The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Improving Breast Cancer Prognosis via Multi-Gene Machine Learning Model*

Fateme Azizi† Bahram Sadeghi Bigham‡ Mahboobe Zarrabi§

Abstract

Breast cancer remains one of the most prevalent and lethal diseases among women, with challenges in treatment stemming from the biological and genetic heterogeneity of tumors. While prior studies have developed grading models using machine learning to improve prognostic accuracy, they reached a peak accuracy of 91%. This paper advances this work by employing a more extensive dataset and refined data selection methods, achieving an accuracy improvement to 92%. Gene expression datasets were collected from the Gene Expression Omnibus (GEO) repository, undergoing preprocessing, integration, and normalization, before being analyzed by the XGBoost algorithm to develop a predictive tumor grading model. Evaluation results show that our expanded dataset and modified biomarker panel of 70 markers contribute to enhanced grading accuracy, particularly in classifying grade 2 and indeterminate tumors, which are often challenging to diagnose and treat. This model underscores the effectiveness of combining expanded transcriptomic data with advanced machine learning techniques. Furthermore, it highlights key genes associated with prognosis, offering insights into potential biomarkers for future research and clinical applications.

Keywords: Gene Expression, Machine Learning, Cancer Prognosis

1 Introduction

Breast cancer remains one of the most prevalent and deadly malignancies affecting women globally, with millions of new diagnoses each year [3]. Despite advancements in detection methods and therapeutic interventions, breast cancer continues to account for a significant proportion of cancer-related deaths. Accurate prognostication has become increasingly critical, as it directly influences patient outcomes and guides clinical decision-making. A key aspect of this challenge lies in the classification of breast tumors, a process central to understanding disease aggressiveness and determining the prognosis for patients. [10]

In recent years, substantial progress has been made in classifying breast cancer subtypes, particularly through the use of immunohistochemical markers. Key markers such as the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are now pivotal in identifying breast cancer subtypes [15]. This molecular classification approach has enhanced the precision of treatment regimens, tailoring interventions based on the specific characteristics of each subtype. While these advancements have improved patient outcomes, breast cancer remains a heterogeneous disease. Tumors progress through unique sets of genetic mutations and biological variations, complicating the prediction of disease progression and treatment responses based on traditional classification systems. This complexity underscores the need for individualized treatment strategies, taking into account the diversity in genetic alterations, gene expression profiles, and signaling pathways. [16]

A significant tool in advancing breast cancer diagnostics is the use of gene expression panels, such as the PAM50 classification system. The PAM50 categorizes tumors into four major subtypes: Luminal A, Luminal B, Basal-like, and HER2-enriched. The system is based on a 50-gene panel that provides deeper insights into tumor biology and their likely treatment responses [15]. By incorporating gene expression data, models like PAM50 have expanded our understanding of the biological variability in breast cancer, leading to more personalized treatment strategies.

As our knowledge of tumor biology expands, the demand for more accurate predictive models that align clinical decisions with molecular data continues to grow. Although efforts have been made to refine these models, challenges remain. The genetic diversity and varying histological features of breast tumors complicate predictions regarding disease progression and therapeutic responses. Despite progress in molecular classification and predictive modeling, the complexity of breast cancer remains a significant barrier to achieving highly accurate prognoses. [12]

To address these challenges, it is essential to develop models that classify breast cancer based on intrinsic tumor features, such as histological grade, which can

**This research has been facilitated by the Data Science Lab of the Faculty of Mathematical Sciences at Alzahra University.

†Department of Computer Science, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran. azizif799@gmail.com

‡Department of Computer Science, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran. b_sadeghi_b@iasbs.ac.ir

§Department of Biotechnology, Faculty of Biological Sciences, Alzahra University, Tehran, Iran. mzarrabi@alzahra.ac.ir

provide more precise predictions than traditional factors like tumor size and stage [11, 6]. By combining both clinical and molecular data, such models can support more personalized diagnostic and treatment strategies. Although considerable advances have been made in understanding breast cancer subtypes, aligning clinical features with molecular patterns remains a critical challenge. This disconnect highlights the need for models that better identify high-risk tumors, optimizing treatment regimens to avoid both undertreatment and overtreatment [14, 7].

Furthermore, the difficulty of correlating clinical features with the molecular profiles of breast tumors is a significant barrier to accurately predicting patient prognosis and selecting appropriate therapies [9, 17]. Despite progress in classification methods and expanding knowledge of tumor biology, there is still a considerable gap between understanding molecular patterns and applying them in real-world clinical practice. Predictive models based on molecular profiling could help bridge this gap by reducing the risk of underdiagnosing high-risk cases while also minimizing unnecessary treatments in low-risk patients. Achieving this balance could reduce reliance on aggressive systemic therapies, ultimately improving treatment outcomes while minimizing adverse effects [2, 4].

In conclusion, developing more accurate and reliable models for breast cancer grading and prognosis is crucial for improving patient outcomes. By integrating transcriptomic profiles, employing advanced machine learning techniques, and incorporating both genetic and histological data, this study aims to enhance diagnostic accuracy and facilitate personalized treatment strategies. Identifying key predictive biomarkers could open new therapeutic pathways, especially for cases that have historically posed significant treatment challenges. As bioinformatics and molecular biology continue to advance, these innovations have the potential to redefine breast cancer treatment and prognostication [5].

Histological grading adds another layer of insight by evaluating how much tumor cells differ from normal cells and how invasive the cancer has become. This grading process is a powerful prognostic tool for assessing the aggressiveness of breast cancer [12]. Grade 2 tumors, in particular, present a unique challenge. These tumors are considered intermediate in terms of cellular morphology, biological characteristics, and invasive potential. As such, treatment decisions for patients with grade 2 tumors are often uncertain, as they fall into a diagnostic gray area. Studies suggest that between 30% and 60% of grade 2 tumors fall within this uncertain range, complicating the decision between aggressive or conservative treatment strategies [8]. This ambiguity has resulted in both undertreatment and overtreatment, highlighting the need for more accurate prognostic tools [13].

Clinicians traditionally have more success in determining treatment strategies for grade 1 or grade 3 tumors, while grade 2 tumors remain challenging due to their intermediate nature. In response, researchers have begun utilizing advanced machine learning and genetic data analysis techniques to refine diagnostic accuracy and improve treatment predictions.

The main objective of this study is to address these diagnostic and therapeutic challenges by developing a robust breast cancer grading model. This model integrates transcriptomic data from multiple datasets, focusing on genetic and histological features, and employs the XGBoost algorithm for analysis. The model aims to improve the classification and prognosis of grade 2 and indeterminate tumors, enhancing predictive accuracy and facilitating personalized treatment strategies. Furthermore, the model seeks to identify key genes that can reliably predict relapse-free survival, overall survival, and distant metastasis-free survival, irrespective of receptor status, which is typically assessed through conventional histopathological methods. Identifying such biomarkers could open new therapeutic targets, especially for cases that currently lack effective treatment options. In summary, although existing models for breast cancer prognosis have made considerable progress, challenges persist, particularly in achieving high predictive accuracy for intermediate-grade tumors [1]. This paper explores a machine learning-based approach that integrates both genetic and histological data, demonstrating improved predictive accuracy and utilizing a more comprehensive dataset compared to previous models. These findings underscore the potential of combining transcriptomic data with advanced machine learning techniques to enhance breast cancer prognosis.

2 Materials and methods

2.1 Computational framework

The computational framework for this study is illustrated in Figure 2.1, which presents a comprehensive overview of the methodology. This workflow encompasses key stages, starting with data preprocessing and integration, followed by the construction of the machine learning model and cross-validation procedures. Additionally, the framework highlights essential stages of model interpretation, including feature prioritization and prognostic data analysis. This structured approach ensures a robust and reproducible process, leveraging state-of-the-art machine learning techniques for enhanced breast cancer grading and prognosis prediction.

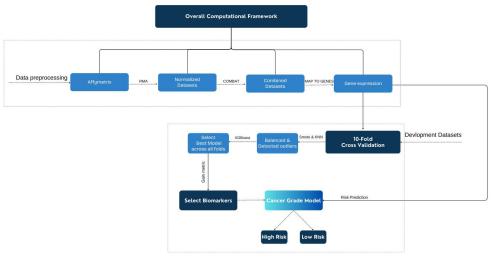

Figure 1: Overview of the framework model

2.2 Dataset, pre-processing and integration

Gene expression data from 33 publicly available breast cancer datasets corresponding to the GPL570 (Genome U133 Plus 2.0) and GPL96 (Genome U133A) platforms were obtained from the Gene Expression Omnibus (GEO) database. A total of 5,334 breast tumor samples and 70 normal breast tissue samples were collected, focusing on distinguishing gene expression profiles between malignant and non-malignant tissues. Samples from patients with prior treatments were excluded to mitigate potential biases introduced by treatment effects.

Data collection followed a rigorous approach, beginning with a focus on datasets specifically related to breast cancer. The inclusion criteria were restricted to datasets examining normal and malignant breast tissues, and only datasets corresponding to the Affymetrix GPL96 or GPL570 platforms were considered. Furthermore, only datasets containing human tissue samples were selected, which yielded an initial pool of 16,044 datasets.

A date filter was applied to include datasets collected between 2006 and 2021, ensuring that only recent and relevant data were used for the study. Subsequently, datasets were sorted by the number of samples, with those containing the highest number of samples prioritized. To ensure statistical robustness, a threshold of at least 25 samples per dataset was applied, narrowing the selection to 243 datasets. Of these, only datasets with available clinical information—such as histological grade, overall survival, distant metastasis-free survival, or relapse-free survival were retained. Finally, samples from patients who had received prior treatment were excluded to minimize confounding variables in gene expression analysis Table 1.

Given the large and diverse sample size, which includes tumor samples from various grades and gene expression profiles, the model has demonstrated strong generalizability and robustness. The dataset's comprehensive inclusion of different tumor grades and clinical features enables the model to generalize effectively to new, unseen data, making it suitable for application across various clinical settings and populations.

Data pre-processing involved the normalization of raw intensity values from Affymetrix CEL files using the Robust Multichip Average (RMA) method. Batch effects, which can arise from integrating datasets collected under different experimental conditions, were corrected using the COMBAT algorithm. These steps were implemented in R version 4.4 with the appropriate bioinformatics libraries. Following normalization and batch correction, probe sets were mapped to their corresponding genes. In instances where multiple probes corresponded to the same gene, the mean expression value of the probes was computed to ensure consistent and accurate gene expression representation.

The data integration and processing workflow are outlined in Figure 1. The resulting integrated dataset was used for training and testing the machine learning model. Initially, the model was trained using grade 1 and grade 3 tumor samples (1,891 samples), representing distinct low- and high-risk categories. Subsequently, the model was applied to grade 2 and unknown-grade tumor samples (3,443 samples), with the aim of stratifying them into low-risk and high-risk categories based on their gene expression profiles and clinical characteristics.

2.3 Machine learning model development

The machine learning model was designed as a binary classifier, using gene expression levels as input features and cancer grades (grade 1 vs. grade 3) as output labels. The development process followed a comprehensive machine learning pipeline to minimize overfitting and improve model generalization. The dataset was split into training 80%, validation 10%, and testing 10% sets. Hyperparameter tuning was performed on the validation set, and model performance was assessed on the testing set.

To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, generating synthetic samples for the minority class (grade 1) to ensure balanced training. Outlier detection was conducted using the K-nearest neighbor algorithm, implemented with the PyOD library, and samples flagged as outliers were excluded from training to improve model robustness.

The core model was constructed using XGBoost (eXtreme Gradient Boosting), an ensemble learning method known for optimizing classification tasks through the combination of weak decision trees. During the training phase, a grid search was employed to optimize hyperparameters, including maximum tree depth, subsample ratio, and gamma. The optimized parameters for the

Cancer Grade Model. were a maximum tree depth of 5, subsample ratio of 0.6, minimum child weight of 1, and gamma of 0.5. Cross-validation was utilized to select the optimal model configuration.

Given the large number of genes in the dataset, using all features directly in the model would increase complexity and potentially reduce accuracy. To address this, Principal Component Analysis (PCA) was employed for dimensionality reduction. PCA allowed us to reduce the high-dimensional gene expression data to a smaller set of principal components while retaining most of the information. This reduction in dimensionality led to several benefits: it decreased the model's complexity, reduced training time, and minimized the risk of overfitting by eliminating noise and irrelevant features.

Feature importance was further assessed using the Gain metric, which highlights the genes contributing most to model performance. Additionally, SHAP (SHapley Additive exPlanations) values were calculated to provide an interpretable understanding of each feature's impact on model predictions, ensuring transparency and interpretability in the decision-making process.

3 Results

3.1 Model Performance and Validation

The dataset was divided into two subsets: a development dataset consisting of grade 1 and grade 3 tumor samples (total of 1,891), which was used to train and test the machine learning model, and a prediction dataset composed of grade 2 and unknown-grade tumors (total of 3,443). The model was applied to the prediction dataset to classify these intermediate-grade and ambiguous tumors into low-risk (1,330) and high-risk (2,113) groups based on their gene expression profiles.

Model validation was conducted using 10-fold cross-validation, with performance metrics such as accuracy, AUC, F1-score, precision, recall, and ROC curve reported (Table1). The results demonstrate the high efficacy of the model in differentiating between tumor grades, while hyperparameter tuning effectively minimized the risk of overfitting. Furthermore, the implementation of SMOTE addressed the class imbalance between grade 1 and grade 3 tumors, ensuring balanced model training. Outlier detection using the K-nearest neighbor algorithm further enhanced the dataset by removing anomalous samples, thereby improving the model's overall stability and predictive power.

Given the large and diverse sample size, the model has demonstrated strong generalizability and robustness. The dataset's comprehensive inclusion of different tumor grades and gene expression profiles allows the model to effectively generalize to new, unseen data.

Table 1: Model Performance and Validation

Performance metric	All genes	Selected genes
Accuracy	0.92	0.92
AUC	0.95	0.92
$F_1 score$	0.90	0.91
Precision	0.90	0.91
Recall	0.89	0.85

This ensures that the model is capable of accurately predicting tumor risk in a variety of clinical settings and populations, extending its applicability beyond the training data. The model was also tested on additional external datasets, where it maintained high predictive performance, further supporting its robust generalizability.

3.2 Feature Selection and Interpretation

The 70-gene panel selected for this study was specifically chosen for its capacity to reflect key tumor clinical characteristics, including histological grade,distant metastasis-free survival, relapse-free survival and overall survival. These genes represent critical biological differences between high-risk and low-risk tumors, enabling more precise classification of heterogeneous tumors, particularly those in grade 2. For example, BIRC5, which is highly expressed in high-risk tumors, is associated with cell proliferation and inhibition of apoptosis, whereas LINC00472, which is more active in low-risk tumors, is linked to tumor-suppressing pathways, such as the p53 pathway.

The selection of these genes was based on the Gain metric derived from the XGBoost algorithm, which evaluates the relative importance of each gene in the classification model. Up to the inclusion of 70 genes, the Gain values allowed for meaningful differentiation between the genes, ensuring robust model performance. However, beyond the 70-gene threshold, the Gain values for additional genes exhibited minimal differences, typically only at the decimal level. These marginal differences indicated that the inclusion of additional genes would not significantly impact the model's predictive performance, thereby justifying the exclusion of further genes. Consequently, the 70-gene panel was determined to be the optimal balance between predictive performance and interpretability.

Key features that significantly contributed to the model's predictive power were identified through the Gain metric, which highlighted the most influential genes in the classification task. These genes are inte-

gral to fundamental cellular processes, such as cancer cell proliferation and mitosis, reinforcing their prognostic relevance in breast cancer. To visualize the effects of feature selection, predictions for both the full gene set and the 70-gene subset are presented in Figure 2 and Figure 3. These results demonstrate that the model, when trained on the full set of genes or the reduced 70-gene panel, shows similar performance in terms of accuracy. This confirms that the 70 genes retain the essential information required for tumor classification while reducing dimensionality.

This approach not only reduces computational complexity and enhances efficiency but also mitigates the risk of overfitting, thereby ensuring the model's ability to generalize effectively to new datasets. Additionally, the manageable size of the 70-gene panel makes it clinically viable, reducing both the costs and logistical challenges typically associated with large-scale genetic testing. Further validation, including survival analysis and comparisons with established clinical methods such as OncotypeDX and EndoPredict, has demonstrated the competitive predictive power of this gene panel, supporting its selection as both scientifically and clinically sound.

To enhance model interpretability, SHAP (SHapley Additive exPlanations) values were calculated, providing insights into the specific influence of each gene on the model's predictions Figure 4. This interpretability analysis revealed that genes such as BIRC5 and CDC20 played a prominent role in distinguishing high-risk tumor samples, thereby making the model's decision-making process more transparent and clinically actionable. This transparency is crucial for ensuring the model's applicability in clinical settings where interpretability is essential for guiding therapeutic decisions.

3.3 Clinical Relevance and Prognostic Accuracy

The model demonstrated its potential to identify high-risk and low-risk groups among grade 2 and unknown-grade tumors, contributing to more accurate prognostic assessment. The identified key genes, particularly BIRC5 and CDC20, are strongly associated with cancer progression and could serve as biomarkers for guiding treatment decisions. These findings highlight the model's potential clinical utility, providing an additional tool for stratifying patients based on the risk of recurrence and aggressiveness. The detailed evaluation of performance metrics, including accuracy, AUC, and F1-score, is summarized in Table1. This comprehensive assessment demonstrates the model's effectiveness in improving risk classification, further discussed in the subsequent sections.

3.4 Application to Grade 2 Tumors

One of the major challenges in breast cancer diagnostics is the precise classification of grade 2 tumors, owing to their intermediate characteristics and heterogeneous behavior. The model was applied to a total of 3,443 grade 2 and unknown-grade tumors from the prediction dataset, categorizing them into high-risk and low-risk groups. This stratification enables the development of more tailored and individualized treatment strategies by providing a clearer prognostic profile for tumors that traditionally present significant diagnostic uncertainty.

The reclassification results obtained from the model were benchmarked against established genomic tests, including OncotypeDX, EndoPredict, and GGI. These genomic assays, which evaluate gene expression levels linked to tumor aggressiveness, are widely recommended in clinical guidelines for determining adjuvant systemic therapies. A Table 2 illustrates the overlap between the biomarkers identified by our model and those reported by the aforementioned tests.

3.5 Genomic Tests for Clinical Assessment of Breast Cancer

The clinical assessment of breast cancer, particularly for determining the aggressiveness and potential recurrence of tumors, often involves the use of established genomic tests. These tests, such as OncotypeDX, EndoPredict, and GGI, evaluate the expression levels of genes associated with tumor biology and are widely integrated into clinical guidelines to guide decisions regarding adjuvant systemic therapy.

To benchmark the performance of the our model, the reclassification of grade 2 and unknown-grade tumors was directly compared with the results from these genomic tests. The model classified 3,443 tumors into high-risk and low-risk categories based on gene expression profiles. A comparative analysis revealed notable overlap between the biomarkers identified by model and those used by OncotypeDX, EndoPredict, and GGI, highlighting shared genetic indicators of cancer aggressiveness.

The Table 2 illustrates the common and unique biomarkers identified across these models, emphasizing the model's alignment with clinically accepted standards while also introducing novel gene markers (full gene list available in Supplementary Table 3). This comparison underscores the potential of the model to serve as a complementary or alternative tool for risk stratification in clinical practice, especially for grade 2 tumors, where the current genomic tests may yield ambiguous results.

4 Discussion

Our study demonstrates that the breast cancer grade model, developed using full gene expression profiles and subsequently optimized with 70 selected key genes, provides highly accurate predictions for breast cancer prognosis. Specifically, the model was effective in classifying grade 2 and unknown-grade tumors into distinct high-risk and low-risk categories, which is a critical advancement in the classification of these ambiguous cases. The identification of key genes, such as BIRC5, CDC20, and CENPN, was particularly important, as these biomarkers significantly contribute to tumor aggressiveness and patient prognosis. The model's performance was comparable to widely accepted genomic tools, such as OncotypeDX and EndoPredict, which are currently used in clinical practice to inform treatment decisions. However, the mpdel's unique strength lies in its ability to handle grade 2 tumors—traditionally challenging cases—more effectively than these existing tests. This ability is particularly valuable because accurate classification of grade 2 tumors is crucial in preventing under- or overtreatment, both of which can lead to poor patient outcomes. One of the most significant findings of this study is the identification of BIRC5, CDC20, and CENPN as influential genes in breast cancer progression. These genes are involved in crucial cellular processes such as cell cycle regulation and mitosis, making them not only predictive of tumor behavior but also potential therapeutic targets. By focusing on these biomarkers, the model model could guide the development of personalized treatment strategies aimed at targeting aggressive tumor phenotypes. The ability to accurately classify tumors based on these molecular signatures may also reduce the use of unnecessary systemic therapies, thereby minimizing side effects for low-risk patients while ensuring high-risk patients receive appropriate treatment. While the model exhibited strong predictive power, several limitations must be acknowledged. First, the reliance on gene expression data from specific platforms (GPL570 and GPL96) could restrict the generalizability of the model to datasets from other platforms or newer genomic technologies. Additionally, although the use of SMOTE effectively handled the class imbalance between grade 1 and grade 3 tumors, the potential for model bias remains, particularly in cases where grade 1 tumors were underrepresented in the training dataset. In future research, efforts should be directed toward validating the model on more diverse datasets that include data from a broader range of platforms and technologies. Expanding the model to include multi-omics data, such as proteomics or histopathological imaging, could also enhance its predictive capacity and clinical relevance. Moreover, further investigation into the functional role of biomarkers like BIRC5 and CDC2 may uncover novel therapeutic strategies, especially for patients with high-risk grade 2 tumors. The model represents a significant advancement in breast cancer prognosis, particularly in the classification of grade 2 and unknown-grade tumors. By integrating gene expression data with machine learning techniques, the model provides an accurate risk stratification that can be directly applied to clinical decision-making. The identification of key biomarkers offers new avenues for therapeutic intervention, while the model's capacity to address challenges in grade 2 classification highlights its potential as an alternative or complementary tool in current clinical practice. Further refinement and validation of the model are necessary to realize its full potential in improving patient outcomes and advancing personalized medicine.

- Full-size view of the figure1
- Full-size view of the Figure 2
- Full-size view of the Figure 3
- Full-size view of the Figure 4
- Full-size view of the Table 1
- Full-size view of the Table 2
- Full-size view of the Table 3

5 Data availability

The datasets supporting the findings of this study were obtained from the Gene Expression Omnibus (GEO) repository [`http://www.ncbi.nlm.nih.gov/geo`].

6 Code availability

The code will be made available at [`https://github.com/users/Fatemeazizi11/projects/1`]

References

[1] E. Amiri Souri, A. Chenoweth, A. Cheung, S. N. Karagiannis, S. Tsoka Cancer Grade Model: a multi-gene machine learning-based risk classification for improving prognosis in breast cancer. *British Journal of Cancer*, 125(5):748–758, 2021.

[2] F. Bertucci, P. Finetti, J. Ostrowski, W. K. Kim, H. Kim, M. A. Pantaleo et al. Genomic Grade Index predicts postoperative clinical outcome of GIST. *British Journal of Cancer*, 107:1433–1441, 2012.

[3] F. Cardoso, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans, I. T. Rubio et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 30:1194–1220, 2019.

[4] C. Christos, P. Lajos Gene-expression signatures in breast cancer. *New England Journal of Medicine*, 11:790–800, 2009. https://doi.org/10.1056/NEJMra0801289

[5] D. Demetriou, Z. Lockhat, L. Brzozowski, K. S. Saini, Z. Dlamini, R. Hull The Convergence of Radiology and Genomics: Advancing Breast Cancer Diagnosis with Radiogenomics. *Cancers*, 16(5):1076, 2024.

[6] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*, 121:2750–2767, 2011.

[7] A. Goldhirsch, J. N. Ingle, R. D. Gelber, A. S. Coates, B. Thürlimann, H.-J. Senn Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2009. *Annals of Oncology*, 20:1319–1329, 2009.

[8] S. G. Komen Contents of a Breast Cancer Pathology Report. *Susan G. Komen®*, 2023. https://www.komen.org

[9] M. T. Phung, S. Tin Tin, J. M. Elwood Prognostic models for breast cancer: a systematic review. *BMC Cancer*, 19:230, 2019.

[10] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans et al. Breast cancer. *Nature Reviews Disease Primers*, 5:66, 2019.

[11] Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27:1160–1167, 2009.

[12] A. Romaniuk-Drapała, E. Totoń, M. Taube, M. Idzik, B. Rubiś, N. Lisiak Breast Cancer Stem Cells and Tumor Heterogeneity: Characteristics and Therapeutic Strategies. *Cancers*, 16(13):2481, 2024.

[13] S. W. Jahn, A. Bösl, O. Tsybrovskyy, C. Gruber-Rossipal, R. Helfgott, F. Fitzal et al. Clinically high-risk breast cancer displays markedly discordant molecular risk predictions between the MammaPrint and EndoPredict tests. *British Journal of Cancer*, 122:1744–1746, 2020.

[14] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, 12:207, 2010.

[15] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the USA*, 98:10869–10874, 2001.

[16] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees et al. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

[17] The Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.