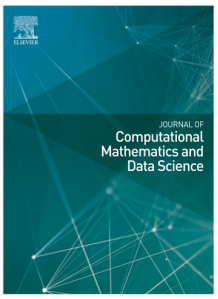
MicroRNA signature for interpretable breast cancer classification with subtype clue

Journal Pre-proof



MicroRNA signature for interpretable breast cancer classification with subtype clue

Paolo Andreini, Simone Bonechi, Monica Bianchini, Filippo Geraci

PII: S2772-4158(22)00011-6
DOI: https://doi.org/10.1016/j.jcmds.2022.100042
Reference: JCMDS 100042

To appear in: *Journal of Computational Mathematics and Data Science*

Received date : 10 March 2022
Revised date : 2 May 2022
Accepted date : 2 May 2022

Please cite this article as: P. Andreini, S. Bonechi, M. Bianchini et al., MicroRNA signature for interpretable breast cancer classification with subtype clue. *Journal of Computational Mathematics and Data Science* (2022), doi: https://doi.org/10.1016/j.jcmds.2022.100042.

Available online at www.sciencedirect.com
ScienceDirect
JCMDS
ELSEVIER
Journal of Computational Mathematics and Data Science 00 (2022) 1–13

MicroRNA Signature for Interpretable Breast Cancer Classification with Subtype Clue

Paolo Andreini[a,1], Simone Bonechi[b,1], Monica Bianchini[a], Filippo Geraci[c,*]

[a] *Department of Information Engineering and Mathematics, University of Siena, Siena*
[b] *Department of Social, Political and Cognitive Sciences, University of Siena, Siena*
[c] *Institute for Informatics and Telematics, CNR, Pisa*

Abstract

MicroRNAs (miRNAs) are short non-coding RNAs engaged in cellular regulation by suppressing genes at their post-transcriptional stage. Evidence of their involvement in breast cancer and the possibility of quantifying the their concentration in the blood has sparked the hope of using them as reliable, inexpensive and non-invasive biomarkers.

While differential expression analysis succeeded in identifying groups of disregulated miRNAs among tumor and healthy samples, its intrinsic dual nature makes it inadequate for cancer subtype detection. Using artificial intelligence or machine learning to uncover complex profiles of miRNA expression associated with different breast cancer subtypes has poorly been investigated and only few recent works have explored this possibility. However, the use of the same dataset both for training and testing leaves the issue of the robustness of these results still open.

In this paper, we propose a two-stage method that leverages on two ad-hoc classifiers for tumor/healthy classification and subtype identification. We assess our results using two completely independent datasets: TGCA for training and GSE68085 for testing. Experiments show that our strategy is extraordinarily effective especially for tumor/healthy classification, where we achieved an accuracy of 0.99. Yet, by means of a feature importance mechanism, our method is able to display which miRNAs lead to every single sample classification so as to enable a personalized medicine approach to therapy as well as the algorithm explainability required by the EU GDPR regulation and other similar legislations.

Keywords: miRNA biomarkers, breast cancer subtype, supervised classification, feature importance

1. Introduction

Cancer statistics collected from the US National Center for Health Statistics [1] show that, although smoothly, the incidence of new cancer cases in women is still increasing. Conversely, mortality is seamlessly decreasing since the beginning of the nineties. The discordance of these two trends is in part due to the astonishing improvements in the surgical and pharmaceutical treatment of cancer, but also to increasingly accurate screening techniques. Nevertheless, breast cancer still remains the most common tumor type (accounting for about 30% of the cases) in women and the second in terms of number of fatal events.

*Corresponding author
Email address: filippo.geraci@iit.cnr.it (Filippo Geraci)
[1] Equal contribution ordered alphabetically

Despite of a fairly consolidated therapeutic protocol, at least for the early stages, screening techniques still remain limited because breast cancer subtypes are not yet well characterized. The main classification for this pathology consists in four intrinsic (or molecular) subtypes based on combinations of the expression levels of three receptors: estrogen-receptor (ER), progesterone-receptor (PR), and HER2. An alternative classification system (PAM50 [2]), instead, uses a panel of 50 genes to discriminate among breast cancer subtypes. Although partially overlapping [3], the differences in the subtype classification induced by these two classification systems, reveals a large subtype heterogeneity [4] or suggests the presence of other (rare) subtypes not well-represented in the classification system [5], [6].

Mammography, a test that produces images of the breast by exposing it to low-energy X-rays, is the current main screening technique for this type of cancer. Since it has been introduced, this test has shown a dramatic advantage in terms of survival [7], even if it is not yet accurate enough in discriminating the cancer subtype [8]. Using expression-based profiling tests as an adjuvant to mammography would provide much richer details endowing screenings with subtype information as well as reducing the risk of overdiagnosis. However, a more detailed test output cannot be obtained with invasive procedures, such as biopsy, since they would be daunting, limiting the population coverage.

Due to a relatively simple procedure to quantify them in serum [9], using inexpensive RT-qPCR [10], microRNAs (short non-coding RNAs sequences), have been extensively investigated in recent years as potential biomarkers for cancer [11]. The established role of miRNAs in gene regulation [12] (in particular suppression) as well as the proved relationship between the expression of circulating miRNAs and cellular miRNAs [13] has confirmed the potential of these small RNAs as non-invasive biomarkers [14]. Yet, miRNAs' expression profiles of breast cancer subtypes have been shown to be different among each other [15, 16].

Previous differential expression studies have identified several miRNAs involved in specific cellular activities (e.g. cell proliferation) that are disregulated among healthy and tumor samples (see [17, 18]). However, the dual nature of differential expression fails to capture complex patterns of expression due to miRNA/miRNA or miRNA/mRNA interactions. Computational approaches based on machine learning address this problem modelling the identification of effective panels of biomarkers as the classic feature selection task so as to improve binary classification [19, 20]. The training process is made possible by the generous abundance of tumor and healthy samples in The Cancer Genome Atlas [21]. Results like that reported in [19] have shown that miRNAs have the potential to be adopted as reliable biomarkers as well as they can be used to identify the primary cancer site. This latter fact is important in view of the possibility of exploiting extracellular miRNAs, extracted from the serum, instead of tissue miRNAs, extracted by means of biopsy.

Narrowing to breast cancer, although classification methods like that in [20, 22] have shown an outstanding power in discriminating healthy from tumor samples, subtype identification still remains an open problem.

Very few recent works extend machine learning approaches to multi-class classification in the attempt to identify subtypes. In [23], e a tree based approach is proposed that identifies a very small set of miRNAs for classification. In spite of its desirable interpretability, the proposed method is unable to distinguish among luminal A and luminal B subtypes. In [24], a set of 5 miRNAs able to separate triple negative breast cancers from other subtypes is found. A pool of classifiers and feature selection methods is employed to identify the minimal set of miRNAs that maximize the validation accuracy in a k-fold validation process. With an in spirit similar approach, [25] identifies 27 miRNAs associated to Luminal A, HER-enriched and basal-like subtypes.

In this paper, we face the problem of subtype identification in addition to breast cancer classification. We fulfill this goal by means of two independent classifiers: a Support Vector Machine (SVM) [26] for the binary healthy/cancer classification and a specialized multi-class Random Forest (RF) [27] for subtypes. As opposed to previous results [19, 20, 23, 25], an important novelty of our work is that for training/validation and testing we used two completely independent datasets, sequenced with different technologies and processed with different bioinformatic pipelines. In particular, the raw counts of the BRCA dataset from *The Cancer Genome Atlas* (TCGA) were used for training and validation, while we performed tests by pre-processing fastq files from a cohort of 114 samples downloaded from the *Gene Expression Omnibus* (accession number GSE68085). This result is particularly important since it proves that our method is robust to changes in the sequencing machines, library preparation and reads pre-processing. A further important novelty of our

method is that we introduce a feature importance mechanism that scores each miRNA of a tested sample according to its contribution in the final classification decision. In addition to the utility of identifying general panels of miRNAs involved in a specific cancer subtype, this method is compliant to the *"right for an explanation"* principle introduced in the UE General Data Protection Regulation (GDPR). Moreover, the per-sample nature of our feature importance mechanism, not only provides the subtype indication but also explains on the basis of the expression of which miRNAs the sample belongs to that subtype. This, in turns, facilitates a personalized medicine approach to therapy.

2. Materials and Methods

Our approach aims to use miRNA fragments as possible biomarkers for breast cancer detection. We addressed this problem in two distinct phases. In the first phase, two machine learning models were trained to distinguish healthy from cancer samples and the cancer subtype (see Section 2.1). Secondly, a feature importance approach has been employed to identify the most relevant features used by the machine learning model to make its predictions (Section 2.2). The overall pipeline of the proposed method is shown in Figure 1.

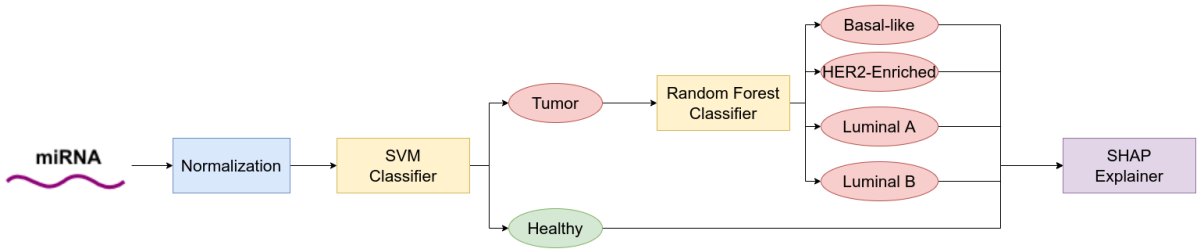


Figure 1: Overall pipeline of the proposed classification approach.

2.1. Classification

Our classification procedure consists of training two distinct classifiers which are subsequently used in cascade. In the first step, we employed a Support Vector Machine (SVM) classifier to recognize healthy from tumor samples. SVMs are a commonly used class of machine learning algorithms particularly indicated in the cases where the number of available training samples is rather limited. Moreover, they have been reported to work well when data are represented as vectors of continuous variables. Due to these characteristics, this method has largely been used in bioinformatics classification problems and, in particular, for cancer genomic classification or subtyping [28]. The SVM is based on the idea of classifying the data by finding the hyperplane that divides the samples with the maximum margin. Indeed, to improve data separability, they map the features into a higher dimensional space employing specific kernel functions. In the second step, the predicted cancer samples are classified into their specific sub-type using a Random Forest (RF) classifier.

RFs are a class of machine learning models consisting of a set of decision trees that are independently learned (bagging). The RF classifier determines the consensus outcome combining the predictions of the individual decision trees. Similarly to SVMs, this class of algorithms is suitable for small datasets with a limited number of labelled samples. In addition, RFs can easily handle high-dimensional feature spaces as, for example, that constituted by human genes [29].

The rationale behind our two-stage classification approach is that the two phases serve different purposes and can be optimized independently. The distinction between healthy and cancer samples can be considered a screening step, in which it is essential to avoid false negatives. Therefore, the classification model should be chosen considering not only its accuracy but also the model recall. For the classification of tumor sub-types, however, this requirement is less stringent, so that a different model with a different set of hyperparameters

can be chosen. In this work, to select the best set of hyperparameters for both the models, a grid-search approach using a predefined grid of values has been employed. Furthermore, an oversample strategy has been used on the training set to alleviate the dataset class imbalance during training. Moreover, we evaluated the performance of the models using the balanced class accuracy. The best model in both stages was finally selected by evaluating the performance with a 4-fold cross-validation. For the healthy/cancer classifier the selection takes into account not only the balanced accuracy but also the recall, in order to obtain a model capable of minimizing false negatives.

2.2. Feature Importance

Being able to understand and interpret the results of machine learning tools is essential for developing systems that are reliable and usable in practice. There are many different ways to increase the understanding of a machine learning model decision, and the feature importance is one of the most useful. Indeed, the feature importance allows to estimate how much each feature contributes to the prediction of the model. In other words, a feature importance method provides a better understanding on which features are having the greatest impact on the decisions made by the model. In this work, we used the SHAP approach [30] that permits to explain the prediction of a single example by computing the contribution of each feature to the model output. In particular, the SHAP method computes the Shapley values from coalitional game theory. In a cooperative multiplayer game, Shapley values aim to quantify each player's contribution to the game or, in our case, to indicate how to fairly distribute the "payout" (which correspond to the prediction) among the features. In SHAP the key idea is, given a specific sample, to calculate the Shapley values associated with each feature. Each Shapley value provides an estimate of the impact of the corresponding feature to the prediction. SHAP uses the kernel Shap method to efficiently calculate the Shapley values. The method can be used to estimate the importance of each feature for each individual example. To give an overall evaluation of the importance of the different features, we averaged the values obtained for each sample in the test set.

3. Experiments

3.1. Training dataset

Data-driven machine learning approaches for biomarker identification require the availability of a large collection of annotated samples for training. As most of the other studies (see e.g. [19, 20, 23, 25, 31]), we leveraged on the generous collection of breast cancer samples available in The Cancer Genome Atlas.

We used the Firebrowse service (http://firebrowse.org/) from the Broad Institute to download the breast cancer data while we derived clinical information from supplementary data of [32]. The Firebrowse repository consists of all 1098 cases from TCGA at the date of 2016/01/28, including 20 samples for which the miRNA-seq data is not available. Each sample is provided as the count of reads aligned to the primary transcript of 503 miRNAs.

One of the benefits of downloading datasets from Firebrowse is that this service makes a big effort to mitigate undesired effects due to systematic biases. Consequently, we did not need to apply any batch correction, but we simply converted raw counts into counts per millions by means of the `cpm()` function of the edgeR [33] R package. As for filtering, we removed miRNAs that are marked as *dead* in the current version of miRBase (http://www.mirbase.org/) version 22 (namely: hsa-mir-1254, hsa-mir-3653, hsa-mir-3687, hsa-mir-3607, hsa-mir-3647, hsa-mir-3676, hsa-mir-1274b) and miRNAs with expression count equals to 0 in at least the 60% of the samples.

In order to limit possible biases due to variability or sample management, we filtered-out males (13 samples) and FFPE (Formalin-Fixed Paraffin-Embedded) samples (12 cases). A consistent fraction of the remaining samples, however, is not provided with enough clinical information to unambiguously derive the subtype and, thus, we could not include it in our study (492 cases).

We labelled the HER2-enriched and basal-like subtypes in accordance to Table 1 of [34] while, in the absence of information on the KI67 index and tumor grade, we extracted the luminal A and luminal B

subtypes from the PAM50 classification, as provided in the supplementary data of [35]. We also removed 26 samples where the molecular subtype classification conflicted with the PAM50 one.

In addition to cancerous samples (431 cases), the dataset includes 104 healthy samples. The subdivision of the samples by subtype is reported in Table 1.

Table 1: Training and test dataset description

Cancer subtype	**TGCA-BRCA**	**GSE68085**
Luminal A (LA)	215	62
Luminal B (LB)	96	-
HER2-Enriched (HER2)	56	-
Basal-Like (BL)	64	29
Healthy	104	11

3.2. Test dataset

In order to measure the degree of dependence of the classification outcome from the sequencing and processing protocols used for data collection, we could not use the k-fold cross validation on the TCGA data, but we needed a different source of testing samples. We thus thoroughly scan the Sequence Read Archive (SRA) and the Gene Expression Omnibus (GEO) database to find suitable datasets of miRNA profiles of breast cancer samples with available raw reads and annotated subtypes.

We identified only one collection (see Table 1 for details) consisting in a cohort of 114 samples (GEO GSE68085), 11 of which belonging to apparently healthy woman subjected to reduction mammoplasty, sequenced with Illumina Genome Analyzer IIx and described in details in [36]. We downloaded raw data from NCBI SRA and used *EZcount* [37] for pre-processing and miRNA counting. As for annotations, we used miRBase version 22 [38], remapping the nomenclature to make it consistent with that of the TCGA. We had to exclude the 23 samples belonging to HER2+ and Luminal B since the samples of these types are mixed in the original classification.

As for the training set, we normalized raw counts applying the count per million `cpm()` function from EdgeR package.

3.3. Classification accuracy

Following the procedure described in Section 2.1, we performed classification in two independent steps: the first to separate healthy from tumor samples and the second to identify the cancer sub-type. For each step, we selected the most accurate model, either SVM or RF, and the corresponding set of hyperparameters by using a 4-fold cross-validation with a grid search approach. For the perfomance evaluation, in order to avoid misinterpretation of the classification results, we used balanced accuracy [39]. Indeed balanced accuracy — defined as the average of the recall value obtained on each class — is especially useful when the classes are imbalanced, for instance in anomaly detection problems. Table 2 shows the best results obtained in cross-validation from the SVM and RF models trained on the TCGA-BRCA dataset, comparing different sets of hyperparameters, to recognize healthy from tumor samples.

Table 2: Cross-validation results on the TCGA-BRCA dataset for the healthy vs cancer classification. Average balanced accuracy and standard deviation are calculated based on the 4 validation folds

Model	**Mean Balanced Accuracy**	**Standard Deviation**
SVM	**0.9926**	**± 0.017**
RF	0.9886	± 0.035

The hyperparameter selection procedure led to the choice of the SVM model[2]. The model was then evaluated on the test set (GSE68085) and the results are reported in Table 3, while the confusion matrix is shown in Table 4.

Table 3: Balanced Accuracy and Accuracy obtained by the SVM for the classification of healthy and tumor samples in the test set (GSE68085)

Model	Balanced Accuracy	Accuracy
SVM	0.9545	0.9901

Table 4: Confusion matrix obtained by the SVM for the classification of healthy and tumor samples in the test set (GSE68085)

	Healthy	Tumor
Pred Healthy	10	0
Pred Tumor	1	91

As for the tumor/healthy classifier, a grid search approach with 4-fold cross-validation was performed also to select the best model for tumor subtype identification. The results of the hyperparameter selection procedure for this second classifier are shown in Table 5.

Table 5: Cross-validation results on the TCGA-BRCA dataset for the classification of tumor subtypes. Average balanced accuracy and standard deviation are calculated based on the 4 validation folds

Model	Mean Balanced Accuracy	Standard Deviation
SVM	0.7442	± 0.119
RF	**0.7800**	**± 0.093**

In this case, the best model selected for the cancer subtype classification is a RF[3]. It is worth noting that our two-stage classification approach allowed to specialize a different model for each of the stages, thus providing a great flexibility. The selected models, the SVM for the healthy-tumor classification and the RF for the tumor subtypes identification, were used in cascade on the test set (GSE68085). In particular, the first model identifies cancer samples while the second classifier identifies their subtype. In Table 6, we show the accuracy and the balanced accuracy obtained at the end of this combined classification pipeline. Finally, in Table 7, the overall confusion matrix is reported.

The results indicate that the proposed approach allows to recognize cancer samples with excellent accuracy and to perform a precise classification of the subtypes. Furthermore, false negatives (i.e. tumor samples predicted as healthy) are completely avoided in the test set (see the confusion matrix in Table 4), which is particularly important to avoid delayed diagnoses with consequent health injury for the patient. All the described experiments were carried out on a computer with a 3.50GHz Intel(R) Core(TM) i9-10920X CPU and 128 GB of RAM, using the Python library scikit-learn [40]. In Table 8 we reported the execution time needed to train the SVM and RF models, and to predict the class and extract the feature importance for a single sample.

3.4. Biological significance

As a direct consequence of the training phase, our classifier learned several co-expression patterns of miRNAs representative of each category. In contrast to differential expression, where a sample is labelled as belonging to a given class only if certain known miRNAs follow a precise pattern, in our case, different profiles and different miRNAs can lead to the same classification. Thanks to the feature importance mechanism, it is possible to identify a posteriori the miRNAs involved in a specific classification process and, consequently, to derive their expression pattern.

In this section, we show the interaction network of miRNAs more often involved in the assignment of samples to their corresponding class. For each sample of the test set that was correctly classified, we run

[2]With the following hyperparameters: linear kernel and C = 0.001.

[3]With the following hyperparameters: 50 trees, max features 30, min sample leaf 1, mean sample split 8 and gini impurity as the split criterion.

Table 6: Balanced Accuracy and Accuracy obtained by the combined classifier (SVM+RF) on the test set (GSE68085)

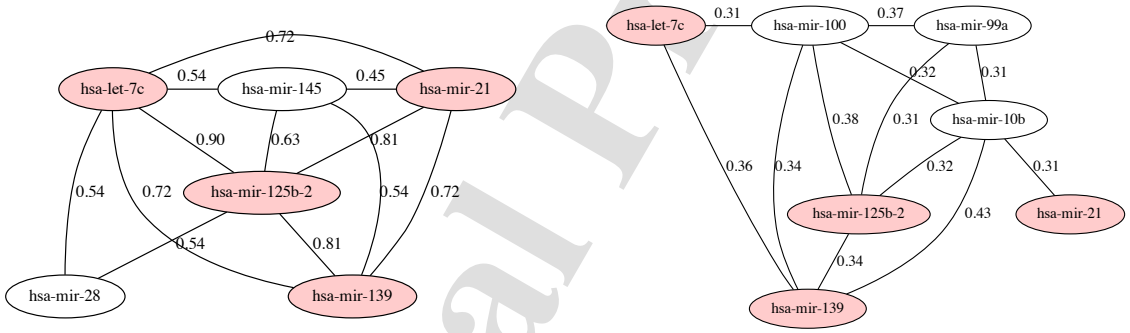
Model	Balanced Accuracy	Accuracy
Combined (SVM+RF)	0.8108	0.7450

Table 7: Confusion matrix obtained by the combined classifier (SVM+RF) on the test set (GSE68085)

	Basal-like	HER2-Enriched	Luminal B	Luminal A	Healthy
Pred Basal-like	25	0	0	4	0
Pred HER2-Enriched	2	0	0	1	0
Pred Luminal B	0	0	0	16	0
Pred Luminal A	2	0	0	41	1
Pred Healthy	0	0	0	0	10

the feature importance mechanism and extracted the list of the 10 miRNAs with the highest score. Then, we built a graph where each node corresponds to a miRNA and edges represent the relationship of co-appearance in the list of relevant miRNAs of the same tested sample. We score the relationship strength using the probability of co-appearance among all samples.

To thin out the graph and narrow the network to the most important miRNA/miRNA relationships, we applied three filters in cascade. Firstly, we removed edges where none of the involved miRNAs appeared in the list of relevant miRNAs for at least half of the tested samples. Then, we trimmed edges with probability score lower than 0.3. Finally, unconnected nodes were withdrawn from the graph.



(a) Healthy samples

(b) Tumor samples

Figure 2: Network of co-occurrence of highly important miRNAs for at least half of the classified samples. In red, miRNAs equally important both for the healthy and tumor classes.

In Figure 2 the networks of miRNAs for healthy (a) and tumor (b) samples are depicted. As the figure reveals, the two networks share a fraction of the nodes (highlighted in red) but each of them has specific ones. Common nodes tend to correspond to dysregulated miRNAs. In fact, analyzing them using edgeR [33], we found that they are all remarkably differentially expressed with very low p-values and absolute log-fold change higher than 2. The other nodes, instead, do not show the same tendency. In particular, the absolute log-fold change in some cases is lower than one. This derives from the fact that the expression is somehow constant within a class but fluctuates within the other.

A closer inspection of the relevant miRNAs reported in Figure 2 reveals that they belong to known tumor pathways. The function of hsa-mir-125b-2, for example, has repeatedly been investigated in relation to cancer. As reported in a recent review [41], it can either acts as an oncogene or a suppressor according to the specific cancer type. In the case of breast cancer, in [42], its molecular behaviour in conjunction with a

Table 8: Execution time of each phase of the proposed multistage classification approach

Step	Seconds
SVM training: Healthy VS Tumor	~ 0.0656
RF training: Tumor subtypes	~ 0.3438
SVM prediction: Healthy VS Tumor	~ 0.0002 per sample
RF prediction: Tumor subtypes	~ 0.0027 per sample
SVM SHAP Feature Importance: Healthy VS Tumor	~ 32.022 per sample
RF SHAP Feature Importance: Tumor subtypes	~ 22.993 per sample

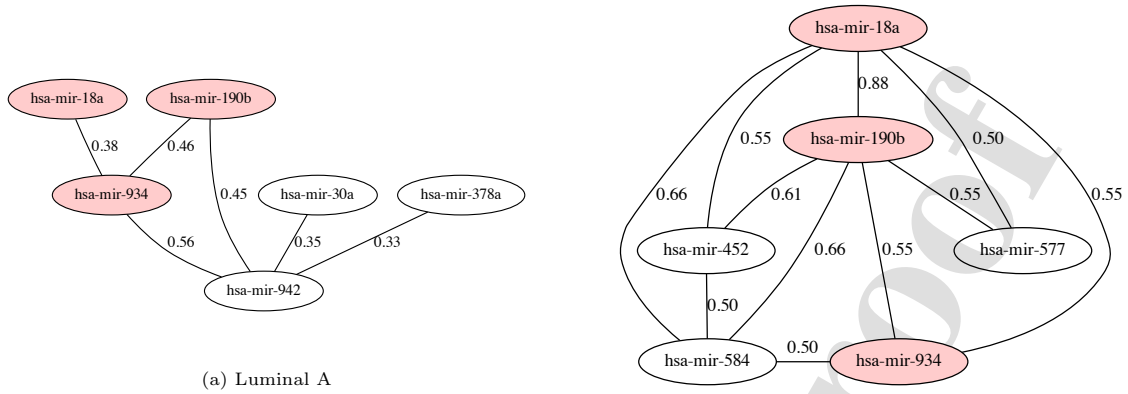
SNP in the binding site of BMPR1B is described, which miR-125b exploits to alter the transcription level. Increasing levels of BMPR1B (thus low levels of hsa-mir-125b-2) are associated to higher risk. Counts in the GSE68085 dataset confirms this hypothesis showing higher concentration of hsa-mir-125b-2 in healthy samples. Hsa-let-7c has attracted the attention as a therapeutic target, due to its tumor suppression properties for several cancer types. In [43], a preliminary analysis of its regulation in breast cancer was reported, even if its role at the pathway level still needs to be investigated. Hsa-mir-21 is known to be correlated to advanced clinical stages of breast cancer [44], being associated to the insurgence of lymph node metastasis [45]. The potential of hsa-mir-139 as a diagnostic biomarker for several types of cancer is reviewed in [46]. In breast cancer, [47] and others describe its suppression function of the proliferation and migration of tumor cells by targeting RAB1A.

Although tumor specific miRNAs (hsa-mir-100, hsa-mir-99a and hsa-mir-10b) have already been investigated in cancer, their involvement is not as ubiquitous as that of miRNAs in common with the healthy class. In particular, recurrent co-expression patterns of hsa-mir-100 with hsa-mir-99b-3p have been found in oral carcinoma [48] and in the resistance mechanisms of colorectal cancer [49]. Similarly, when in combination with hsa-mir-491, hsa-mir-99a modulates drug-resistance in gastric cancer [50]. For both miRNAs, however, little is known about the interaction with breast cancer. The literature on hsa-mir-10b role in cancer focuses only on gastric cancer [51], even if its role in breast cancer has been proved with in-silico analyses in [52]. Claiming a causal relationship of hsa-mir-100, hsa-mir-99a and hsa-mir-10b with breast cancer or understanding the molecular mechanisms that connect patterns of them with this pathology is beyond the scope of this paper. However, we can hypothesize that their central role in healthy/tumor classification is a clue of their role in breast cancer even if their missing differential expression is responsible for the limited support in the scientific literature.

On the other hand, although the property of being differentially expressed of the miRNAs in common between healthy and tumor classes would appear appealing, relegating specific miRNAs to a marginal role, we notice that these miRNAs activate general cancer pathways and are not specific for breast cancer (in particular hsa-mir-125b-2, Hsa-let-7c, and hsa-mir-139). Consequently, their quantification in blood can help in determining the presence of a tumor but is unable to discriminate its type.

Figure 3 shows the interaction networks emerging from the identification of Luminal A (a) and basal-like (b) subtypes. The absence of Luminal B and HER2-enriched types in the GSE68085 dataset prevented us from showing similar networks for these two cancer types.

A first observation about miRNAs involved in the subtype identification is that, according to differential analysis made using edgeR, they are all dysregulated among healthy and tumor samples with very small p-values and, except for hsa-mir-584, with absolute log-fold change higher (or much higher) than 1. Interestingly, all but hsa-mir-378a and hsa-mir-584 have already been reported in tumor pathways of other forms of cancer (in particular cell proliferation), even if limited information of their role in BRCA is available the literature. The two not cancer-specific miRNAs, namely hsa-mir-378a and hsa-mir-584, are both involved in angiogenesis (see [53] and [54]). Even if understanding their role is beyond the scope of this paper, their presence in the networks of Figure 3 would suggest that their role could be in support of the growth of the tumor lesions. In particular, belonging them to two different networks shown in Figure 3, we could also hypothesize that Luminal A and basal-like subtypes use different pathways for vascularization. Effectively, in [55], hsa-miR-584 was found to be down-regulating TGF-β in BC cells. PHACTR1, an actin-binding



(a) Luminal A

(b) Basal-like

Figure 3: Network of co-occurrence of miRNAs highly important for at least half of the classified samples. In red, miRNAs equally important for both the tested cancer subtypes.

protein, is also regulated by hsa-miR-584. Overexpression of hsa-miR-584 and knockdown of PHACTR1 resulted in a drastic rearrangement of the actin cytoskeleton and in a loss of TGF-β-induced cell migration. The drastic reorganization of the actin cytoskeleton is important in axonal guidance signalling, playing a role in tumour cell migration, tumour cell survival and tumour angiogenesis.

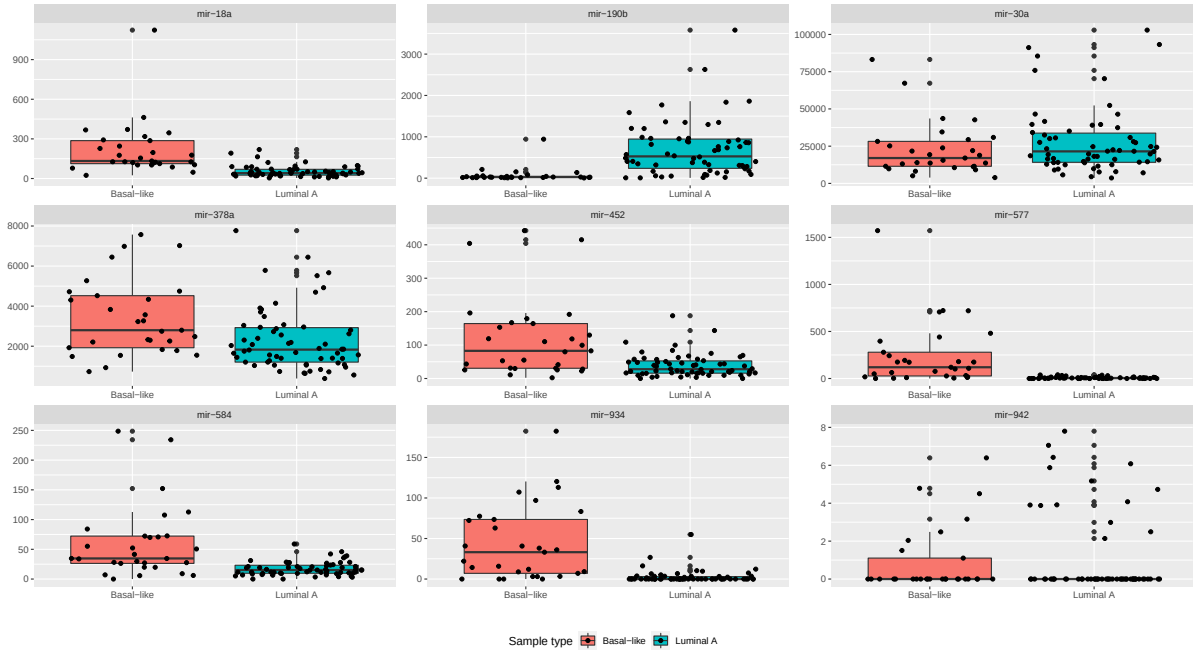


Figure 4: Comparison of the distributions of the expression levels for Luminal A and basal-like tumors.

A direct comparison of the distributions of the expression level among Luminal A and basal-like tumors (see Figure 4) shows that common miRNAs (in particular hsa-mir-18a and hsa-mir-190b) tend to have

differential patterns, while subtype specific patterns for the other miRNAs are not evident. This result is in line with our hypothesis that the mutual influence of miRNAs in the cellular regulation cannot be fully captured with differential expression but requires the inspection of more complex patterns.

4. Discussion

In this paper, we addressed the problem of using complex miRNA signatures as a tool to discriminate breast cancer samples from healthy ones. In addition, we investigated the possibility of deriving intrinsic subtypes as well.

A vast literature on this subject has identified several miRNAs whose differential concentration interferes with oncogenes and/or suppressors and, thus, correlates with breast cancer. In several cases, however, the same miRNA is not specific but promotes/suppresses general pathways in common among different cancer types. As a result, although using these miRNAs as biomarkers could be useful to determine the presence of a tumor, they leave open the problem of identifying its type.

Recent studies have argued that, despite useful, differential expression analysis is not enough powerful for tumor classification since it does not take co-regulation into consideration. Machine learning approaches such as [19, 20, 22] (and many others) have shown how the ability of supervised classifiers to learn complex patterns can improve the classification accuracy. On the basis of these results, few authors have faced the challenge of extending classification to cancer subtype. Partial results (where not all breast cancer subtypes are considered) have been described in [23, 24, 25].

The main limit we impute to these works is that, due to the reduced availability of training samples, the classification accuracy was evaluated on the same dataset used for training by means of the k-fold cross-validation technique. However, as we know that miRNA quantification suffers from several biases, due to sequencing and raw data processing, using the same dataset for training and testing could fail to prove the robustness of the method. What we believe to be one of the major merits of our work is the use of two completely independent datasets for training and test. The datasets have been produced with different sequencing machines and also pre-processing of raw data has been done with different bioinformatic pipelines. In the healthy/tumor classification, we achieve an accuracy in line with the best published results. Interestingly, no tumor samples have been classified as healthy. In our view, this is particularly important considering the negative implications of false negatives that could cause delayed diagnoses.

Another important aspect of our work is that the explanation of the classification results is not based on a general model but it is tailored on the specific sample under consideration. Methods like that described in [23], which are also explainable, define general rules that do not take peculiarities of the individual into consideration. Despite defining general rules fulfills the goal of inferring relationships among miRNAs and cancer, ex-ante explanation is not suitable for personalized medicine. Our approach, instead, pursues both goals. Results depicted in Figure 2 and 3 prove that running our classifier on a large cohort of samples allows to derive general knowledge, while the per-sample feature importance mechanism can be used for personalized medicine purposes.

5. Conclusions

MiRNAs are short fragments of non-coding RNA that influence cellular activity by means of a suppression mechanism. Due to their ubiquitous involvement in almost all molecular functions, miRNAs dysregulation have been investigated in conjunction with several diseases including cancer. Current research in bioinformatics is devoted to extend the standard differential expression model thanks to machine learning approaches that are able to infer complex expression patterns forecasting a disease. In this paper, we followed a similar approach with reference to breast cancer. Unlike the majority of other approaches, however, we do not limit ourselves to just recognizing the two health/disease classes, but, through a two-step classification, we also identify the molecular subtype of the cancer. In addition, we employ a feature importance method that enables personalized identification of the miRNAs responsible for a particular classification.

MiRNA quantification is known to be greatly influenced by sequencing protocols and bioinformatic pipeline for aligning reads and perform counting. In this regard, the current practice (driven by the lack

of labelled examples) of using partitions of the same dataset for training and test leaves some doubts on the actual robustness of the proposed methods. In the present work, we overcome this limit by using two independent datasets for training and test. In particular we perform training on the BRCA collection of TCGA while tests are made on the GSE68085 dataset form Gene Expression Omnibus. Experimental results have demonstrated the efficacy of our two step classification. The healthy/tumor step is as reliable as displaying an accuracy of 0.99 with only one healthy sample misclassified as tumor and no errors in the opposite direction. Overall, classification and subtype detection reach a balanced accuracy of 0.81.

References

[1] R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2019, CA: a cancer journal for clinicians.

[2] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, Journal of clinical oncology 27 (8) (2009) 1160.

[3] C. Sweeney, P. S. Bernard, R. E. Factor, M. L. Kwan, L. A. Habel, C. P. Quesenberry, K. Shakespear, E. K. Weltzien, I. J. Stijleman, C. A. Davis, et al., Intrinsic subtypes from pam50 gene expression assay in a population-based breast cancer cohort: differences by age, race, and tumor characteristics, Cancer Epidemiology and Prevention Biomarkers 23 (5) (2014) 714–724.

[4] J. Holm, L. Eriksson, A. Ploner, M. Eriksson, M. Rantalainen, J. Li, P. Hall, K. Czene, Assessment of breast cancer risk factors reveals subtype heterogeneity, Cancer research 77 (13) (2017) 3708–3717.

[5] M. V. Dieci, E. Orvieto, M. Dominici, P. Conte, V. Guarneri, Rare breast cancer subtypes: histological, molecular, and clinical peculiarities, The oncologist 19 (8) (2014) 805–813.

[6] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups, Nature 486 (7403) (2012) 346.

[7] E. R. Myers, P. Moorman, J. M. Gierisch, L. J. Havrilesky, L. J. Grimm, S. Ghate, B. Davidson, R. C. Mongtomery, M. J. Crowley, D. C. McCrory, et al., Benefits and harms of breast cancer screening: a systematic review, Jama 314 (15) (2015) 1615–1634.

[8] G. Farshid, D. Walters, Molecular subtypes of screen-detected breast cancer, Breast cancer research and treatment 172 (1) (2018) 191–199.

[9] I. S. Sourvinou, A. Markou, E. S. Lianidou, Quantification of circulating mirnas in plasma: effect of preanalytical and analytical parameters on their isolation and stability, The Journal of Molecular Diagnostics 15 (6) (2013) 827–834.

[10] M. G. Kok, A. Halliani, P. D. Moerland, J. C. Meijers, E. E. Creemers, S.-J. Pinto-Sietsma, Normalization panels for the reliable quantification of circulating micrornas by rt-qpcr, The FASEB Journal 29 (9) (2015) 3853–3862.

[11] J. Lu, G. Getz, E. Miska, E. Saavedra, J. Lamb, D. Peck, A. Cordero, B. Ebert, R. Mak, A. Ferrando, J. Downing, T. Jacks, H. Horvitz, T. Golub, MicroRNA expression profiles classify human cancers, Nature 435 (7043) (2005) 834–838.

[12] R.Shalgi, D. Lieber, M. Oren, Y. Pilpel, Global and local architecture of the mammalian microRNA-transcription factor regulatory network, PLoS Computational Biology 3 (7) (2007) e131.

[13] R. Duttagupta, R. Jiang, J. Gollub, R. C. Getts, K. W. Jones, Impact of cellular mirnas on circulating mirna biomarker signatures, PloS One 6 (6) (2011) e20769.

[14] S. Gilad, E. Meiri, Y. Yogev, S. Benjamin, D. Lebanony, N. Yerushalmi, H. Benjamin, M. Kushnir, H. Cholakh, N. Melamed, et al., Serum micrornas are promising novel biomarkers, PloS One 3 (9) (2008) e3148.

[15] X. Dai, A. Chen, Z. Bai, Integrative investigation on breast cancer in er, pr and her2-defined subgroups using mrna and mirna expression profiling, Scientific reports 4 (1) (2014) 1–10.

[16] S. Kurozumi, Y. Yamaguchi, M. Kurosumi, M. Ohira, H. Matsumoto, J. Horiguchi, Recent trends in microrna research into breast cancer with particular focus on the associations between micrornas and intrinsic subtypes, Journal of human genetics 62 (1) (2017) 15–24.

[17] C. A. Andorfer, B. M. Necela, E. A. Thompson, E. A. Perez, Microrna signatures: clinical biomarkers for the diagnosis and treatment of breast cancer, Trends in molecular medicine 17 (6) (2011) 313–319.

[18] M. Adhami, A. A. Haghdoost, B. Sadeghi, R. Malekpour Afshar, Candidate mirnas in human breast cancer biomarkers: a systematic review, Breast Cancer 25 (2) (2018) 198–205.

[19] S. S. Bhowmick, I. Saha, D. Bhattacharjee, L. M. Genovese, F. Geraci, Genome-wide analysis of ngs data to compile cancer-specific panels of mirna biomarkers, PloS one 13 (7) (2018) e0200353.

[20] O. Rehman, H. Zhuang, A. Muhamed Ali, A. Ibrahim, Z. Li, Validation of mirnas as breast cancer biomarkers with a machine learning approach, Cancers 11 (3) (2019) 431.

[21] K. Tomczak, P. Czerwińska, M. Wiznerowicz, The cancer genome atlas (tcga): an immeasurable source of knowledge, Contemporary oncology 19 (1A) (2015) A68.

[22] I. Saha, S. S. Bhowmick, F. Geraci, M. Pellegrini, D. Bhattacharjee, U. Maulik, D. Plewczynski, Analysis of next-generation sequencing data of mirna for the prediction of breast cancer, in: International Conference on Swarm, Evolutionary, and Memetic Computing, Springer, 2015, pp. 116–127.

[23] M. Sherafatian, Tree-based machine learning algorithms identified minimal set of mirna biomarkers for breast cancer diagnosis and molecular subtyping, Gene 677 (2018) 111–118.

[24] A. Lopez-Rincon, L. Mendoza-Maldonado, M. Martinez-Archundia, A. Schönhuth, A. D. Kraneveld, J. Garssen, A. Tonda, Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification, Cancers 12 (7) (2020) 1785.
[25] J. P. Sarkar, I. Saha, A. Sarkar, U. Maulik, Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific mirna biomarkers, Computers in Biology and Medicine 131 (2021) 104244.
[26] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.
[27] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.
[28] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (svm) learning in cancer genomics, Cancer genomics & proteomics 15 (1) (2018) 41–51.
[29] Y. Qi, Random forest for bioinformatics, in: Ensemble machine learning, Springer, 2012, pp. 307–323.
[30] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.
URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
[31] G. Tamilmani, V. B. Devi, T. Sujithra, F. H. Shajin, P. Rajesh, Cancer mirna biomarker classification based on improved generative adversarial network optimized with mayfly optimization algorithm, Biomedical Signal Processing and Control 75 (2022) 103545.
[32] D. Koboldt, R. Fulton, M. McLellan, H. Schmidt, J. Kalicki-Veizer, J. McMichael, L. Fulton, D. Dooling, L. Ding, E. Mardis, et al., Comprehensive molecular portraits of human breast tumours, Nature 490 (7418) (2012) 61–70.
[33] M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2010) 139–140.
[34] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, B. Shi, Breast cancer intrinsic subtype classification, clinical use and future trends, American journal of cancer research 5 (10) (2015) 2929.
[35] D. Netanely, A. Avraham, A. Ben-Baruch, E. Evron, R. Shamir, Expression and methylation patterns partition luminal-a breast tumors into distinct prognostic subgroups, Breast Cancer Research 18 (1) (2016) 1–16.
[36] P. Krishnan, S. Ghosh, B. Wang, D. Li, A. Narasimhan, R. Berendt, K. Graham, J. R. Mackey, O. Kovalchuk, S. Damaraju, Next generation sequencing profiling identifies mir-574-3p and mir-660-5p as potential novel prognostic markers for breast cancer, BMC genomics 16 (1) (2015) 1–17.
[37] F. Geraci, G. Manzini, Ezcount: An all-in-one software for microrna expression quantification from ngs sequencing data, Computers in Biology and Medicine 133 (2021) 104352.
[38] A. Kozomara, S. Griffiths-Jones, mirbase: annotating high confidence micrornas using deep sequencing data, Nucleic acids research 42 (D1) (2013) D68–D73.
[39] K. H. Brodersen, C. S. Ong, K. E. Stephan, J. M. Buhmann, The balanced accuracy and its posterior distribution, in: 2010 20th international conference on pattern recognition, IEEE, 2010, pp. 3121–3124.
[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[41] B. Peng, P. Y. Theng, M. T. Le, Essential functions of mir-125b in cancer, Cell Proliferation 54 (2) (2021) e12913.
[42] P. Sætrom, J. Biesinger, S. M. Li, D. Smith, L. F. Thomas, K. Majzoub, G. E. Rivas, J. Alluin, J. J. Rossi, T. G. Krontiris, et al., A risk variant in an mir-125b binding site in bmpr1b is associated with breast cancer pathogenesis, Cancer research 69 (18) (2009) 7459–7465.
[43] E. Bozgeyik, Bioinformatic analysis and in vitro validation of let-7b and let-7c in breast cancer, Computational Biology and Chemistry 84 (2020) 107191.
[44] A. Amirfallah, H. Knutsdottir, A. Arason, B. Hilmarsdottir, O. T. Johannsson, B. A. Agnarsson, R. B. Barkardottir, I. Reynisdottir, Hsa-mir-21-3p associates with breast cancer patient survival and targets genes in tumor suppressive pathways, PloS one 16 (11) (2021) e0260327.
[45] L.-X. Yan, X.-F. Huang, Q. Shao, M.-Y. Huang, L. Deng, Q.-L. Wu, Y.-X. Zeng, J.-Y. Shao, Microrna mir-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis, Rna 14 (11) (2008) 2348–2360.
[46] N. Khalili, M. Nouri-Vaskeh, Z. H. Segherlou, A. Baghbanzadeh, M. Halimi, H. Rezaee, B. Baradaran, Diagnostic, prognostic, and therapeutic significance of mir-139-5p in cancers, Life sciences 256 (2020) 117865.
[47] W. Zhang, J. Xu, K. Wang, X. Tang, J. He, mir-139-3p suppresses the invasion and migration properties of breast cancer cells by targeting rab1a, Oncology reports 42 (5) (2019) 1699–1708.
[48] M. Jakob, L. M. Mattes, S. Küffer, K. Unger, J. Hess, M. Bertlich, F. Haubner, F. Ihler, M. Canis, B. G. Weiss, et al., Microrna expression patterns in oral squamous cell carcinoma: hsa-mir-99b-3p and hsa-mir-100-5p as novel prognostic markers for oral cancer, Head & Neck 41 (10) (2019) 3499–3515.
[49] X.-D. Yang, X.-H. Xu, S.-Y. Zhang, Y. Wu, C.-G. Xing, G. Ru, H.-T. Xu, J.-P. Cao, Role of mir-100 in the radioresistance of colorectal cancer cells, American journal of cancer research 5 (2) (2015) 545.
[50] Y. Zhang, W. Xu, P. Ni, A. Li, J. Zhou, S. Xu, Mir-99a and mir-491 regulate cisplatin resistance in human gastric cancer cells by targeting capns1, International journal of biological sciences 12 (12) (2016) 1437.
[51] Y.-Y. Wang, Z.-Y. Ye, Z.-S. Zhao, L. Li, Y.-X. Wang, H.-Q. Tao, H.-J. Wang, X.-J. He, Clinicopathologic significance of mir-10b expression in gastric carcinoma, Human pathology 44 (7) (2013) 1278–1285.
[52] J. Wang, Y. Yan, Z. Zhang, Y. Li, Role of mir-10b-5p in the prognosis of breast cancer, PeerJ 7 (2019) e7728.
[53] B. Krist, U. Florczyk, K. Pietraszek-Gremplewicz, A. Józkowicz, J. Dulak, The role of mir-378a in metabolism, angiogen-

esis, and muscle biology, International journal of endocrinology 2015.
[54] B. Mopidevi, S. Maharjan, S. Jain, V. G. Pandey, A. Kumar, Micrornas hsa-mir-584 and hsa-mir-31 regulate expression of human angiotensnogen gene (2012).
[55] C. Cava, A. Colaprico, G. Bertoli, G. Bontempi, G. Mauri, I. Castiglioni, How interacting pathways are regulated by miRNAs in breast cancer subtypes, BMC Bioinformatics 17 (348).

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: