applied sciences

MDPI

Systematic Review

Machine Learning and miRNAs as Potential Biomarkers of Breast Cancer: A Systematic Review of Classification Methods

Jorge Alberto Contreras-Rodríguez [1], Diana Margarita Córdova-Esparza [1], María Zenaida Saavedra-Leos [2] and Macrina Beatriz Silva-Cázares [2,*]

[1] Facultad de Informática, Universidad Autónoma de Querétaro, Queretaro 76230, Mexico; jcontreras19@alumnos.uaq.mx (J.A.C.-R.); diana.cordova@uaq.mx (D.M.C.-E.)
[2] Coordinación Académica Región Altiplano, Universidad Autónoma de San Luis Potosí, San Luis Potosí 78760, Mexico; zenaida.saavedra@uaslp.mx
* Correspondence: macrina.silva@uaslp.mx

Abstract: This work aims to offer an analysis of empirical research on the automatic learning methods used in detecting microRNA (miRNA) as potential markers of breast cancer. To carry out this study, we consulted the sources of Google Scholar, IEEE, PubMed, and Science Direct using appropriate keywords to meet the objective of the research. The selection of interesting articles was carried out using exclusion and inclusion criteria, as well as research questions. The results obtained in the search were 36 articles, of which PubMed = 14, IEEE = 8, Science Direct = 4, Google Scholar = 10; among them, six were selected, since they met the search perspective. In conclusion, we observed that the machine learning methods frequently mentioned in the reviewed studies were Support Vector Machine (SVM) and Random Forest (RF), the latter obtaining the best performance in terms of precision.

Keywords: machine learning; micro-RNA; breast cancer; biomarkers; classification methods

Citation: Contreras-Rodríguez, J.A.; Córdova-Esparza, D.M.; Saavedra-Leos, M.Z.; Silva-Cázares, M.B. Machine Learning and miRNAs as Potential Biomarkers of Breast Cancer: A Systematic Review of Classification Methods. *Appl. Sci.* **2023**, *13*, 8257. https://doi.org/10.3390/app13148257

Academic Editors: Giorgio Leonardi and Enno van der Velde

Received: 19 June 2023
Revised: 10 July 2023
Accepted: 14 July 2023
Published: 17 July 2023

1. Introduction

Cancer is characterized by cellular dysregulation that can be modified by genetic control at the post-transcriptional and transductional levels, which can be regulated through cell cycle control over the expression levels of related genes. Therefore, modifications are mainly described by microRNA (miRNA) methylation and transcriptional processes [1]. On the other hand, breast cancer (BRCA) is a type of cancer that affects the epithelial cells of the mammary gland, where cell multiplication occurs abnormally and in an uncontrolled manner, thus developing the formation of malignant tumors. Breast cancer cells arise from milk-producing glands called lobules and ducts, which are channels responsible for transporting milk secreted by the lobules to the nipple [2]. In this regard, miRNAs are small non-coding RNAs that function as important post-transcriptional genetic regulators of various biological functions. In general, miRNAs downregulate gene expression by binding to their selective messenger RNAs (mRNAs), which can lead to the degradation or inhibition of mRNA translation, depending on the levels of complementation with the target sequence. Abnormal expression of these miRNAs has been implicated in the etiology of several human diseases [3]. However, health-related processes generate a large amount of complex information to analyze. This is mainly due to the amount of data, the speed of production, and the variety, e.g., text, images, and administrative files. Tools such as machine learning or other data analysis techniques can overcome these difficulties by providing fast and reliable information to help make decisions [4]. Adding to the above, the expression of miRNAs was identified through probability under null distributions the sample result equal to or more extreme than the one observed, which is defined as the *p*-value and is interpreted as the smallest level of significance, i.e., the "cut-off level," since the observed result would be considered significant at all levels greater than or equal to

the *p*-value, but not significant at the smallest levels [5]. However, in most gene expression studies, the genes of greatest interest are those with large relative differences. The relative difference, or fold change, is a basic and widely used measure to identify differential gene expression [6–8].

Machine literacy refers to the capability of a computer system to use statistical styles to "learn" and "acclimate" data for the purpose of prognosticating issues without unequivocal programming (ML) [9]. This approach involves the creation and training of one or further classifiers using training data attained from model organisms that retain both significant and insignificant inheritable traits. The trained classifier is also applied to prognosticate the significance of genes within the target organism. It can be inferred that accurate prognostications bear the vacuity of high-quality data and robust machine literacy ways. The generally employed ways include supervised, semi-supervised, unsupervised, and underpinning literacy [10,11].

ML algorithms are generally distributed as supervised, unsupervised, and deep learning. When considering ML, clinicians should consider several crucial generalities. A simplified approach to developing and enforcing ML algorithms involves dividing the available data into three subsets training data for optimizing the named algorithm and estimating parameters, a test dataset for assessing the performance of the trained algorithm, and a confirmation dataset from a different source rather than the training and test datasets. The confirmation step, although occasionally challenging due to limited data vacuity, provides a more dependable assessment of the algorithm's performance beyond the training dataset [12]. In classification tasks with balanced datasets, standard performance measures similar as delicacy, perceptivity, particularity, and perfection are generally used [13,14]. Nevertheless, for imbalanced datasets where the number of cases is significantly lower than controls, further robust performance observers for class distribution are recommended. Exemplifications include the F1 score, area under the curve (AUC), and Cohen's Kappa [13–15].

In this article, we report on a systematic review of studies on machine learning methods used to classify potential miRNA targets in breast cancer. To offer an overview of empirical research in this field, the types of machine learning methods that exist in the field of breast cancer, characteristics of miRNAs used for BRCA prediction, databases used to carry out the study, metrics used in the performance of the machine learning method and the main results.

2. Materials and Methods

The research questions were: What are the machine learning methods currently applied in the classification of BRCA miRNAs? What is the performance obtained by applying machine learning methods for the classification of BRCA miRNAs? What are the machine learning methods that use the *p*-value and fold change as features to determine the differential expression and classification of BRCA miRNAs?

In response to the above-mentioned research questions, a systematic review of the published scientific literature on technology and in relation to the biology of miRNAs as potential biomarkers of breast cancer has been carried out in the present study. For its preparation, the guidelines of the declaration followed PRISMA to show the completion of the systematic review in Figure 1 [16,17].

Initial Search

We started the literature search in September 2022 using the terms "machine learning for breast cancer miRNA targets" in the PubMed and Google Scholar databases. Although these searches yielded 21 results (Supplementary Materials) on the PubMed platform and 17,000 on the Google Scholar platform, only the majority did not include the keywords above, and we decided to review each of the pages of the results obtained; we observed that the most relevant articles were those who were on the first result pages of the aforementioned search engines.

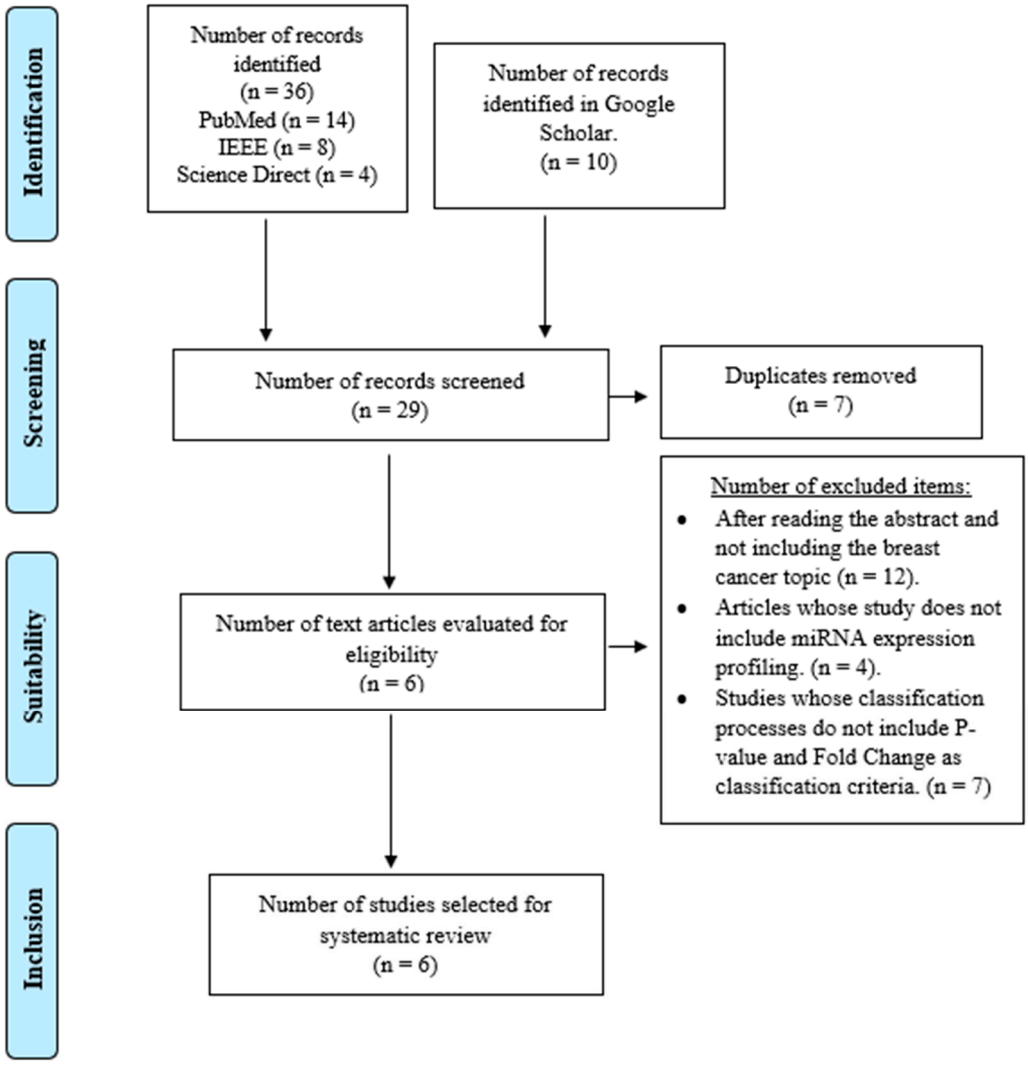

Figure 1. PRISMA flow diagram at four levels shows the elaboration process and different phases of the research work [17].

It was also decided to search the IEEE repository with the terms "machine learning cancer miRNA", which yielded 18 results, but unlike the PubMed and Google Scholar platforms, these included the most used terms in the search.

We also consulted the CONRICyT and Science Direct search platforms in which we used the terms "*Machine learning, breast cancer, miRNA expression*". The CONRICyT search engine determined resulted in different search resources that included the number of articles found. Related to what was stated in the search terms, some of the repositories with the highest results are the following: biblat (*n* = 6747), BMC (*n* = 11,048), DOAJ (*n* = 60,649), Hindawi (10,000), LIPPINCOT (*n* = 4476), PQDT OPEN (*n* = 1861), PubMed Central (*n* = 3522), REDALYC (*n* = 2721), ECLAC Repository (*n* = 1538), SPRINGER (*n* = 3687), and The National Academies Press (*n* = 6642). It is worth mentioning that in most of the results of each repository, only the keyword "machine learning" was included, except for the PubMed Central resource, where the relationship of the results with all the search words was abundant in the results. On the other hand, the number of results obtained by Science Direct was 952, which in turn showed the year of publication from 2003 to 2022, with the characteristic of the articles found being that most were related only to the term "machine learning" and just a few with all the research words.

Systematic search
Search inclusion criteria.

- Empirical investigations;
- Articles written in English and Spanish;
- Include the term breast cancer;
- Research that was published between 2018 and 2022;
- About machine learning methods used in miRNA classification;
- Includes the characteristic of the miRNA expression profile for their classification;
- Articles including the *p*-value and fold change as miRNA classification features.

Exclusion criteria.

- Articles that do not include the topic of breast cancer;
- Studies published in a period of less than 2018;
- Those that do not contain the miRNA expression profile as a classification feature;
- Studies that do not include the *p*-value and fold change as miRNA classification features.

The systematic search was carried out in September 2022 on the PubMed, Google Scholar, CONRICyT, IEEE, and Science Direct platforms, delimiting the publications carried out between 2018 and 2022, respectively.

Also taking these criteria into account, and only by reading the title, 36 articles were considered interesting in Figure 2, of which 7 were eliminated, as when reviewing the authors and the content, duplication was observed. We proceeded to read the summary and content where 12 were excluded, since they did not address the issue of breast cancer in particular. Those that did not contain the concepts of miRNA expression profile, *p*-value, and fold change as a classification characteristic were also separated, with a total of 11.

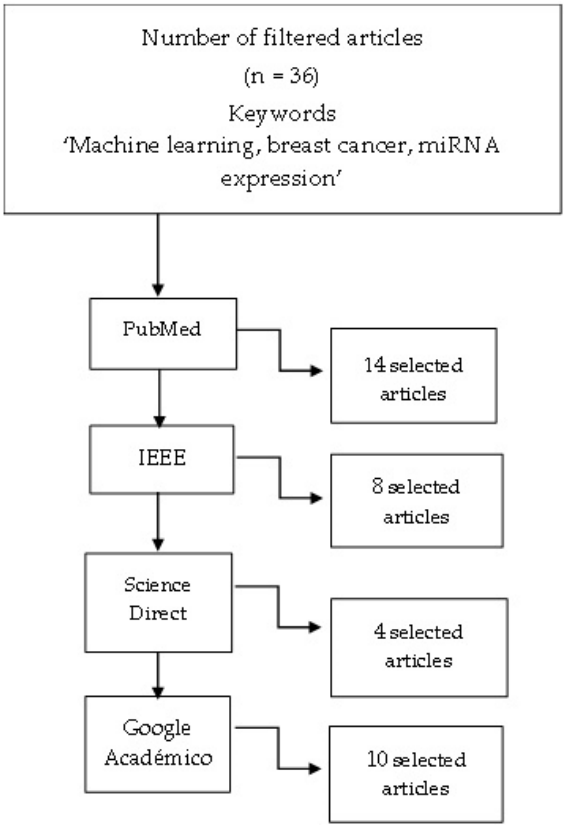

Figure 2. Thirty-six identified studies were obtained: fourteen were selected in PubMed, eight were obtained in IEEE, four were obtained in Science Direct, and ten were obtained in Google Scholar. Before proceeding to the selection of articles, these results were obtained based on the inclusion and exclusion criteria listed above.

Finally, six articles that met the inclusion criteria were chosen and selected to carry out the systematic review. All of them include the topic of breast cancer and use features such as miRNA expression profile, *p*-value, and fold change that, with the application of machine learning allow for miRNA classification in the detection of potential targets in BRCA.

3. Results

An analysis of the selected studies can be found in Table 1, including the synthesis that follows the order that we have considered most pertinent to facilitate the understanding and integration of the results.

Naorem et al. [18] designed a study that focused on triple-negative breast cancer (TNBC), a subtype of breast cancer with a poor clinical outcome for which no specific approved treatment exists. MicroRNAs (miRNAs) have been identified as promising biomarkers with an important role in human cancer tumorigenesis. Due to the growing dataset of TNBC miRNA profiles, their investigation requires proper analysis. The focal point of this exploration lies in the regulation of miRNAs, which involves their over- and down-regulation, determined by specific criteria like fold change and *p*-value. Different studies collect lists of up-regulated and down-regulated miRNAs, prioritizing them grounded on their expression change and *p*-value statistics. The significance of miRNA expression in TNBC is determined by the *p*-value or fold change of each miRNA. Several studies have linked a miRNA metasignature (hsa-miR-135b-5p, hsa-miR-18a-5p, hsa-miR-9-5p, hsa-miR-522-3p, hsa-miR-190b, hsa-miR-9a) that exhibits significant differences and demonstrates high prophetic delicacy. These linked miRNAs are implicit as individual biomarkers for CMTN, warranting farther analysis of their pivotal part in TNBC.

Yu et al. [19] conducted exploration concentrated on exploring the mechanisms of gene relations specific to each subtype of breast cancer, as these mechanisms play a pivotal part in acclimatizing substantiated treatments. To achieve this, they incorporated the natural significance of genes deduced from gene nonsupervisory networks into the analysis of discriminatory gene expression. This integration allowed them to identify weighted differentially expressed genes (weighted DEGs) that retain natural significance deduced from the gene nonsupervisory network. By exercising weighted SDR computations, they developed double classifiers that displayed strong performance in terms of "perceptivity", "particularity", "delicacy", "F1 score", and "AUC" criteria. These classifiers were suitable to distinguish between control and experimental groups, furnishing new perceptivity through gene ontology (GO) enrichment analysis. The fortified GO terms uncovered specific natural functions associated with colorful subtypes of BRCA.

Sherafatian [20] participated in the implementation of the study where the miRNA expression dataset of patients with breast cancer from the TCGA database was used to develop predictive models, with which they identified miRNA biomarkers for diagnosis and the molecular subtyping of BRCA. For the purposes of this article, to gain empirical negative control miRNAs, they used an in silico approach and calculated the *p*-value for all miRNAs. Three tree-grounded algorithms (Random Forest, Rpart, and treebag) were employed to model the breast cancer status grounded on the regularized expression of the filtered miRNAs in a balanced training dataset. The significance of each point was determined during the construction of the bracket models, and the results were compared to identify miRNAs that constantly showed significance across all models. Among them, hsa-miR-139 and has-miR-96 were set up to be constantly significant in all three models. Also, the top ten miRNAs for classifying breast cancer from a normal solid towel using tree-grounded machine literacy algorithms included hsa-miR-139 and has-miR-96, as well as miRNAs 15, 183, 592, 20,125 b, 2, 21, 11, and 125b.1.

Table 1. Characteristics of selected studies.

Author, Reference and Year	Type of Cancer	Feature	Data Origin	*p*-Value, Fold Change	ML Method	Performance
Naorem, Muthaiyan and Venkatesan [18] (2019)	Triple-negative breast cancer	miRNA expression data	Gene Expression Omnibus (GEO)	*p*-value < 0.05 Fold change ≥ 1.0	* Naïve Bayes—NB * Sequential minimal optimization [SMO] * Random forest [RF])	NB = 96.8447% SMO = 96.966% RF = 96.4806%
Yu et al. [19]. (2020).	Breast cancer subtype classification	The differential expression analysis	TCGA database	p adjusted ≤ 0.01 Fold change ≥ 0.5	* NB * RF * Radia Vector Support Machines l" (SVM with radial basis kernel)	NB = 0.96 RF = 0.98 SMVRadial = 0.97 (Basal-like)
Sherafatian [20]. (2018).	Breast cancer subtype classification	miRNA expression data	TCGA database	*p*-value > 0.5	Three tree-based algorithms (RF, Rpart and treebag)	RF = 0.845 (Basal-like)
Qiu et al. [21]. (2020).	Breast cancer	The expression profile of mRNA and miRNA	Genomic Data Commons data portal (GDC1); the investigated sets were differential miRNAs in TCGA BRCA database cohort	*p*-value < 0.05	* SVM	Area under the curve (AUC) = 0.9633
Sarkar et al. [22]. (2021).	Breast cancer subtype classification	Next-generation sequencing (NGS)-based miRNA expression values	The Cancer Genome Atlas (TCGA)	*p*-value < 0.05	* Machine (SVM) * Artificial neural network (ANN) * K Nearest Neighbour (KNN) * Decision Tree (DT) * Random Forest (RF) * Naive Bayes(NB) and Discriminant analysis (DISCR)	SVM = 74.9094% ANN = 74.9094% KNN = 67.1014% DT = 64.4565% RF = 76.5761% NB = 70.5978% DISCR = 73.1884%
Andreini et al. [23]. (2022).	Breast cancer subtype classification	Uncover complex profiles of miRNA expression	BRCA dataset from The Cancer Genome Atlas (TCGA)	Fold change > 2	* SVM * Specialized multi-class Random Forest (RF)	SVM = 0.926 RF = 0.9886

Qiu et al. [21] constructed a dysregulated miRNA target network (DMTN) using miRNA–mRNA dyads that displayed significant dysregulation scores, indicating dysregulated nonsupervisory connections between miRNAs and target genes. The mRNA and miRNA expression biographies used in the study were attained from the genomic data commons (GDC1) data gate. All statistical and graphical analyses were conducted in the R terrain. The study linked 588 miRNAs and 3,146 genes associated with medicines, where the expression of miRNA genes showed significant associations with the response of cancer cells to anticancer medicines. The findings suggest that the expression situations of specific threat miRNAs and their neighboring monophasic genes could serve as pointers of cancer cell perceptivity to anticancer medicines. The experimenters proposed that with farther experimental and clinical confirmation, these miRNAs could potentially be used as biomarkers to guide the treatment of breast cancer cases.

Sarkar et al. [22] conducted a study using NGS data from breast cancer to identify the most important miRNA biomarkers. Selected miRNA biomarkers are strongly associated with multiple breast cancer subtypes. To do this, they used data from The Cancer Genome Atlas (TCGA) and proposed a two-step technique called feature selection methods embedded in machine learning, followed by survival analysis. In the first phase, to obtain this list of miRNAs, they selected the best among seven machine learning techniques (machine (SVM), artificial neural network (ANN), K-nearest neighbor (KNN), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB) and Discriminant Analysis (DISCR)), using the entire set of features; the best machine learning technique in this RF case was selected. In the second phase, based on the classification accuracy values, the most important features from each selection method are considered to make a set that provides miRNAs such as 8*, 7*, and even 1*. These analytical results confirmed the fact that the selected miRNAs are potential biomarkers for the diagnosis of cancer subtypes. Furthermore, GO (gene ontology) enrichment analysis also revealed selected miRNA biological, molecular, and cellular processes related to breast cancer. Overall, this study identified all 27 miRNAs as potential biomarkers and found that they are responsible for different subtypes of breast cancer.

Andreini et al. [23] proposed a new approach to use miRNA fractions as implicit biomarkers for the discovery of breast cancer. Their approach comprised two distinct stages. In the first stage, they employed two machine literacy models: a Support Vector Machine (SVM) to separate between healthy samples and cancer cells, and Random Forest (RF) to classify different subtypes of cancer. Through a comprehensive evaluation process using four-way cross-validation and grid hunt, they linked the most accurate model for each step, achieving a perfection of 0.9926 for the SVM and 0.9886 for the RF, along with the corresponding set of optimized hyperparameters.

In the alternate stage, they employed a significance-grounded point selection approach to determine the crucial features employed by the machine literacy models to make their prognostications. This two-step approach involved using devoted classifiers for tumor/health classification and subtype discovery. A notable advantage of their study was the application of two entirely independent datasets for training and testing. These datasets were generated from different sequencing machines and passed distinct bioinformatics processes for data preprocessing. Their results in the bracket of healthy and tumorous samples yielded perfect situations, in line with the best-published findings. In particular, none of the tumor samples were misclassified as healthy, pressing the robustness of their approach.

4. Discussion

Several machine learning methods exist to identify potential miRNA targets in BRCA. With this review, we intend to analyze the studies of different methods according to the level of bioinformatics technique, and ensure their stability according to the characteristics used.

Firstly, in all the analyzed studies, it was possible to assess the scope of the contribution of bioinformatics for the detection and treatment of breast cancer, and it was perceived over time that, together with other types of analyses (in vivo, in vitro), there is support in the fight against this disease.

The study by Naorem et al. [18] coincides with the research of Wang et al. [24] that addresses the topic of triple-negative breast cancer (TNBC). Regarding the role of fibroblasts, it also uses the Gene Expression Omnibus (GEO) dataset and Cancer Genome Atlas (TCGA) datasets: Random Forest (RF), Decision Tree (DT), and K-nearest neighbors (KNN) from the R package "caret", in contrast to the machine learning methods used, to create predictions for cancer-associated fibroblast (CAF) subtypes, resulting in an area under the curve of 0.921 with an RF model integrated into the experimental dataset. Significant genes and their differential expression were obtained with p-value ≤ 0.05 and fold change > 0.5, which was different from that applied by Naorem et al. [18], whereby it was $\geq$1.

On the other hand, research by Yu et al. [19] does not agree with Cascianelli et al. [25], as they focused on the robustness and portability of PAM50. The closest central classifier was developed using microarray data to identify five "intrinsic subtypes" of breast cancer. They also proposed a strategy called "average within class" (AWCA) that improves classification power with more than 90 matches and predictive power. They agreed that they used data from the Cancer Genome Atlas (TCGA); however, they did not use p-value and fold-change parameters for significant gene expression.

Sherafatian [20] used a tree-based and matching classifier Tabl et al. [26] that has a hierarchical machine learning system that assigns each node a relative class other, making the model tree-based. This study analyzed datasets available on the cBioPortal. Naive Bayes and Random Forest were used as classifiers. Finally, several biomarker genes are identified to predict the appropriate treatment for the patient. They obtained a 97.9% accuracy for the decision tree. For samples of live hormones based on the correlation coefficient between gene expressions, they used $p < 0.05$ that differs from that proposed by Sherafatian [20].

Qiu et al. [21] used a support vector machine classifier that coincides with the work of Yerukala Sathipati and Ho [27] that proposes a Support Vector Machine (SVM)-based classifier to classify patients with early and advanced breast cancer. SVM-BRC uses an optimal feature selection method, a legacy bio-target combinatorial genetic algorithm to identify a miRNA signature, which is a small set of informative miRNAs. They also agree to process the breast cancer miRNA expression profile data from the Cancer Genome Atlas. They show a significant association with the prognosis of the breast cancer patient using the p-value and fold change at different significant intervals. Regarding the performance of SVM-BRC, they achieved an accuracy of 83.16%, which was the best compared to other classifiers.

In relation to the work of Sarkar et al. [22], some coincidences are observed with the study by Denkiewicz et al. [28]. The first is that they used to validate seven well-known machine learning methods computationally: Logistic Regression (LR), Decision Tree (DT), Artificial Neural Network (RNA), also known as Multilayer Perceptron, Support Vector Machine with linear kernel (SVM), K-nearest neighbor (K-NN), Random Forest (RF), and finally Naive Bayes Classifier (NB). The second is that the expression and clinical data of miRNA-seq of invasive breast carcinoma (BRCA) were obtained from The Cancer Genome Atlas. It is even worth mentioning that the highest performance was obtained by the RF method, as well as in the study of Sarkar et al. [22]. Likewise, a p-value < 0.05 was used to determine significant miRNAs in a subtype of cancer.

Finally, the study by Andreini et al. [23] agrees with MotieGhader et al. [29], since both use a SVM (Support Vector Machine) classifier, although the first used it to distinguish healthy samples from cancerous ones and the second also included the support of 11 efficient and popular meta-heuristic algorithms to stratify breast cancer molecular subtypes using mRNA and micro-RNA expression data, as well as data downloaded from the NCBI Gene Expression Omnibus (GEO) database, which contains miRNA and mRNA expression profiles across five molecular subtypes of breast cancer, including Normal-Like, LUM A, LUM B, Basal, and ERBB2.

The results show that most selected miRNAs are related to the miR-190, miR-129, miR-34, and miR-181 families, where the p-value < 0.05 corresponds to miRNA families that

may play an important role in breast cancer, especially in studies of breast cancer subtypes. The accuracy obtained was 95%, and a five-fold cross-validation was used to assess the performance of the miRNA dataset.

In the analyzed studies, it is worth mentioning that we found different machine learning methods, some of which were mentioned in various articles. Because the miRNA BRCA classification performance was optimal, we also noticed that there are various types of data sources with open access that provide reliable information on the subject that support the development of future projects.

Finally, this work is not free of limitations, which depends on the interest in a particular aspect that needs to be improved to guide research toward the desired objective. On the other hand, the continuous technological and scientific innovations in bioinformatics bring us new essential and even more reliable advances in treating and diagnosing breast cancer.

5. Conclusions

After the effort to include the results analyzed in this study, it is important to mention that the research questions posed for this work were answered during analysis. Regarding most of the selected articles, it was observed that they use a p-value < 0.05 to determine the differential expression of miRNA. Finally, the automatic learning methods SVM (Support Vector Machine) and RF (Random Forest) were the most mentioned in the studies to carry out the classification; with this, they provide an important overview on deciding the automatic learning method to use in these types of investigations. There is no doubt that the advancement of technology will open many doors to develop better studies in bioinformatics, and breast cancer in particular.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app13148257/s1.

Author Contributions: Conceptualization, M.B.S.-C. and J.A.C.-R.; methodology, J.A.C.-R., D.M.C.-E. and M.Z.S.-L.; software, J.A.C.-R. and D.M.C.-E.; validation, M.B.S.-C., J.A.C.-R. and D.M.C.-E.; formal analysis, J.A.C.-R., D.M.C.-E. and M.Z.S.-L.; investigation, M.B.S.-C. J.A.C.-R., D.M.C.-E. and M.Z.S.-L.; resources, M.B.S.-C.; data curation, J.A.C.-R. and D.M.C.-E.; writing—original draft preparation, M.B.S.-C., J.A.C.-R., D.M.C.-E. and M.Z.S.-L.; writing—review and editing, M.B.S.-C., J.A.C.-R., D.M.C.-E. and M.Z.S.-L.; visualization, M.B.S.-C. and J.A.C.-R.; supervision, M.B.S.-C.; project administration, M.B.S.-C.; funding acquisition, M.B.S.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Consejo Nacional de Ciencia y Tecnologías (CONACYT), Ciencia Fronteras, number 319921.

Informed Consent Statement: Not applicable.

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

References

1. Hu, Z.-Z.; Huang, H.; Wu, C.H.; Jung, M.; Dritschilo, A.; Riegel, A.T.; Wellstein, A. Omics-Based Molecular Target and Biomarker Identification. *Methods Mol. Biol.* **2011**, *719*, 547–571. [CrossRef] [PubMed]
2. Breastcancer.Org.—Breast Cancer Information and Support. Breastcancer.Org. 2021. Available online: http://Breastcancer.org (accessed on 12 December 2022).
3. Loh, H.-Y.; Norman, B.P.; Lai, K.-S.; Rahman, N.M.A.N.A.; Alitheen, N.B.M.; Osman, M.A. The Regulatory Role of MicroRNAs in Breast Cancer. *Int. J. Mol. Sci.* **2019**, *20*, 4940. [CrossRef]
4. Pedrero, V.; Reynaldos-Grandón, K.; Ureta-Achurra, J.; Cortez-Pinto, E. Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Rev. Med. Chil.* **2021**, *149*, 248–254. [CrossRef]
5. Gibbons, J.D.; Pratt, J.W. *P*-values: Interpretation and Methodology. *Am. Stat.* **1975**, *29*, 20–25. [CrossRef]
6. Durbin, B.; Hardin, J.; Hawkins, D.; Rocke, D. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **2002**, *18* (Suppl. S1), S105–S110. [CrossRef] [PubMed]
7. Ritchie, M.E.; Silver, J.; Oshlack, A.; Holmes, M.; Diyagama, D.; Holloway, A.; Smyth, G.K. A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **2007**, *23*, 2700–2707. [CrossRef] [PubMed]
8. Rocke, D.M.; Durbin, B. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **2003**, *19*, 966–972. [CrossRef]

9. Sakr, S.; Elshawi, R.; Ahmed, A.M.; Qureshi, W.T.; Brawner, C.A.; Keteyian, S.J.; Blaha, M.J.; Al-Mallah, M.H. Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry ford exercIse testing (FIT) project. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 174. [CrossRef]
10. Başţanlar, Y.; Özuysal, M. Introduction to Machine Learning. *Methods Mol. Biol.* **2014**, *1107*, 105–128. [CrossRef]
11. Yu, Y.; Yang, L.; Liu, Z.; Zhu, C. Gene essentiality prediction based on fractal features and machine learning. *Mol. Biosyst.* **2017**, *13*, 577–584. [CrossRef]
12. Wong, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit.* **2015**, *48*, 2839–2846. [CrossRef]
13. Bendavid, A. Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Syst. Appl.* **2008**, *34*, 825–832. [CrossRef]
14. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process Manag.* **2009**, *45*, 427–437. [CrossRef]
15. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
16. Urrútia, G.; Bonfill, X. La declaración PRISMA: Un paso adelante en la mejora de las publicaciones de la Revista Española de Salud Pública. *Rev. Esp. Salud Publica* **2013**, *87*, 99–102. [CrossRef]
17. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann. Intern. Med.* **2009**, *151*, 264–269. [CrossRef]
18. Naorem, L.D.; Muthaiyan, M.; Venkatesan, A. Identification of Dysregulated MiRNAs in Triple Negative Breast Cancer: A Meta-analysis Approach. *J. Cell. Physiol.* **2019**, *234*, 11768–11779. [CrossRef]
19. Yu, Z.; Wang, Z.; Yu, X.; Zhang, Z. RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches. *Comput. Intell. Neurosci.* **2020**, *2020*, 4737969. [CrossRef]
20. Sherafatian, M. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene* **2018**, *677*, 111–118. [CrossRef]
21. Qiu, M.; Fu, Q.; Jiang, C.; Liu, D. Machine Learning Based Network Analysis Determined Clinically Relevant miRNAs in Breast Cancer. *Front. Genet.* **2020**, *11*, 615864. [CrossRef]
22. Sarkar, J.P.; Saha, I.; Sarkar, A.; Maulik, U. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. *Comput. Biol. Med.* **2021**, *131*, 104244. [CrossRef] [PubMed]
23. Andreini, P.; Bonechi, S.; Bianchini, M.; Geraci, F. MicroRNA signature for interpretable breast cancer classification with subtype clue. *J. Comput. Math. Data Sci.* **2022**, *3*, 100042. [CrossRef]
24. Wang, M.; Feng, R.; Chen, Z.; Shi, W.; Li, C.; Liu, H.; Wu, K.; Li, D.; Li, X. Identification of Cancer-Associated Fibroblast Subtype of Triple-Negative Breast Cancer. *J. Oncol.* **2022**, *2022*, 6452636. [CrossRef] [PubMed]
25. Cascianelli, S.; Molineris, I.; Isella, C.; Masseroli, M.; Medico, E. Machine Learning for RNA Sequencing-Based Intrinsic Subtyping of Breast Cancer. *Sci. Rep.* **2020**, *10*, 14071. [CrossRef]
26. Tabl, A.A.; Alkhateeb, A.; Elmaraghy, W.; Rueda, L.; Ngom, A. A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Front. Genet.* **2019**, *10*, 256. [CrossRef]
27. Sathipati, S.Y.; Ho, S.-Y. Identifying a miRNA signature for predicting the stage of breast cancer. *Sci. Rep.* **2018**, *8*, 16138. [CrossRef]
28. Denkiewicz, M.; Saha, I.; Rakshit, S.; Sarkar, J.P.; Plewczynski, D. Identification of Breast Cancer Subtype Specific MicroRNAs Using Survival Analysis to Find Their Role in Transcriptomic Regulation. *Front. Genet.* **2019**, *10*, 1047. [CrossRef]
29. MotieGhader, H.; Masoudi-Sobhanzadeh, Y.; Ashtiani, S.H.; Masoudi-Nejad, A. mRNA and microRNA selection for breast cancer molecular subtype stratification using meta-heuristic based algorithms. *Genomics* **2020**, *112*, 3207–3217. [CrossRef]