УДК 618.19-006.6:577.217
https://doi.org/10.20538/1682-0363-2026-2-91-100

The Potential Application of the Random Forest Method to Explore the Association between MicroRNAs and Tissue and Molecular Genetic Biomarkers of Breast Cancer

Studenikina A.A.[1, 2], Mihailovsky M.A.[3], Timofeev V.S.[3], Arkhipov S.A.[1, 2], Autenshlyus A.I.[1, 2]

[1] *Novosibirsk State Medical University*
52 Krasny Ave., 630091 Novosibirsk, Russian Federation

[2] *Institute of Molecular Biology and Biophysics, Federal Research Center of Fundamental and Translational Medicine*
2 Timakova St., 630060 Novosibirsk, Russian Federation

[3] *Novosibirsk State Technical University*
20 Karl Marx Ave., 630073 Novosibirsk, Russian Federation

ABSTRACT

Aim. Exploring the potential of using a random forest method (RFM) to identify associations between serum levels of microRNAs miR-181a and miR-25 and the expression of the following markers: E-cadherin (CDH1), type II collagen (CII), integrin beta-1 (ITGB1), cyclin D1 (CCND1), and T-lymphocyte costimulator (CD86), as well as their relationship with clinically relevant molecular genetic markers (ER, PR, HER2, Ki-67) in patients with breast cancer (BC).

Materials and methods. We recruited 44 BC patients with invasive ductal carcinoma who had not received neoadjuvant treatment. Serum levels of miR-181a and miR-25 were measured using digital droplet PCR, and the expression of CDH1, CII, ITGB1, CCND1, CD86, ER, PR, HER2, and Ki-67 in tumor tissue samples was assessed via immunohistochemistry. Spearman's correlation analysis and the RFM were applied to assess associations between variables.

Results. Spearman's correlation analysis did not reveal any significant linear relationships between the studied parameters ($p > 0.05$). However, the RFM identified nonlinear associations: for miR-181a, the optimal predictive model included CDH1 and CCND1 (MAE = 4.267); for miR-25, ITGB1 and CCND1 (MAE = 16.255). Among clinical and pathological characteristics, statistically significant positive correlations were found between CII and ER ($r = 0.498$; $p = 0.001$) and PR ($r = 0.354$; $p = 0.018$). The RFM models demonstrated the highest accuracy in predicting the HER2 status (82.6%) when using a combination of ITGB1, miR-181a, and CDH1.

Conclusion. The RFM revealed complex nonlinear associations between microRNAs and tissue markers of tumor progression that were not detected by conventional statistical methods. These findings are exploratory in nature and provide a foundation for further research aimed at validating these biomarkers and elucidating their role in the pathogenesis of various breast cancer subtypes.

Keywords: breast cancer, HER2, microRNA, machine learning, random forest, molecular markers, epithelial – mesenchymal transition

Conflict of interest. The authors declare the absence of obvious or potential conflict of interest related to the publication of this article.

Source of financing. The study was conducted within the state assignment of Novosibirsk State Medical University, Ministry of Health of the Russian Federation (Registration No. 115060410035), and the research project of the Federal Research Center for Fundamental and Translational Medicine, Research Institute of Molecular Biology and Biophysics (Registration No. 125031203556-7).

✉ *Studenikina Anastasiia A.,* aastudenikina@frcftm.ru

Conformity with the principles of ethics. All patients signed an informed voluntary consent to participate in the study. The study was approved by the Ethics Committee at the Research Institute of Molecular Biology and Biophysics of the Federal Research Center for Fundamental and Translational Medicine (Minutes No. 28 dated September 27, 2023).

For citation: Studenikina A.A., Mihailovsky M.A., Timofeev V.S., Arkhipov S.A., Autenshlyus A.I. The Potential Application of the Random Forest Method to Explore the Association between MicroRNAs and Tissue and Molecular Genetic Biomarkers of Breast Cancer. *Bulletin of Siberian Medicine.* 2026;25(2):91–100. https://doi.org/10.20538/1682-0363-2026-2-91-100.

Возможность использования метода «случайного леса» для поиска взаимосвязи микроРНК с тканевыми и молекулярно-генетическими маркерами рака молочной железы

Студеникина А.А.[1, 2], Михайловский М.А.[3], Тимофеев В.С.[3], Архипов С.А.[1, 2], Аутеншлюс А.И.[1, 2]

[1] *Новосибирский государственный медицинский университет (НГМУ)*
Россия, 630091, г. Новосибирск, Красный пр., 52

[2] *Научно-исследовательский институт молекулярной биологии и биофизики (НИИМББ), Федерального исследовательского центра фундаментальной и трансляционной медицины (ФИЦ ФТМ)*
Россия, 630060, г. Новосибирск, ул. Тимакова, 2

[3] *Новосибирский государственный технический университет (НГТУ)*
Россия, 630073, г. Новосибирск, пр. Карла Маркса, 20

РЕЗЮМЕ

Цель. Поиск возможности выявления с помощью метода «случайного леса» (МСЛ) взаимосвязи между уровнем микроРНК miR-181a, miR-25 и экспрессией маркеров: Е-кадгерина (CDH1), коллагена II (CII), интегрина бета-1 (ITGB1), циклина D1 (CCND1) и белка-костимулятора Т-лимфоцитов (CD86), а также их ассоциации с клинически значимыми молекулярно-генетическими маркерами (ER, PR, HER2, Ki-67) у пациенток с раком молочной железы (РМЖ).

Материалы и методы. У 44 пациенток с инвазивной протоковой карциномой молочной железы (без неоадъювантной терапии) определяли экспрессию микроРНК miR-181a и miR-25 в сыворотке крови, используя цифровую капельную полимеразную цепную реакцию (ПЦР). Экспрессию CDH1, CII, ITGB1, CCND1, CD86, ER, PR, HER2 и Ki-67 анализировали с помощью иммуногистохимии в образцах опухолевой ткани. Для анализа взаимосвязей использовали корреляционный анализ Спирмена и алгоритм МСЛ.

Результаты. Корреляционный анализ не выявил статистически значимых линейных взаимосвязей между исследуемыми параметрами ($p > 0,05$). Метода «случайного леса» показал наличие нелинейных ассоциаций: для miR-181a оптимальная прогностическая модель включала CDH1 и CCND1 (MAE = 4,267), для miR-25 – ITGB1 и CCND1 (MAE = 16,255). Из клинико-патологических характеристик статистически значимые положительные корреляции выявлены для CII с ER ($r = 0,498$; $p = 0,001$) и PR ($r = 0,354$; p = 0,018). Модели МСЛ продемонстрировали наибольшую точность предсказания HER2-статуса (82,6%) при использовании комбинации ITGB1, miR-181a и CDH1.

Заключение. Алгоритм МСЛ выявил сложные нелинейные взаимосвязи между miR-181a, miR-25 и тканевыми маркерами опухолевой прогрессии, которые ранее не обнаружены стандартными статистическими методами. Полученные результаты носят поисковый характер и формируют основу для дальнейших исследований по валидации данных биомаркеров и их роли в патогенезе различных подтипов РМЖ.

Ключевые слова: рак молочной железы, HER2, микроРНК, машинное обучение, метод «случайного леса», молекулярные маркеры, эпителиально-мезенхимальный переход

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Источник финансирования. Исследование выполнено в рамках темы государственного задания ФГБОУ ВО «Новосибирский государственный медицинский университет» Министерства здравоохранения Российской Федерации (регистрационный номер 115060410035) и в рамках темы НИР ФГБНУ «Федеральный исследовательский центр фундаментальной и трансляционной медицины», Научно–исследовательский институт молекулярной биологии и биофизики (регистрационный номер 125031203556-7).

Соответствие принципам этики. Все пациенты подписали информированное согласие на участие в исследовании. Исследование одобрено этическим комитетом НИИМББ, НИЦ АТМ (протокол № 28 от 27.09.2023).

Для цитирования: Студеникина А.А., Михайловский М.А., Тимофеев В.С., Архипов С.А., Аутеншлюс А.И. Возможность использования метода «случайного леса» для поиска взаимосвязи микроРНК с тканевыми и молекулярно-генетическими маркерами рака молочной железы. *Бюллетень сибирской медицины.* 2026;25(2):91–100. https://doi.org/10.20538/1682-0363-2026-2-91-100.

INTRODUCTION

Breast cancer (BC) annually claims the lives of thousands of women, underscoring the need for an in-depth study of the molecular mechanisms of its pathogenesis. Particular attention should be paid to analyzing the association of potential markers with classical molecular genetic indicators: estrogen receptors (ER), progesterone receptors (PR), human epidermal growth factor receptor 2 (HER2), and the Ki-67 proliferation index.

MicroRNAs (miRs) are small non-coding RNAs that regulate gene expression at the post-transcriptional level [1]. They function as both oncogenes and tumor suppressors, affecting mRNA stability and translation. Dysregulation of microRNAs is associated with various types of cancer [2]. Specifically, miR-181a and miR-25 are involved in proliferation, invasion, and modulation of the immune response [3, 4]. It is known that miR-181a can reduce ER expression in BC cell culture, and its high level correlates with low PR expression, HER2 overexpression, and a high Ki-67 index [5–7]. MiR-25, in turn, is involved in modulating ER and PR signaling pathways [5].

The study by J. Feng et al. confirms that abnormal microRNA expression is accompanied by the development of cellular transformation towards carcinogenesis, which is also associated with the ability of microRNAs to promote epithelial – mesenchymal transition (EMT) [8]. This process is characterized by a decrease in the expression of epithelial markers, such as E-cadherin (CDH1), and the activation of mesenchymal markers, including integrin beta-1 (ITGB1) [9]. During this process, tumor cells lose intercellular connections and acquire increased mobility and invasiveness, which is associated with a higher degree of malignancy and metastasis to lymph nodes [10]. According to A.H. Ali et al., low CDH1 expression was characteristic of patients with negative ER and HER2 status [11].

We previously showed that BC patients had elevated levels of miR-181a, miR-25, CD86, ITGB1, and cyclin D1 (CCND1) along with reduced CDH1 expression compared to patients with benign breast diseases [12–14]. However, no statistically significant relationships were found between these biomarkers, indicating the need for more sophisticated analytical tools. It should be emphasized that each of these biomarkers plays an important role in BC pathogenesis. CCND1 is a key regulator of the cell cycle and can increase the transcriptional activity of ER, contributing to the formation of resistance to endocrine therapy [15]. CD86 is a co-stimulatory molecule involved in T-lymphocyte activation and the formation of an anti-tumor immune response [16]. Type II collagen (CII), as a major component of the cartilage matrix, may be involved in extracellular matrix remodeling and the creation of metastatic niches in BC [17]. Despite the fact that CCND1, CD86, and CII are not classic inducers of EMT, their influence on key signaling pathways, such as TGF-β, Wnt/β-catenin, and integrin-dependent signaling, suggests their indirect role in EMT progression [17, 18].

Machine learning algorithms offer objective and reproducible tools for identifying hidden patterns and relationships between biomarkers that may be overlooked by traditional analysis, which is a promising direction for studying the molecular mechanisms of the development and progression of oncological diseases [19]. The Random Forest Method (RFM) is a machine learning algorithm based

on constructing an ensemble of decision trees, each of which is built on its own unique subset of data and features (bootstrap sample), and then averaging their predictions. This significantly reduces prediction variance and the tendency to overfit compared to single trees [20].

The first step is to form the training sample $D = \{(X_i, y_i)\}_{t=1}^{N}$, where N is the number of objects, X_i is a set of k features for the i-th observation, and y_i is the value of the target variable. Then, a function that calculates the final prediction of the values of the variable of interest is sought. An ensemble of T regression trees, $\{h_t\}_{t=1}^{T}$, is formed, where h_t is an individual decision tree. Next, a bootstrap sample is formed for each tree h_t: from the training sample D , a training set D_t is created by random sampling of N observations; some observations will appear in D_t several times, and some will not appear at all. The construction of a regression decision tree on D_t begins with the root node, which contains the entire sample D_t . Then, for each node, a subset of features is randomly selected from all features, and among the selected features, the best feature and the best split value (point) are found to maximize the reduction of the variance of the target variable y in the two child nodes. The criterion is the minimization of the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ is the obtained prediction of the target variable, and n is the number of observations. The sample is then split into two based on the found point and features. The splitting procedure is recursively repeated for each child node until one of the stopping criteria is met (maximum tree depth reached, split does not reduce MSE sufficiently, etc.). When the stopping criterion is reached, the prediction in that node becomes the average value of the target variable y of all training observations that fell into that node. The prediction of the trained model for a new observation x is obtained by finding the terminal node for x in each tree h_t of the ensemble. The terminal node is found by following the splitting rules compiled during the training of the decision tree. The prediction of the terminal node is considered as the model's prediction. The final RFM prediction is found by averaging the predictions of all trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t.$$

RFM has shown promising results in predicting the risk of BC development, early detection of recurrences, and predicting a response to chemotherapy [21–23].

The present study is exploratory in nature and aims at determining the possibility of effectively using machine learning algorithms to analyze the relationships between microRNAs and molecular tumor markers in the context of a limited sample size.

The aim of the study was to explore the potential of using the RFM algorithm to identify the association between the levels of miR-181a, miR-25 and the expression of the molecular markers CDH1, CII, ITGB1, CCND1, and CD86, as well as their association with key clinicopathological characteristics (ER, PR, HER2, Ki-67) in BC patients.

MATERIALS AND METHODS

The study involved 44 female patients with the diagnosis of ductal breast carcinoma of no special type, aged 23 to 72 years (mean age 55 years), who received treatment at the Breast Tumor Department of City Clinical Hospital (CCH) No. 1 (Novosibirsk). Inclusion criteria: primary BC without neoadjuvant therapy; histologically confirmed diagnosis of ductal breast carcinoma of no special type; histological tumor grade G2. Exclusion criteria: concomitant malignant neoplasms; acute infectious diseases or exacerbation of chronic infections; distant metastases. The expression of ER, PR, HER2, and Ki-67 was determined according to standard clinical protocols of CCH No. 1 and provided in patient charts. A detailed description of the patients is presented in Table 1.

Table 1

Clinical and Pathological Characteristics of BC Patient Tumors

Parameter	Category	Abs. Number (n)	Share (%)
Tumor Size (T)	T1 (<2 cm)	13	29.5
	T2 (2–5 cm)	31	70.5
Lymph Node Involvement (N)	N0	28	63.6
	N1	12	27.3
	N2	4	9.1
Estrogen Receptors (ER)	ER+ (≥1%)	25	56.8
	ER– (<1%)	19	43.2
Progesterone Receptors (PR)	PR+ (≥1%)	22	50.0
	PR– (<1%)	22	50.0
Human Epidermal Growth Factor Receptor 2 (HER2)	3+	9	20.5
	0 or 1+	35	79.5
Ki-67 Proliferative Index	<20%	18	40.9
	≥20%	26	59.1

MicroRNAs were isolated from 300 µl of blood serum using the NucleoSpin miRNA Plasma Kit (Macherey–Nagel, Germany). M–MuLV–RH (Biolabmix, Russia) and miR-specific primers were used for reverse transcription. Detailed primer characteristics were presented in a previously published work [16]. Quantitative expression was assessed by digital droplet PCR on the QX200 AutoDG system (Bio–Rad, USA) with TaqMan detection. U6 small nuclear RNA served as an internal control sample.

Immunohistochemistry. The expression of the studied markers was determined on deparaffinized tumor tissue sections using monoclonal antibodies: Anti-E-Cadherin (Cat. No. 610181, BD Biosciences, USA), Anti-CD29/Clone 18 (Cat. No. 610467, BD Transduction Laboratories, USA), Anti-CII (COL2A1/M2139, Cat. No. sc-52658, Santa Cruz Biotechnology, Inc., USA), Anti-Cyclin D1 and Anti-B7–2 (PAA585Hu01, Cloud–Clone Corp., China). Visualization was performed using the VECTASTAIN ABC Kit (Vector Laboratories, USA) and analyzed with the ImageJ software (National Instruments Corporation, USA). Marker expression levels were calculated as the percentage of positively stained cells, using at least 10 fields of view at ×400 magnification for each patient.

The expression levels of Her2/neu, estrogen (ER) and progesterone (PR) receptors, as well as the Ki-67 proliferation marker were assessed in BC samples according to the methods [24] recommended for determining molecular genetic subtypes of BC (Gallen International Expert Consensus, 2011). The receptor status for ER and PR was determined using the ER/PR pharmDx Kit (K4071, Dako). The Ultra View Universal DAB Detection Kit (4B5 antibodies) was used to determine the HER2/neu status. The Ki-67 protein (MIB-1 clone) was detected using a set of diagnostic antibodies from Dako. All antibodies were used at a dilution of 1:100. ER and PR expression levels were assessed using the Allred scoring system [25]. Her2/neu expression levels were assessed using the generally accepted scoring system (from 0 to 3+).

Data processing was carried out using the Python programming language (version 3.12). The Scikit – learn package (version 1.4.0) for Data Science and Machine Learning was used to calculate Spearman's rank correlation coefficients (r) and develop the RFM model. Normality of distribution was checked using the Shapiro – Wilk test. The non-parametric Mann – Whitney U-test was used to assess differences in marker expression levels between the patient groups. Data were presented as the median, the 25th and the 75th percentiles Me (Q_{25}; Q_{75}). Differences were considered to be statistically significant at $p < 0.05$.

The quality of RFM regression models was assessed using the Mean Absolute Error (MAE), which indicates the average deviation of predicted values from actual values. The smaller the MAE, the more accurately the model describes the data [26]. MAE is expressed in the same units as the target variable. The MAE formula is:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1} |y_i - \hat{y}_i|,$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of observations. For RFM classification tasks, the precision metric was used, calculated as the ratio of correctly predicted positive cases to the total number of predicted positive cases.

Given the limited sample size ($n = 44$), which did not allow for splitting the data into training and test subsets without losing statistical representativeness, RFM model training and validation were performed on the entire dataset. The relatively small number of predictors (molecular markers) allowed for finding the most suitable model to describe the existing relationships through complete enumeration of all predictor combinations. For each possible combination of predictors, 500 models with different randomness parameters were built, and the quality was assessed by averaging the MAE across all models.

RESULTS

Initial analysis using the Spearman's rank correlation coefficient did not reveal significant relationships between the levels of miR-181a, miR-25, and markers of adhesion, proliferation, and immune response (Table 2, $p > 0.05$ for all comparisons). This indicated the absence of simple linear dependencies and the need for more complex analytical approaches.

Table 2

Correlations between MicroRNA Levels and Molecular Markers

Marker	miR-181a		miR-25	
	r	*p*	*r*	*p*
CDH1	– 0.049	0.753	– 0.018	0.910
ITGB1	0.036	0.815	– 0.071	0.648

End of table 2

Marker	miR-181a		miR-25	
	r	*p*	*r*	*p*
CCND1	0.006	0.970	– 0.124	0.421
CII	0.069	0.657	– 0.091	0.557
CD86	– 0.081	0.600	0.053	0.731

The RFM was used to identify more complex nonlinear associations. The optimal model corresponded to the combination of predictors yielding the smallest MAE. For miR-181a, the best predictive model included the combination of markers CDH1 and CCND1 (MAE = 4.267). For miR-25, the optimal set of predictors was ITGB1 and CCND1 (MAE = 16.255). Figure presents a comparison of predicted and actual values for both models, demonstrating their ability to detect hidden patterns in the data.

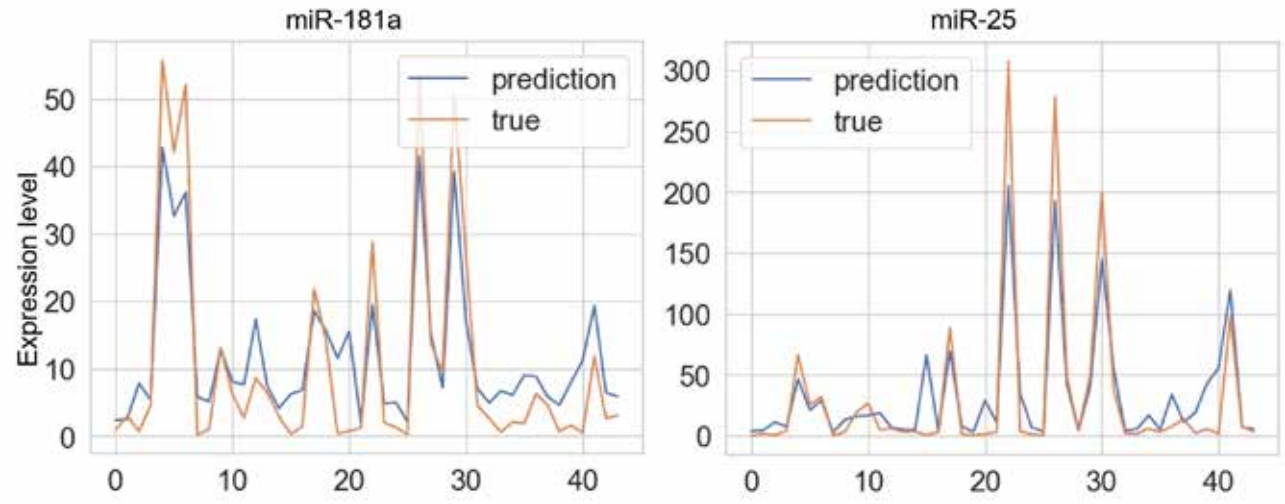

Figure. Comparison of predicted and actual levels of miR-181a and miR-25: X-axis: patients, $n = 44$; Y-axis: microRNA concentration

The analysis of the relationships between the studied biomarkers and key clinicopathological tumor parameters (ER, PR, HER2, Ki-67) showed statistically significant correlations only for CII. Moderate positive correlations were found between CII expression and estrogen receptor status ($r = 0.498$; $p = 0.001$) and progesterone receptor status ($r = 0.354$; $p = 0.018$, Table 3).

Table 3

Correlation between MicroRNA Levels, Molecular Markers, and Molecular Genetic Tumor Parameters

Marker		ER	PR	HER2	Ki-67
miR-181a	*r*	– 0.156	– 0.036	0.150	0.176
	p	0.311	0.819	0.330	0.253
miR-25	*r*	– 0.166	– 0.007	0.057	0.300
	p	0.282	0.967	0.712	0.048
CDH1	*r*	0.089	0.047	0.024	– 0.051
	p	0.568	0.763	0.879	0.744
ITGB1	*r*	0.326	0.180	0.168	– 0.121
	p	0.031	0.243	0.275	0.433
CII	*r*	0.498*	0.354*	0.076	– 0.196
	p	0.001	0.018	0.623	0.202
CCND1	*r*	– 0.016	0.118	0.076	– 0.164
	p	0.917	0.446	0.622	0.288
CD86	*r*	0.030	– 0.058	– 0.277	0.204
	p	0.846	0.708	0.069	0.183

*$p < 0.05$.

Further analysis of marker level differences between the patient groups showed that CII expression was statistically significantly higher in tumors with positive hormone receptor status (Table 4). In ER-positive samples, the *Me* expression level was 20.3% compared to 6.9% in ER-negative samples ($p = 0.010$). A similar trend was noted for the PR status: 20.2% in PR-positive versus 9.8% in PR-negative tumors ($p = 0.016$). For ITGB1, there was a trend toward higher expression in ER-positive tumors (*Me* 18.8% vs. 11.5%), but the difference did not reach statistical significance ($p = 0.052$).

The application of the RFM for predicting the BC receptor status showed varying levels of accuracy depending on the marker. The best result was achieved for the HER2 status (82.6% accuracy) using the combination of ITGB1, miR-181a, and CDH1.

Table 4

Analysis of the Expression of the Studied Markers Depending on Molecular Genetic Tumor Parameters, *Me* (Q_{25}; Q_{75})

Parameters	ER–, *n* = 19	ER+, *n* = 25	PR–, *n* = 22	PR+ *n* = 22	HER2–0 или 1+, *n* = 35	HER2+ 3+, *n* = 9	Ki-67 <20%, *n* = 18	Ki-67 ≥20%, *n* = 26
miR-181a	2.9 (1.6–14.4)	3.0 (0.8–11.3)	3.0 (1.5–18.0)	2.9 (0.9–10.3)	2.9 (1.0–9.4)	3.1 (1.4–52.8)	2.6 (0.8–18.1)	4.5 (1.6–13.1)
	p = 0.414		*p* = 0.438		*p* = 0.383		*p* = 0.557	
miR-25	5.0 (3.1–26.7)	3.3 (1.3–25.8)	3.6 (2.2–28.1)	4.5 (1.3–24.2)	3.6 (1.8–25.5)	5.0 (0.8–49.2)	3.3 (1.0–16.6)	5.5 (3.1–38.6)
	p = 0.337		*p* = 0.656		*p* = 0.977		*p* = 0.162	
CDH1	81.5 (30.8–94.0)	90.3 (59.5–96.4)	84.0 (36.8–94.5)	89.4 (67.1–96.3)	88.5 (58.9–94.8)	74.1 (31.4–96.3)	88.5 (49.3–95.4)	86.4 (38.8–94.8)
	p = 0.507		*p* = 0.664		*p* = 0.942		*p* = 0.925	
ITGB1	11.5 (5.4–19.1)	18.8 (10.5–80.6)	12.0 (6.1–20.4)	18.5 (8.2–80.1)	13.5 (6.3–46.1)	16.5 (9.8–24.9)	14.3 (5.9–26.4)	16.4 (7.4–46.1)
	p = 0.052		*p* = 0.236		*p* = 0.827		*p* = 0.690	
CII	6.9 (6.3–15.7)	20.3 (12.7–30.0)	9.8 (6.4–15.9)	20.2 (11.2–27.3)	14.4 (6.4–26.5)	14.4 (9.0–18.5)	15.7 (8.4–27.2)	14.4 (6.5–20.5)
	p = 0.010*		*p* = 0.016*		*p* = 0.942		*p* = 0.259	
CCND1	93.9 (77.1–97.2)	93.9 (83.1–97.6)	93.9 (74.6–96.8)	93.9 (85.8–97.7)	93.9 (80.7–97.3)	93.9 (87.9–97.4)	93.9 (83.1–98.0)	93.9 (77.1–96.7)
	p = 0.849		*p* = 0.451		*p* = 0.759		*p* = 0.207	
CD86	21.1 (10.9–26.4)	19.9 (15.2–29.2)	21.3 (10.9–27.3)	19.0 2.4–29.0)	20.7 (12.8–30.2)	19.0 (8.9–23.6)	20.7 (11.2–29.2)	19.9 (16.1–26.4)
	p = 0.924		*p* = 0.673		*p* = 0.244		*p* = 0.991	

The accuracy for the ER status was 69.5% using miR-25, CDH1, and CII. The accuracy for the PR status was 66.3% with predictors miR-181a, ITGB1, and CCND1. In contrast, predicting the Ki-67 level using miR-25 and CCND1 was the least accurate, indicated by high mean absolute error (MAE = 30.411).

DISCUSSION

The obtained data demonstrate the complex nature of the associations between the expression of miR-25, miR-181a and molecular markers in BC. The ability of the RFM algorithm to detect nonlinear associations that were not detectable by traditional correlation analysis methods emphasizes the potential of machine learning in studying the molecular mechanisms of carcinogenesis.

The identified association of miR-181a with CDH1 and CCND1 has a biological rationale. MiR-181a can influence E-cadherin by activating TGF-β/Smad-dependent mechanisms, promoting EMT [27]. Furthermore, miR-181a activates the PI3K-Akt signaling pathway, which enhances cell growth and proliferation, thereby influencing CDH1 expression [28]. The involvement of both CCND1 and CDH1 in regulating the Wnt/β-catenin signaling pathway may explain their joint inclusion in the predictive model for miR-181a [29]. A study on multiple myeloma by X. Yan et al. outlined the mechanisms underlying the regulation of the cell cycle by miR-181a through influencing CCND1 expression [30].

The connection of miR-25 with ITGB1 and CCND1 also finds confirmation at the molecular level. MiR-25 is involved in activating the Wnt/β-catenin pathway, promoting the nuclear accumulation of β-catenin and enhancing the transcription of target genes, including CCND1 [31]. ITGB1 is involved in the ILK-Akt signaling pathway, which enhances EMT, potentially indicating common regulatory mechanisms with miR-25 [32].

The statistically significant associations of CII with hormonal receptors (ER and PR) are of particular interest, as this marker is not well-studied in the context of BC. In breast cancer, different types of collagen demonstrate varied expression profiles [33]. It has been shown that different types of collagen have varying effects on the development of malignant growth, particularly BC, by impacting crucial cellular processes, including apoptosis, proliferation, and metastasis [33, 34]. For example, in triple-negative BC, decreased expression of COL2A1 (CII) is associated with an increased grade of tumor malignancy, metastasis, poor prognosis, and decreased survival [33]. Conversely, a decrease in COL1A1 expression may be associated with a decrease in BC malignancy. It has been established that COL1A1 expression is linked to ER/PR expression and metastasis status. In patients with ER-positive BC, higher COL1A1 expression correlates with aggressive cellular behavior and poor prognosis [34]. Our findings of increased CII expression in hormone-positive tumors may reflect

specific characteristics of extracellular matrix remodeling in luminal BC subtypes, which requires further investigation.

The high accuracy of HER2 status prediction (over 80%) using the combination of ITGB1, miR-181a, and CDH1 points to their potential diagnostic significance. This combination likely reflects the biological characteristics of the aggressive HER2-positive BC subtype comprehensively. ITGB1 enhances tumor aggressiveness through cellular remodeling, while miR-181a further amplifies proliferation and invasive potential by suppressing CDH1 expression. In contrast, the models for ER and PR status showed only moderate accuracy. This is most likely due to the significant heterogeneity of luminal BC subtypes, which vary greatly in both molecular characteristics and clinical behavior.

It should be emphasized that this study is exploratory in nature and has several significant limitations that must be considered when interpreting the results. The main limitations are the relatively small sample size and its heterogeneity in terms of clinical, morphological, and molecular genetic parameters. Due to the small number of observations, we did not split the data into training and test samples, which increases the risk of model overfitting and limits the generalizability of the results. Therefore, the findings should be considered as preliminary, requiring validation in larger and independent patient cohorts.

CONCLUSION

The obtained data demonstrate that the application of the random forest machine learning method is an effective approach for identifying complex, nonlinear associations between circulating microRNAs and markers in breast tumor tissue. The established associations (specifically between miR-181a, CDH1, and CCND1, as well as between miR-25, ITGB1, and CCND1) were not detected using standard correlation analysis methods. These results are of fundamental importance as they allow for the hypothesis of coordinated regulation of key processes in tumor progression: loss of cell adhesion (CDH1), enhanced migration (ITGB1), and disruption of the cell cycle (CCND1). Although the resulting models are not intended for direct clinical application at this stage, they represent an important starting point for future research in this direction. The findings underscore the potential of biomarker combinations for patient stratification and prediction of molecular BC subtypes, especially the HER2 status.

REFERENCES

1. Rani V., Sengar R.S. Biogenesis and mechanisms of microRNA-mediated gene regulation. *Biotechnol. Bioeng*. 2022;119 (3):685–692. DOI: 10.1002/bit.28029.
2. Kumar S., Ranga A. Role of miRNAs in breast cancer development and progression: Current research. *BioFactors*. 2025;51(1):e2146. DOI: 10.1002/biof.2146.
3. Zhai Z., Mu T., Zhao L., Li Y., Zhu D., Pan Y. MiR-181a-5p facilitates proliferation, invasion, and glycolysis of breast cancer through NDRG2-mediated activation of PTEN/AKT pathway. *Bioengineered*. 2022;13(1):83–95. DOI: 10.1080/21655979.2021.2006974.
4. Zhao T., Meng W., Chin Y., Gao L., Yang X., Sun S. et al. Identification of miR-25-3p as a tumor biomarker: Regulation of cellular functions via TOB1 in breast cancer. *Mol. Med. Rep*. 2021; 23(6):406. DOI: 10.3892/mmr.2021.12045.
5. Kalinina T.S., Kononchuk V.V., Alekseenok E.Y., Obukhova D., Sidorov S.V., Strunkin D. et al. Expression of estrogen receptor- and progesterone receptor-regulating microRNAs in breast cancer. *Genes*. 2021;12(4):582. DOI: 10.3390/genes12040582.
6. Andreeva O.E., Sorokin D.V., Mikhaevich E.I., Bure I.V., Shchegolev Y.Y., Nemtsova M.V. et al. Towards unravelling the role of ERα-targeting miRNAs in the exosome-mediated transferring of the hormone resistance. *Molecules*. 2021;26(21):6661. DOI: 10.3390/molecules26216661.
7. Gilam A., Shai A., Ashkenazi I., Sarid L.A., Drobot A., Bickel A. et al. MicroRNA Regulation of Progesterone Receptor in Breast Cancer. *Oncotarget*. 2017;8(16):25963–25976. DOI: 10.18632/oncotarget.15657.
8. Feng J., Hu S., Liu K., Sun G., Zhang Y. The role of microRNA in the regulation of tumor epithelial-mesenchymal transition. *Cells*. 2022;11(13):1981. DOI: 10.3390/cells11131981.
9. Mendez M.J., Hoffman M.J., Cherry E.M., Lemmon C.A., Weinberg S.H. Cell fate forecasting: a data-assimilation approach to predict epithelial-mesenchymal transition. *Biophys. J.* 2020;118(7): 1749–1768. DOI: 10.1016/j.bpj.2020.02.011.
10. Manfioletti G., Fedele M. Epithelial-mesenchymal transition (EMT). *Int. J. Mol. Sci.* 2023;24(14):11386. DOI: 10.3390/ijms241411386.
11. Ali A.H., Haider G., Ahmed K., Hashmi M., Kerio P., Irshad R. Correlation between e-cadherin and hormone receptor status among breast cancer patients. *Journal of the College of Physicians and Surgeons Pakistan*. 2020;30(10):1030–1034. DOI: 10.29271/jcpsp.2020.10.1030.
12. Autenshlyus A.I., Perepechaeva M.L., Studenikina A.A., Grishanova A.Y., Lyakhovich V.V. Serum miR-181a and miR-25 Levels in Patients with Breast Cancer or Benign Breast Disease. *Dokl. Biochem. Biophys.* 2023;512(1):279–283. DOI: 10.1134/S1607672923700436.
13. Studenikina A.A., Arkhipov S.A., Varaksin N.A., Proskura A.V., Autenshlyus A.I. Breast Biomarkers for the Detection of Malignancy in Benign Diseases in Women. *Russian Jour-*

nal of Preventive Medicine. 2024;27(1):63–66. (In Russ.). DOI: 10.17116/profmed20242701163.
14. Studenikina A.A., Arkhipov S.A., Zhurakovsky I.P., Proskura A.V., Autenshlyus A.I. Expression of Markers of Epithelial – Mesenchymal Transition in Breast Cancer. *Archive of Pathology*. 2023;85(4):12–17. (In Russ.). DOI: 10.17116/patol20238504112.
15. Jeffreys S.A., Becker T.M., Khan S., Soon P., Neubauer H., de Souza P. et al. Prognostic and predictive value of CCND1/Cyclin D1 Amplification in Breast Cancer with a Focus on Postmenopausal Patients: A Systematic Review and Meta-Analysis. *Frontiers in Endocrinology*. 2022;13:895729. DOI: 10.3389/fendo.2022.895729.
16. Wennhold K., Thelen M., Lehmann J., Schran S., Preugszat E., Garcia-Marquez M. et al. CD86+ antigen-presenting B cells are increased in cancer, localize in tertiary lymphoid structures, and induce specific T-cell responses. *Cancer Immunol. Res*. 2021;9(9):1098–1108. DOI: 10.1158/2326-6066.CIR-20-0949.
17. Yang H.L., Tsai Y.C., Korivi M., Chang C.T., Hseu Y.C. Lucidone Promotes the cutaneous wound healing process via activation of the PI3K/AKT, Wnt/β-catenin and NF-κB signaling pathways. *Biochimica et Biophysica Acta - Molecular Cell Research*. 2017;1864(1):151–168. DOI: 10.1016/j.bbamcr.2016.10.021.
18. Li X., Jin Y., Xue J. Unveiling collagen's role in breast cancer: insights into expression patterns, functions and clinical implications. *International Journal of General Medicine*. 2024;17:1773–1787. DOI: 10.2147/IJGM.S463649.
19. Khan M.J., Singh A.K., Sultana R., Singh P.P., Khan A., Saxena S. Breast cancer: A comparative review for breast cancer detection using machine learning techniques. *Cell Biochem. Funct.* 2023;41(8):996–1007. DOI: 10.1002/cbf.3868.
20. Breiman L. Random forests. Mach. Learn. 2001;45:5–32. DOI: 10.1023/A:1010933404324.
21. Zuo D., Yang L., Jin Y., Qi H., Liu Y., Ren L. Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Med. Inform. Decis. Mak.* 2023;23(1):276. DOI: 10.1186/s12911-023-02377-z.
22. Jin Y., Lan A., Dai Y., Jiang L., Liu S. Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. *Eur. J. Med. Res.* 2023;28(1):394. DOI: 10.1186/s40001-023-01361-7.
23. Arika R.N., Mindila A., Cheruiyo W. Machine learning algorithms for breast cancer diagnosis: challenges, prospects and future research directions. *J. Oncol.* 2023;5(1):1–31. DOI: 10.30564/jor.v5i1.4977.
24. Kondov B., Milenkovikj Z., Kondov G., Petrushevska G., Basheska N., Bogdanovska-Todorovska M. et al. Presentation of the molecular subtypes of breast cancer detected by immunohistochemistry in surgically treated patients. *J. Med. Sci.* 2018;6(6): 961–967. DOI: 10.3889/oamjms.2018.231.
25. Yumpu [Internet]. ER/PR PharmDx Interpretation Manual. – Dako. URL: https://www.yumpu.com/en/document/read/18500132/er-pr-pharmdxtm-interpretation-manual-dako. Available at: 09.20.2013.
26. Willmott C.J., Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* 2005;30(1):79–82. DOI: 10.3354/cr030079.
27. Li L., Xu Q.H., Dong Y.H., Li G.X., Yang L., Wang L.W. et al. MiR-181a upregulation is associated with epithelial-to-mesenchymal transition (EMT) and multidrug resistance (MDR) of ovarian cancer cells. *Eur. Rev. Med. Pharmacol. Sci.* 2016;20(10):2004–2010.
28. Strotbek M., Schmid S., Sánchez-González I., Boerries M., Busch H., Olayioye M.A. miR-181 elevates Akt signaling by co-targeting PHLPP2 and INPP4B phosphatases in luminal breast cancer. *Int. J. Cancer.* 2017;140(10):2310–2320. DOI: 10.1002/ijc.30661.
29. Govatati S., Singamsetty G.K., Nallabelli N., Malempati S., Rao P.S., Madamchetty V.K.K. et al. Contribution of cyclin D1 (CCND1) and E-cadherin (CDH1) alterations to colorectal cancer susceptibility: a case-control study. *Tumor Biol.* 2014;35(12):12059–12067. DOI: 10.1007/S13277-014-2505-9.
30. Yan X., Gao M., Zhang P., Ouyang G., Mu Q., Xu K. MiR-181a functions as an oncogene by regulating CCND1 in multiple myeloma. *Oncol. Lett.* 2020;20(1):758–764. DOI: 10.3892/OL.2020.11579.
31. Yao J., Li G., Liu M., Yang S., Su H., Ye C. lnc-MICAL2 1 sponges miR 25 to regulate DKK3 expression and inhibits activation of the Wnt/β catenin signaling pathway in breast cancer. *Int. J. Mol. Med.* 2022;49(2):23. DOI: 10.3892/ijmm.2021.5078.
32. Jiang X., Wang J., Zhang K., Tang S., Ren C., Chen Y. The role of ITGB1-ILK-Akt signaling-mediated epithelial-mesenchymal transition of liver epithelial cells and chemoresistance and radioresistance in hepatocellular carcinoma cells. *Med. Oncol.* 2015;32(5):141. DOI: 10.1007/S12032-015-0595-X.
33. Vora H., Gavali M., Patel N. A Study of Col2A1 expression in triple-negative brest cancer patients. *Glob. J. Clinic. Medici. Medica. Res.* 2023;1:1–5.
34. Li X., Jin Y., Xue J. Unveiling collagen's role in breast cancer: insights into expression patterns, functions and clinical implications. *Int. J. Gen. Med.* 2024;17:1773–1787. DOI: 10.2147/IJGM.S463649.

Author Contribution

Studenikina A.A. – carrying out of the experiments, discussion of the results of the studies, drafting of the article. Mihailovsky M.A. – carrying out of the experiments, discussion of the results of the studies. Timofeev V.S., Autenshlyus A.I. – conception of the study, supervision of the work, discussion of the results of the studies, editing of the article. Arkhipov S.A. – carrying out of the experiments, discussion of the results of the studies, editing of the article.

Author Information

Studenikina Anastasiia A. – Cand. Sci. (Med.), Senior Researcher, Central Research Laboratory, Novosibirsk State Medical University, Novosibirsk; Researcher, Institute of Molecular Biology and Biophysics, Federal Research Center of Fundamental and Translational Medicine, Novosibirsk, aastudenikina@frcftm.ru, https://orcid.org/0000-0003-3936-1316

Mihailovsky Michael A. – Bachelor's Degree Student, Novosibirsk State Technical University, Novosibirsk, maihailovsky18@mail.ru, https://orcid.org/0009-0009-2272-6295

Timofeev Vladimir S. – Dr. Sci. (Engineering), Professor, Department of Theoretical and Applied Computer Science, Dean of the Faculty of Applied Mathematics and Computer Science, Novosibirsk State Technical University, Novosibirsk, v.timofeev@corp.nstu.ru, https://orcid.org/0000-0002-6352-3218

Arkhipov Sergey A. – Dr. Sci. (Biology), Leading Researcher, Central Research Laboratory, Novosibirsk State Medical University, Novosibirsk; Senior Researcher, Institute of Molecular Biology and Biophysics, Federal Research Center of Fundamental and Translational Medicine, Novosibirsk, arhipowsergei@yandex.ru, https://orcid.org/ 0000-0002-1390-4426

Autenshlyus Aleksandr I. – Dr. Sci. (Biology), Professor, Head of the Central Research Laboratory, Novosibirsk State Medical University, Novosibirsk; Chief Researcher, Institute of Molecular Biology and Biophysics, Federal Research Center of Fundamental and Translational Medicine, Novosibirsk, lpciip@211.ru, https://orcid.org/0000-0001-7180-010X

(✉) **Studenikina Anastasiia A.,** aastudenikina@frcftm.ru

Received on June 23, 2025;
approved after peer review on September 22, 2025;
accepted on October 30, 2025