

Prediction of Metastatic Risk in Breast Cancer by the Expression of Mechanobiological Markers [†]

Ksenia Maksimova ^{1,*}, Margarita Pustovalova ¹, Sergey Leonov ¹  and Yulia Merkher ^{1,2,*} 

¹ Institute of Future Biophysics, Moscow Center for Advanced Studies, 123592 Moscow, Russia; pustovalova.mv@mipt.ru (M.P.); leonov.sv@mipt.ru (S.L.)

² Department of Biomedical Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel

* Correspondence: maksimova.kv@phystech.edu (K.M.); merkher@technion.ac.il (Y.M.)

[†] Presented at the 6th International Electronic Conference on Applied Sciences, 9–11 December 2025; Available online: <https://sciforum.net/event/ASEC2025>.

Abstract

Distant metastasis is the leading cause of breast cancer-related mortality, yet its prediction from primary tumor profiles remains challenging. Cytoskeletal remodeling and cell motility are central to metastatic dissemination, suggesting mechanobiological genes as biologically relevant biomarkers. Here, we evaluated the ability of supervised machine learning models to distinguish metastatic from non-metastatic breast cancer samples using expression profiles of 11 actin cytoskeleton-related genes from the TCGA cohort. Ensemble models, particularly Random Forest and XGBoost, demonstrated strong discriminative ability and consistently outperformed other approaches after employing SMOTE for class balancing in exploratory analyses. *CFL1*, *ANXA2*, and *MYH9* consistently emerged as the most informative predictors, highlighting mechanobiological processes as key drivers of metastatic risk.

Keywords: metastasis; mechanobiology; machine learning; cytoskeletal remodeling; gene markers

1. Introduction

Breast cancer (BC) is among the most common malignancies worldwide, with metastasis being the main cause of cancer-related death [1]. Accurate prediction of metastatic risk at the time of diagnosis is crucial for improving patient outcomes through risk stratification and treatment optimization. Current prediction methods rely on conventional staging systems (TNM classification, tumor grade, and lymph node status) that assess anatomical and histopathological features. Commercial gene expression assays such as OncotypeDX [2] (21-gene panel for HR+/HER2- breast cancer) and MammaPrint [3] (70-gene panel) have improved prognostic stratification and treatment planning. These assays are primarily based on proliferation, hormone receptor signaling, and cell cycle-related pathways. However, their predictive performance varies across cancer types and subtypes, and these molecular tests are costly and often overlook dynamic cellular behaviors such as cytoskeletal remodeling and mechanotransduction that are critical for metastatic dissemination.

Migration, motility, and actin cytoskeleton remodeling are fundamental features of cells with high metastatic potential [4,5]. The acquisition of this phenotype is accompanied by changes in gene expression profiles. Based on our previous research [4,6], we selected a panel of actin cytoskeleton-related genes as potential biomarkers for metastasis prediction: *ACTN4*, *ANXA2*, *ANXA6*, *CFL1*, *ECM1*, *EZR*, *FSCN1*, *GSN*, *MYH9*, *PFN1*, and *TAGLN*.



Academic Editor: Roger Narayan

Published: 10 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Machine learning approaches have shown promise in cancer outcome prediction [7]. However, the predictive potential of focused mechanobiological gene panels using state-of-the-art ensemble algorithms has not been systematically evaluated in clinical patient cohorts.

In this study, we investigated whether expression levels of 11 mechanobiological genes in primary breast tumors can accurately predict distant metastasis. We applied supervised machine learning models to transcriptomic data from the TCGA Breast Invasive Carcinoma dataset and compared five algorithms, including linear methods (Logistic Regression), tree-based approaches (Decision Trees, Random Forest, and XGBoost), and instance-based methods (k-Nearest Neighbors), to identify both high-performing predictive models and the most informative genes contributing to metastatic risk.

2. Materials and Methods

2.1. Dataset

The Breast Invasive Carcinoma dataset (TCGA, Firehose Legacy) was obtained from the cBioPortal database [8], containing genomic and clinical information from primary breast cancer tumors. Gene expression data for 11 previously identified genes [6] (*ACTN4*, *ANXA2*, *ANXA6*, *CFL1*, *ECM1*, *EZR*, *FSCN1*, *GSN*, *MYH9*, *TAGLN*, and *PFN1*) were analyzed. These genes encode proteins involved in cytoskeleton regulation, cell motility and extracellular matrix remodeling. The predictive target was distant metastasis status according to the American Joint Committee on Cancer (AJCC) staging [9], where M0 indicates the absence of distant metastases and M1 indicates their presence.

2.2. Data Preprocessing

Expression data was normalized as $\log_2(\text{TPM} + 1)$, where TPM represents transcripts per million. Due to class imbalance (902 M0 vs. 22 M1 samples), Synthetic Minority Over-sampling Technique (SMOTE) [10] was applied to the metastatic samples to prevent model bias toward the majority class. Following SMOTE application, the dataset contained 902 samples per class. The dataset was split into training (80%) and test (20%) sets. SMOTE was applied prior to the train–test split in order to increase the number of metastatic samples available for model training and evaluation. We acknowledge that oversampling the full dataset before splitting may lead to optimistic performance estimates due to information leakage; this limitation is explicitly addressed in the Discussion section. Per-gene z-score standardization was performed using scaling parameters (mean and standard deviation) derived from the training set, which were then applied to the test set to prevent information leakage.

2.3. Machine Learning Models

Five supervised machine learning algorithms were evaluated: k-Nearest Neighbors (kNN), Logistic Regression, Decision Trees, Random Forest, and XGBoost. Hyperparameters were optimized using Grid Search and Optuna with 10-fold cross-validation on the training set. Models were compared on the test set using accuracy, precision, sensitivity, specificity and F1-score. Following class balancing via SMOTE, accuracy was considered an appropriate primary metric due to the equal representation of both classes in the dataset. Additionally, F1-score was used to assess the balance between precision and sensitivity, which is particularly relevant in clinical settings where both false negatives and false positives carry significant consequences. Feature importance was assessed using model-specific methods to identify the most informative genes associated with metastasis. For Logistic Regression, feature importance was derived from the magnitude of model coefficients. For tree-based models feature importance was computed based on impurity reduction or gain,

depending on the algorithm. Since feature importance metrics differ across model types, importance scores were normalized using min-max scaling to a 0–1 range within each model to facilitate qualitative comparison of feature rankings. Given the exploratory nature of this study and the extreme rarity of metastatic cases, model evaluation was primarily used to compare algorithms and identify informative features rather than to establish clinically deployable predictive performance.

2.4. Software

All analyses were performed in RStudio (2023 Posit Software, PBC, Boston, MA, USA) using R (v.4.3.3) and packages: smotefamily (v.1.4.0), caret (v.7.0-1), rpart (v.4.1.24), xgboost (v.1.7.11.1), reticulate (v.1.43.0), tidyr (v.1.3.1), and ranger (v.0.17.0).

3. Results

We analyzed expression data for 11 mechanobiological genes from 924 primary breast cancer tumors in the TCGA dataset (902 M0; 22 M1). To address class imbalance, SMOTE was applied, resulting in 902 samples for both M0 and M1 classes. We employed five supervised machine learning algorithms to evaluate the predictive potential of mechanobiological gene signatures: Logistic Regression, a linear model estimating the probability; Decision Trees, which recursively partition the feature space using threshold-based rules; Random Forest, an ensemble method aggregating multiple decision trees to improve predictive performance and reduce overfitting; XGBoost, a gradient boosting algorithm that builds additive decision trees; and k-Nearest Neighbors, an instance-based method classifying samples based on similarity to neighboring data points.

3.1. Model Performance

Model performance was evaluated by the following parameters: accuracy (overall classification correctness), precision (positive prediction reliability), sensitivity (true positive rate), specificity (true negative rate), and F1-score (balance between precision and sensitivity). Comprehensive performance metrics for all five machine learning models are presented in Figure 1; higher score represents better performance.

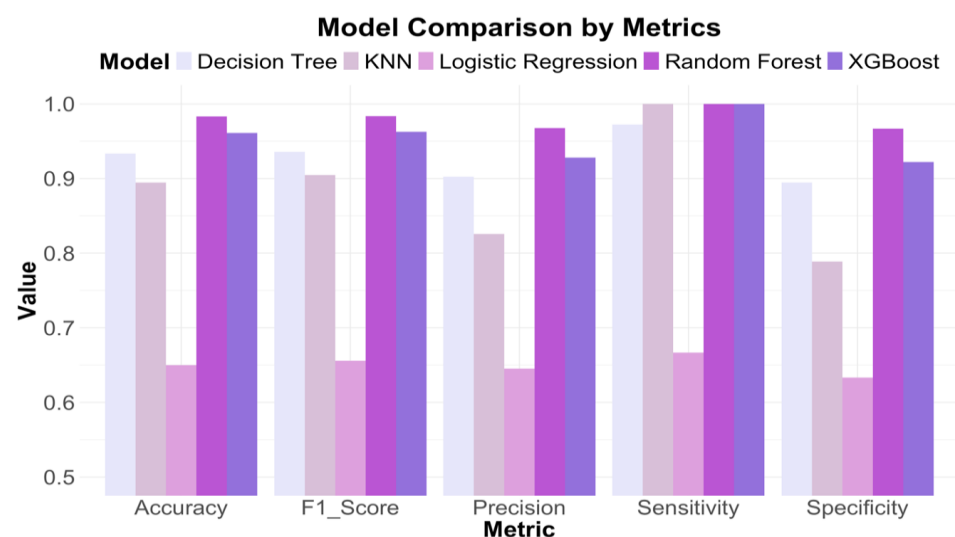


Figure 1. Comparison of classification model performance across five evaluation metrics: accuracy, F1-score, precision, sensitivity, and specificity. The bar chart summarizes results for Decision Tree, KNN, Logistic Regression, Random Forest, and XGBoost models. Higher score represents better performance.

The ensemble methods (Random Forest and XGBoost) substantially outperformed all other approaches across all evaluation metrics, with all scores exceeding 0.92. Random Forest emerged as the best-performing model, achieving near-perfect scores: accuracy = 0.98; sensitivity = 1.0; specificity = 0.97; precision = 0.97; F1-score = 0.98. Notably, Random Forest demonstrated the most balanced performance across sensitivity and specificity, which is clinically important for minimizing both false negatives (missed metastases) and false positives (unnecessary patient interventions). XGBoost ranked second with accuracy = 0.96; sensitivity = 1.0; specificity = 0.92; precision = 0.92; F1-score = 0.96. While slightly lower than Random Forest, XGBoost maintained excellent predictive performance across all metrics.

An interesting pattern emerged with k-Nearest Neighbors, which achieved high sensitivity (1.0) comparable to ensemble methods, but substantially lower specificity (0.79). This imbalance suggests that kNN tends to over-predict metastatic cases, potentially leading to false positives. Decision Trees showed moderate performance (accuracy = 0.93), demonstrating the substantial benefit of ensemble approaches over single decision trees.

Logistic Regression exhibited the poorest performance across all metrics (accuracy = 0.65; sensitivity = 0.67; specificity = 0.63), strongly suggesting that the relationship between mechanobiological gene expression and metastasis status is highly non-linear.

3.2. Gene Importance Analysis

Given the strong predictive performance, we examined which genes contribute most significantly to metastasis to prioritize candidates for further validation and identify potential therapeutic targets. Normalized feature importance scores are presented in Figure 2.

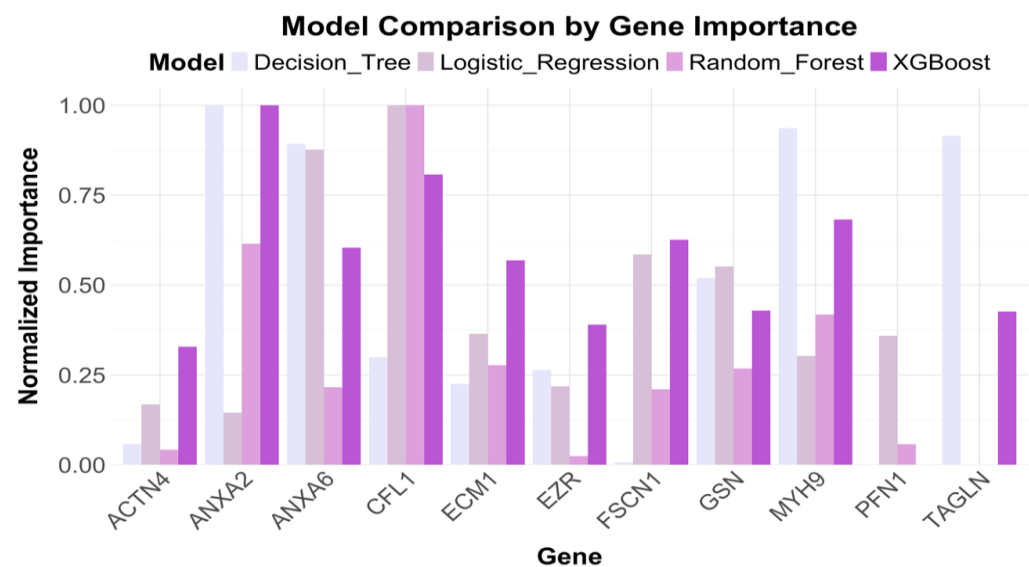


Figure 2. Comparison of normalized gene importance scores across machine learning models. The bar chart displays feature importance for 11 mechanobiological genes evaluated by Decision Tree, Logistic Regression, Random Forest, and XGBoost models. Higher scores indicate greater contribution to metastasis prediction. Importance values were normalized using min-max scaling (0–1 range) for cross-model comparison.

CFL1 emerged as the most influential predictor across all models, achieving near-maximal importance in Random Forest (1.0) and XGBoost (0.81). *ANXA2* and *MYH9* also demonstrated consistently high importance in top-performing models, suggesting their critical role in metastasis prediction and making them high-priority candidates for further investigation. Interestingly, *TAGLN*, *FSCN1* and *ECM1* exhibited model-dependent patterns, showing substantially higher importance in XGBoost compared to Random Forest (0.63 and 0.28 for *FSCN1*; 0.57 and 0.28 for *ECM1*). This finding is particularly noteworthy as

these genes were previously identified in DepMap cell line analyses [11] as having the highest prognostic value. Other genes showed consistently lower importance, suggesting their expression levels are less directly associated with distant metastasis development.

4. Discussion

This study demonstrates that machine learning models trained on mechanobiological gene expression can accurately predict distant breast cancer metastasis. Random Forest achieved the highest performance (accuracy = 0.98) with balanced sensitivity and specificity, while *CFL1*, *ANXA2*, and *MYH9* emerged as the most robust predictors. These results align with previous studies demonstrating the superior performance of Random Forest in cancer gene expression analysis. For example, Duan et al. reported that Random Forest outperformed other classifiers, achieving accuracy of 0.936 in predicting distant metastasis in breast cancer using a 21-gene expression panel [12]. Similarly, Ayubi et al. found that Random Forest reached the highest accuracy (0.97) in predicting colorectal cancer metastasis [13].

The ensemble methods achieved high accuracy (>0.96), comparing favorably with recent machine learning approaches while using a focused 11-gene panel rather than broader signatures [14]. The superior performance of Random Forest over Logistic Regression confirms that gene–metastasis relationships are highly non-linear, consistent with the complex molecular mechanisms underlying metastatic progression [15]. Class imbalance is a common challenge in biomedical datasets, and oversampling approaches such as SMOTE are widely used in biological and clinical research to improve classification performance on minority classes [16,17]. In this study, the balanced performance achieved through SMOTE and ensemble methods addresses a critical clinical requirement: minimizing both false negatives (missed metastases) and false positives (unnecessary interventions).

The consistent high importance of *CFL1*, *ANXA2*, and *MYH9* aligns with their established biological roles in metastases. *CFL1* (cofilin-1) regulates actin filament dynamics and is essential for cell migration and invasion [18]. *ANXA2* (annexin A2) facilitates membrane trafficking and has been implicated in promoting tumor cell invasion and angiogenesis [19]. *MYH9* (myosin-9) is a cytokine with an important role in the formation of cellular pseudopodia [20]. The robustness of these genes across different model architectures strongly suggests they represent core mechanobiological processes required for metastatic dissemination, making them high-priority candidates for further investigation and potential therapeutic targeting.

Interestingly, *TAGLN*, *FSCN1*, and *ECM1* showed higher importance in XGBoost than Random Forest, despite both being ensemble methods. This pattern mirrors our previous cell line findings [11], where these genes demonstrated the highest prognostic value and represented optimal therapeutic targets. The concordance between XGBoost feature importance and prior findings suggests that this model better captures the biological relevance of these biomarkers in patient samples, highlighting them as promising candidates for clinical validation. This pattern aligns with studies showing that XGBoost's feature importance ranking effectively identifies biologically relevant biomarkers: Li et al. used XGBoost feature importance to discover a 6-gene signature for breast cancer metastasis, with subsequent in vitro validation confirming the biological roles of identified genes [21], while Ghuriani et al. demonstrated that XGBoost-ranked features consistently revealed cancer-specific biomarkers across multiple cancer types, with pathway analysis validating their biological significance [22].

These findings enable several clinical applications: a streamlined diagnostic panel for risk stratification, identification of therapeutic targets, and a balanced predictive tool suitable for clinical decision-making. However, several limitations of this study should be

acknowledged. First, the number of patients with distant metastasis in the TCGA cohort is extremely small (22 cases), reflecting the rarity of metastatic disease at the time of primary tumor sampling. To mitigate this imbalance, SMOTE was applied prior to the train–test split, which allowed stable model training but may introduce information leakage and artificially optimistic performance estimates. When we tested a leakage-free alternative strategy, where the dataset was first split and SMOTE was applied only to the training set, the test set contained only 3–4 metastatic samples. Under this setting, model performance became highly unstable and substantially decreased, particularly for non-linear ensemble models, making reliable estimation of generalization performance infeasible. Therefore, the reported performance metrics should be interpreted as exploratory, upper-bound estimates. The primary aim of this study is biomarker identification and hypothesis generation, supported by the consistent ranking of informative genes across models, rather than the development of a clinically validated predictive classifier. For an accurate assessment and clinical generalizability, future studies should involve larger independent patient cohorts with enough metastatic cases.

In conclusion, mechanobiological gene signatures analyzed using ensemble machine learning show promise for identifying metastasis-associated expression patterns and prioritizing candidate biomarkers, warranting further validation in larger independent cohorts before clinical translation.

Author Contributions: Conceptualization, Y.M. and S.L.; methodology, Y.M.; software, K.M.; formal analysis, K.M. and Y.M.; investigation, K.M.; resources, M.P.; writing—original draft preparation, K.M.; writing—review and editing, Y.M. and S.L.; visualization, K.M.; supervision, Y.M. and S.L.; project administration, M.P.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of the Russian Federation (State Task) 075-03-2026-305, Project No. FSMG-2026-0018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.cbiportal.org/study/summary?id=brca_tcg (accessed on 12 January 2025).

Acknowledgments: During the preparation of this manuscript/study, the authors used Claude (Sonnet 4.5) and ChatGPT (OpenAI, GPT-4.1) to assist with language editing and manuscript structuring. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BC	Breast cancer
AJCC	American Joint Committee on Cancer
TPM	Transcripts per million
SMOTE	Synthetic Minority Over-sampling Technique

References

1. Fares, J.; Fares, M.Y.; Khachfe, H.H.; Salhab, H.A.; Fares, Y. Molecular principles of metastasis: A hallmark of cancer revisited. *Signal Transduct. Target. Ther.* **2020**, *5*, 28. [\[CrossRef\]](#)
2. Sparano, J.A.; Gray, R.J.; Makower, D.F.; Pritchard, K.I.; Albain, K.S.; Hayes, D.F.; Geyer, C.E., Jr.; Dees, E.C.; Goetz, M.P.; Olson, J.A.; et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N. Engl. J. Med.* **2018**, *379*, 111–121. [\[CrossRef\]](#)
3. Cardoso, F.; van't Veer, L.J.; Bogaerts, J.; Slaets, L.; Viale, G.; Delaloge, S.; Pierga, J.Y.; Brain, E.; Causeret, S.; Delorenzi, M.; et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N. Engl. J. Med.* **2016**, *375*, 717–729. [\[CrossRef\]](#)
4. Merkher, Y.; Kontareva, E.; Bogdan, E.; Achkasov, K.; Maximova, K.; Grolman, J.M.; Leonov, S. Encapsulation and adhesion of nanoparticles as a potential biomarker for TNBC cells metastatic propensity. *Sci. Rep.* **2023**, *13*, 12289. [\[CrossRef\]](#)
5. Jonas, O.; Mierke, C.T.; Käs, J.A. Invasive cancer cell lines exhibit biomechanical properties that are distinct from their noninvasive counterparts. *Soft Matter* **2011**, *7*, 11488–11495. [\[CrossRef\]](#)
6. Leonov, S.; Inyang, O.; Achkasov, K.; Bogdan, E.; Kontareva, E.; Chen, Y.; Fu, Y.; Osipov, A.N.; Pustovalova, M.; Merkher, Y. Proteomic Markers for Mechanobiological Properties of Metastatic Cancer Cells. *Int. J. Mol. Sci.* **2023**, *24*, 4773. [\[CrossRef\]](#)
7. Abbas, G.H.; Khouri, E.R.; Thaher, O.; Taha, S.; Vladimirov, M.; Oviedo, R.J.; Schmidt, J.; Bausch, D.; Pouwels, S. Predictive modeling for metastasis in oncology: Current methods and future directions. *Ann. Med. Surg.* **2025**, *87*, 3489–3508. [\[CrossRef\]](#) [\[PubMed\]](#)
8. cBioPortal. Available online: <https://www.cbioportal.org> (accessed on 12 January 2025).
9. Zhu, H.; Doğan, B.E. American Joint Committee on Cancer's Staging System for Breast Cancer, Eighth Edition: Summary for Clinicians. *Eur. J. Breast Health* **2021**, *17*, 234–238. [\[CrossRef\]](#)
10. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
11. Maksimova, K.; Leonov, S.; Pustovalova, M.; Merkher, Y. Mechanobiological Markers for Theragnostics in Breast Cancer. *JCO Glob. Oncol.* **2024**, *10*, 119. [\[CrossRef\]](#)
12. Duan, H.; Zhang, Y.; Qiu, H.; Fu, X.; Liu, C.; Zang, X.; Xu, A.; Wu, Z.; Li, X.; Zhang, Q.; et al. Machine learning-based prediction model for distant metastasis of breast cancer. *Comput Biol Med.* **2024**, *169*, 107943. [\[CrossRef\]](#)
13. Ayubi, E.; Farashi, S.; Tapak, L.; Afshar, S. Development and validation of a biomarker-based prediction model for metastasis in patients with colorectal cancer: Application of machine learning algorithms. *Heliyon* **2025**, *11*, e41443. [\[CrossRef\]](#)
14. Yagin, B.; Yagin, F.H.; Colak, C.; Inceoglu, F.; Kadry, S.; Kim, J. Cancer Metastasis Prediction and Genomic Biomarker Identification through Machine Learning and eXplainable Artificial Intelligence in Breast Cancer Research. *Diagnostics* **2023**, *13*, 3314. [\[CrossRef\]](#)
15. Sethi, N.; Kang, Y. Unravelling the complexity of metastasis—Molecular understanding and targeted therapies. *Nat. Rev. Cancer* **2011**, *11*, 735–748. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Hong, J.E.; Kim, Y.E.; Kang, Y.S.; Choi, D.H.; Ahn, S.H.; An, J. SMOTE-augmented machine learning model predicts recurrent and metastatic breast cancer from microbiome analysis. *Sci. Rep.* **2025**, *15*, 33096. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Yang, Y.; Mirzaei, G. Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification. *PLoS ONE* **2024**, *19*, e0293607. [\[CrossRef\]](#)
18. Mousavi, S.; Ng, O.; Saunders, J.H.; Acheson, A.G.; Parsons, S.L.; Safaralizadeh, R.; Hosseinpour-Feizi, M.; Azimzadeh-Isfajani, A.; Hashemzadeh, S. Study of Cofilin 1 Gene Expression in Colorectal Cancer. *J. Gastrointest. Oncol.* **2018**, *9*, 791–796. [\[CrossRef\]](#)
19. Wang, C.Y.; Lin, C.F. Annexin A2: Its molecular regulation and cellular expression in cancer development. *Dis. Markers* **2014**, *2014*, 308976. [\[CrossRef\]](#)
20. Wang, Y.; Liu, S.; Zhang, Y.; Yang, J. Myosin Heavy Chain 9: Oncogene or Tumor Suppressor Gene? *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **2019**, *25*, 888–892. [\[CrossRef\]](#)
21. Li, Q.; Yang, H.; Wang, P.; Liu, X.; Lv, K.; Ye, M. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *J. Transl. Med.* **2022**, *20*, 177. [\[CrossRef\]](#)
22. Ghuriani, V.; Wassan, J.T.; Tripathi, P.; Chauhan, A. XGB-BIF: An XGBoost-Driven Biomarker Identification Framework for Detecting Cancer Using Human Genomic Data. *Int. J. Mol. Sci.* **2025**, *26*, 5590. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.