to mitigate this, but future work with stratified or repeated cross-validation would allow tighter confidence bounds on this estimate.

The study has limitations. Samples were aggregated from multiple sources with different mutational backgrounds; future work should design classifiers able to distinguish individual subgroups (HEALTHY-MT, HEALTHY-WT, BRCA-MT, BRCA-WT, PRE-BRCA, HEALTHY-UNK). A formal feature-importance analysis on the AdaBoost model and pathway enrichment over the top CpG islands are necessary to confirm the underlying biology (notably BRCA1 promoter hypermethylation, cell-cycle and DNA-repair pathways). Finally, the relatively small absolute size of the HEALTHY and BRCA classes still calls for additional samples or synthetic-data strategies in future studies.

5 CONCLUSIONS

This work investigated detectable anomalies in peripheral-blood DNA methylation that may anticipate breast cancer years before tumor manifestation. A pipeline comprising pre-processing, stratified splitting, SMOTE balancing of the minority class, training of five tree-based ensembles, exploratory analysis (PCA and t-SNE) and Kruskal-Wallis hypothesis testing was implemented. AdaBoost delivered the best performance, with 89.5% test accuracy, Kappa of 0.80 and a high F1 of 0.991 for the RISK class. PCA and t-SNE projections visually confirmed that RISK samples cluster apart from both HEALTHY and BRCA. Statistical tests reinforced that the RISK group differs significantly from the others on most of the most important CpG sites. Together, these results support the existence of a pre-tumorigenic epigenetic signature and pave the way for risk-screening biomarkers enabling personalized prevention strategies for high-risk women. Future work includes collecting more HEALTHY and BRCA samples, investigating the two sub-clusters observed within the RISK group and extending the methodology to other cancers associated with DNA-repair gene mutations (ovarian, colon, endometrial).

Author contributions. The authors contributed equally to all stages of the article: conception and design of the study; data collection, analysis, and interpretation; drafting and critical revision of the manuscript; and approval of the final version.

Acknowledgments. The authors thank the Programa de Pós-Graduação em Engenharia de Computação of Universidade de Pernambuco for the academic support.

Funding. The authors did not declare any source of funding.

Data and materials availability. The authors did not provide a data and materials availability statement.

Conflicts of interest. The authors did not declare any conflicts of interest.

Use of artificial intelligence. The authors declare that the use of AI in this work complies with the ethical guidelines of the institution. AI tools were employed for language refinement and text revision only. All AI-generated or AI-modified content was rigorously reviewed, verified and approved by the authors. Predominant intellectual contribution is human and the scientific integrity of the work has been preserved.

REFERENCES

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3): 229-63. doi: https://doi.org/10.3322/caac.21834
2. Dong Y, Zhao H, Li H, Li X, Yang S. DNA methylation as an early diagnostic marker of cancer (Review). Biomed Rep. 2014;2(3):326-30. doi: https://doi.org/10.3892/br.2014.237
3. Hum M, Lee ASG. DNA methylation in breast cancer: early detection and biomarker discovery through current and emerging approaches. J Transl Med. 2025;23(1):465. doi: https://doi.org/10.1186/s12967-025-06495-2
4. Tahara T, Maegawa S, Chung W, Garriga J, Jelinek J, Estécio MR, et al. Examination of whole blood DNA methylation as a potential risk marker for gastric cancer. Cancer Prev Res (Phila). 2013;6(10):1093-100. doi: https://doi.org/10.1158/1940-6207.CAPR-13-0034
5. Wachsman JT. DNA methylation and the association between genetic and epigenetic changes: relation to carcinogenesis. Mutat Res. 1997;375(1):1-8. doi: https://doi.org/10.1016/s0027-5107(97)00003-1
6. Sharma M, Verma RK, Kumar S, Kumar V. Computational challenges in detection of cancer using cell-free DNA methylation. Comput Struct Biotechnol J. 2022;20:26-39. doi: https://doi.org/10.1016/j.csbj.2021.12.001

7. Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785-94. doi: https://doi.org/10.1145/2939672.2939785

8. Schapire R. Explaining AdaBoost. In: Empirical Inference. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 37-52. doi: https://doi.org/10.1007/978-3-642-41136-6_5

9. Ma B, Chai B, Dong H, Qi J, Wang P, Xiong T, et al. Diagnostic classification of cancers using DNA methylation of paracancerous tissues. Sci Rep. 2022;12(1):10646. doi: https://doi.org/10.1038/s41598-022-14786-7

10. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207-10. doi: https://doi.org/10.1093/nar/30.1.207

11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair. 2002;16:321-57. doi: https://doi.org/10.1613/jair.953

12. Davalos V, Esteller M. Cancer epigenetics in clinical practice. CA Cancer J Clin. 2023;73(4):376-424. doi: https://doi.org/10.3322/caac.21765

13. Anjum S, Fourkala EO, Zikan M, Wong A, Gentry-Maharaj A, Jones A, et al. A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. Genome Med. 2014;6(6):47. doi: https://doi.org/10.1186/gm567

EDITORIAL NOTES

The Journal of Health Informatics is a scientific journal maintained by the Brazilian Society of Health Informatics (Sociedade Brasileira de Informática em Saúde), a non-profit scientific association dedicated to the development of Health Informatics and Digital Health in Brazil. The editorial process is conducted with scientific independence, free from institutional, commercial, or sponsor interference, and follows international best practices in publication ethics.

This article was double-blind peer reviewed by specialists in the field. The editorial team used artificial intelligence tools, under human supervision, in the initial screening, in checking the text, references and numbers and, when needed, in redrawing figures, charts and tables, which the authors approved in the proof. The tools and safeguards are described in the journal's Policy on the use of artificial intelligence. Editorial decisions were made by people.

SBIS Sociedade Brasileira de Informática em Saúde