Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(3), 597-603
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

# Machine Learning-Based Selection of Key miRNA Biomarkers for Breast Cancer Diagnostics

**Abderrahim Chafik**

Independent Researcher, Rabat, Morocco
Email: ab.chafiki@gmail.com

**How to cite this paper:** Chafik, A. (2025) Machine Learning-Based Selection of Key miRNA Biomarkers for Breast Cancer Diagnostics. *Open Journal of Applied Sciences*, **15**, 597-603.
https://doi.org/10.4236/ojapps.2025.153038

**Received:** February 13, 2025
**Accepted:** March 14, 2025
**Published:** March 17, 2025



## Abstract

MicroRNAs (miRNAs) play a pivotal role in gene expression regulation and are closely linked to cancer development. In this study, we employ machine learning techniques to identify critical miRNA biomarkers for breast cancer diagnostics using a dataset of 941 patient samples with 1,882 miRNA features. By addressing class imbalance and applying robust feature selection, we developed an optimized Random Forest model that achieved a perfect classification accuracy of 1.0. Analyzing feature importance revealed 51 miRNAs as potential biomarkers, offering a valuable panel for precision diagnostics and personalized treatment strategies.

## Keywords

RNA, Breast Cancer, Machine Learning, Feature Selection, Random Forest

## 1. Introduction

MicroRNAs (miRNAs) are small, non-coding RNA molecules, typically 18 - 25 nucleotides long, that play crucial roles in regulating gene expression at the post-transcriptional level. By binding to complementary sequences in messenger RNA (mRNA), miRNAs either degrade mRNA or inhibit its translation, effectively suppressing gene expression [1]-[4]. They are essential in many biological processes, including cell proliferation, differentiation, apoptosis, and stress responses. In cancer, miRNAs can function as either oncogenes or tumor suppressors, depending on the genes they target. Aberrant miRNA expression can disrupt normal cellular regulation, contributing to cancer initiation and progression.

Breast cancer is among the most common cancers affecting women worldwide, accounting for a significant share of cancer diagnoses and mortality. Advances in

early detection such as mammography and in treatments, including surgery, chemotherapy, radiation, and targeted therapies, have greatly improved outcomes. However, breast cancer remains a complex disease with diverse molecular subtypes, each requiring distinct prognostic and therapeutic strategies, emphasizing the importance of continued research for personalized treatments that improve survival rates and quality of life. The relationship between miRNA and cancers has been already established [5] [6]. Certain miRNAs, such as miR-21 [7], miR-155 [8], and the let-7 family [9], are often dysregulated in breast cancer, influencing tumor growth, metastasis, and treatment response. For example, miR-21 is frequently overexpressed in breast cancer and is associated with a poor prognosis due to its role in promoting cell proliferation and inhibiting apoptosis. Studying miRNAs in breast cancer provides opportunities to identify new therapeutic targets and biomarkers for early detection, prognosis, and monitoring treatment responses.

This study aims to harness machine learning to identify the most effective miRNA biomarkers for breast cancer detection. By using feature selection and random forest algorithms, we sought to identify a panel of miRNAs that could accurately distinguish breast cancer from normal tissues. Combining machine learning with miRNA profiling has the potential to enhance breast cancer diagnosis, prognosis, and treatment selection, contributing to the field of precision medicine by providing more accurate and personalized approaches to breast cancer management.

## 2. Related Work

Advances in computational methods, particularly machine learning (ML), have revolutionized the field of biomarker discovery. ML algorithms are adept at handling large, high-dimensional datasets and identifying subtle patterns in miRNA expression profiles that differentiate cancerous from normal tissues. Techniques such as support vector machines, neural networks, and ensemble methods like random forests have been successfully applied to classify cancer types, predict patient outcomes, and identify potential therapeutic targets [10]. For instance, Rehman *et al.* validated miRNAs as breast cancer biomarkers using ML approaches, achieving significant accuracy in distinguishing tumor subtypes [11]. Similarly, Contreras-Rodrguez *et al.* systematically reviewed ML methods for miRNA classification, highlighting their potential to enhance diagnostic precision in breast cancer [12].

Feature selection methods are integral to these ML workflows, allowing researchers to reduce data dimensionality and focus on the most relevant features. Techniques such as chi-squared tests and mutual information have been employed to identify key miRNAs from expression datasets, improving model interpretability and performance [10]. Random forest models, in particular, have demonstrated robustness in processing high-dimensional data, offering insights into feature importance while mitigating overfitting [13]. These models have been

widely used in cancer diagnostics due to their ability to handle class imbalances and diverse data distributions effectively [14] [15].

In breast cancer research, integrating miRNA profiling with ML has shown great promise in developing non-invasive diagnostic tools with high sensitivity and specificity. By leveraging computational techniques, researchers have identified miRNA signatures associated with specific molecular subtypes, aiding in early detection and personalized treatment planning. For example, miRNA panels identified through ML have shown potential in distinguishing triple-negative breast cancer from other subtypes, facilitating targeted therapeutic strategies [11] [12].

## 3. Method

### 3.1. Dataset and Preprocessing

The study utilized a miRNA expression dataset from the National Cancer Institutes Genomic Data Commons (GDC) portal. The dataset included 941 samples 873 tumors, 68 healthy, and 1 metastatic[1], comprising 1882 miRNA features. Initial preprocessing involved removing low-expression miRNAs, resulting in 1603 retained features.

To address class imbalance, we employed the Synthetic Minority Oversampling Technique (SMOTE) [14] [15], balancing the dataset to 873 samples per class. This approach mitigated bias, enhancing the classifier's ability to identify patterns in minority-class samples.

### 3.2. Feature Selection

We applied the Chi-squared (CHI) test to identify miRNAs most relevant to breast cancer classification. Features with a p-value below 0.05 were retained, further reducing the dataset to 1560 features.

### 3.3. Model Development

A Random Forest classifier was chosen for its robustness in high-dimensional spaces [13]. The models hyperparameters were optimized using grid search to achieve maximum classification accuracy. This ensemble method mitigates overfitting by aggregating predictions from multiple decision trees.

## 4. Results

The optimized Random Forest model achieved perfect accuracy (1.0) (**Figure 1**), demonstrating its capability to distinguish between cancerous and healthy samples.

Since different thresholds give different number of miRNAs, we choose 0.0035 to be the threshold to the feature importance (**Figure 2**) so as to get 51 miRNAs as the most discriminative biomarkers for breast cancer since this number is the minimum one that includes clinically validated biomarkers like miR-21, miR-145,

[1]In the following we include this single metastatic sample in tumor samples since we are interested in considering just healthy and tumor cases.

and let-7c.

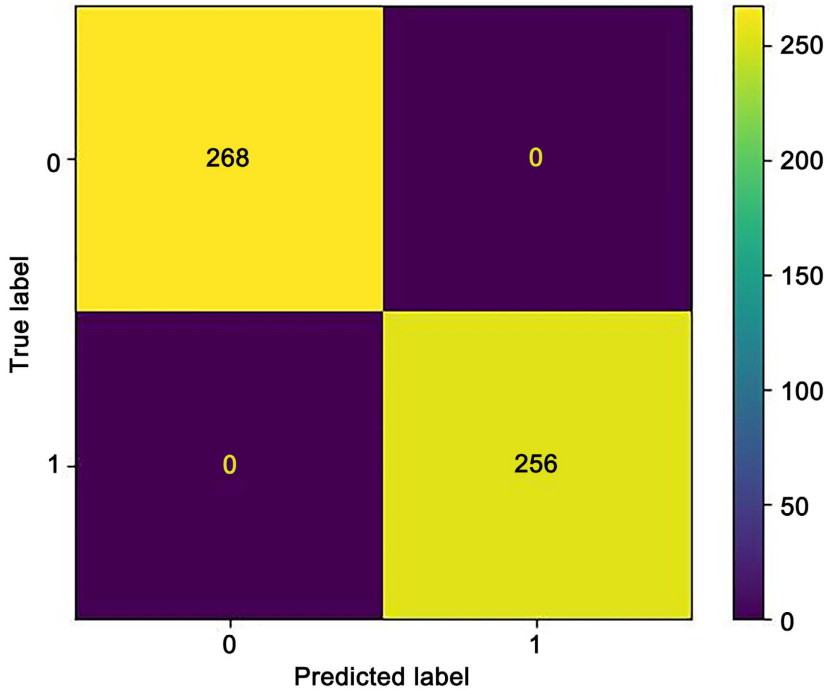


Figure 1. The confusion matrix of the classifier model.

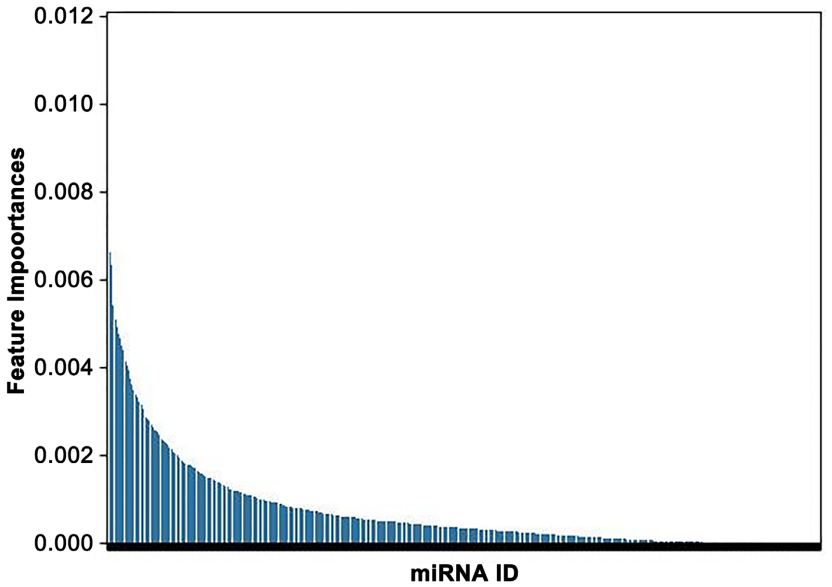


Figure 2. The feature importances deduced from the random forest model.

The identified panel includes several clinically validated miRNAs, such as miR-21, miR-145, and let-7c, [11]. reaffirming their significance in breast cancer diagnostics. While computational analysis is valuable, experimental validation is crucial for confirming the biological and clinical relevance of the identified biomarkers, so the other miRNAs need experimental validation in future studies to

confirm the biological relevance of these biomarkers. These findings highlight the potential of integrating computational methods with molecular biology to advance precision medicine.

hsa-mir-145, hsa-mir-196a-1, hsa-mir-483, hsa-mir-1258, hsa-mir-140, hsa-mir-3677, hsa-mir-100 hsa-mir-139, hsa-mir-518c, hsa-mir-493, hsa-mir-4678, hsa-mir-329-2, hsa-mir-4802, hsa-let-7c, hsa-mir-218-2, hsa-mir-335, hsa-mir-183, hsa-mir-584, hsa-mir-486-1, hsa-mir-337, hsa-mir-148a, hsa-mir-1468, hsa-mir-4788, hsa-mir-21, hsa-mir-192, hsa-mir-6720, hsa-mir-224, hsa-mir-556, hsa-mir-3065, hsa-mir-2355, hsa-mir-193a, hsa-mir-365a, hsa-mir-375, hsa-mir-934, hsa-mir-378d-1, hsa-mir-33b, hsa-mir-758, hsa-mir-381, hsa-mir-196a-2, hsa-mir-665, hsa-mir-6715a, hsa-mir-3199-1, hsa-mir-497, hsa-mir-215, hsa-mir-4638, hsa-mir-200a, hsa-mir-592, hsa-mir-8072, hsa-mir-99a, hsa-mir-7706, hsa-mir-548o-2.

## 5. Discussion

This study demonstrates the potential of combining computational methods with molecular biology to improve breast cancer diagnostics. Utilizing a miRNA expression dataset, we successfully identified a set of 51 miRNAs that hold promise as reliable biomarkers for breast cancer. The rigorous preprocessing pipeline, including the removal of low-expression miRNAs and the application of SMOTE for class balancing, ensured a robust dataset for analysis. Chi-squared feature selection further refined the dataset, isolating the most statistically significant miRNAs.

By optimizing a Random Forest classifier, we achieved a model with perfect accuracy, underscoring the capability of machine learning to discern subtle patterns in high-dimensional data. The analysis of feature importance not only validated the classifier's effectiveness but also identified specific miRNAs that could serve as diagnostic markers.

These results highlight the transformative role of miRNAs in precision medicine, particularly for complex diseases like breast cancer. The 51 identified miRNAs provide a foundation for developing non-invasive diagnostic tools with high sensitivity and specificity. Future work should focus on validating these biomarkers in larger and independent cohorts and exploring their biological roles in cancer progression. Such advancements could significantly enhance early detection and personalized treatment strategies, improving outcomes for breast cancer patients worldwide.

## 6. Conclusion

In this study, we employed machine learning techniques to identify key miRNA biomarkers for breast cancer diagnostics, achieving a high level of accuracy using a Random Forest classifier. Our analysis identified 51 miRNAs with significant discriminatory power, some of which have known roles in cancer progression. While these findings demonstrate the potential of computational approaches in

biomarker discovery, further validation in independent datasets and experimental studies is essential to confirm their clinical relevance. Future research should focus on refining feature selection methods, integrating multi-omics data, and exploring the biological functions of these miRNAs to enhance personalized medicine in breast cancer diagnostics and treatment.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

[1] Bartel, D.P. (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, **116**, 281-297. https://doi.org/10.1016/s0092-8674(04)00045-5

[2] He, L. and Hannon, G.J. (2004) MicroRNAs: Small RNAs with a Big Role in Gene Regulation. *Nature Reviews Genetics*, **5**, 522-531. https://doi.org/10.1038/nrg1379

[3] Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P. and Bartel, D.P. (2008) The Impact of MicroRNAs on Protein Output. *Nature*, **455**, 64-71. https://doi.org/10.1038/nature07242

[4] Fabian, M.R., Sonenberg, N. and Filipowicz, W. (2010) Regulation of mRNA Translation and Stability by MicroRNAs. *Annual Review of Biochemistry*, **79**, 351-379. https://doi.org/10.1146/annurev-biochem-060308-103103

[5] Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., *et al.* (2002) Frequent Deletions and Down-Regulation of MicroRNA Genes *mir*15 and *mir*16 at 13q14 in Chronic Lymphocytic Leukemia. *Proceedings of the National Academy of Sciences*, **99**, 15524-15529. https://doi.org/10.1073/pnas.242606799

[6] McManus, M.T. (2003) MicroRNAs and Cancer. *Seminars in Cancer Biology*, **13**, 253-258. https://doi.org/10.1016/s1044-579x(03)00038-5

[7] Wang, H., Tan, Z., Hu, H., Liu, H., Wu, T., Zheng, C., *et al.* (2019) MicroRNA-21 Promotes Breast Cancer Proliferation and Metastasis by Targeting LZTFL1. *BMC Cancer*, **19**, Article No. 738. https://doi.org/10.1186/s12885-019-5951-3

[8] Wang, J., Wang, Q., Guan, Y., Sun, Y., Wang, X., Lively, K., *et al.* (2022) Breast Cancer Cell-Derived MicroRNA-155 Suppresses Tumor Progression via Enhancing Immune Cell Recruitment and Antitumor Function. *Journal of Clinical Investigation*, **132**, e157248. https://doi.org/10.1172/jci157248

[9] Thammaiah, C.K. and Jayaram, S. (2016) Role of Let-7 Family MicroRNA in Breast Cancer. *Non-Coding RNA Research*, **1**, 77-82. https://doi.org/10.1016/j.ncrna.2016.10.003

[10] Malik, Y. and Jens, A. (2014) MiRNomics: MicroRNA Biology and Computational Analysis. Springer.

[11] Rehman, O., Zhuang, H., Muhamed Ali, A., Ibrahim, A. and Li, Z. (2019) Validation of MiRNAs as Breast Cancer Biomarkers with a Machine Learning Approach. *Cancers*, **11**, Article No. 431. https://doi.org/10.3390/cancers11030431

[12] Contreras-Rodríguez, J.A., Córdova-Esparza, D.M., Saavedra-Leos, M.Z. and Silva-Cázares, M.B. (2023) Machine Learning and MiRNAs as Potential Biomarkers of Breast Cancer: A Systematic Review of Classification Methods. *Applied Sciences*, **13**, Article No. 8257. https://doi.org/10.3390/app13148257

[13] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

https://doi.org/10.1023/a:1010933404324

[14] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[15] Haibo He, and Garcia, E.A. (2009) Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263-1284. https://doi.org/10.1109/tkde.2008.239