Cancer Screening and Prevention **2022** vol. 1(1) | 32–38
DOI: 10.14218/CSP.2021.00001

Original Article

miRNA in Machine-learning-based Diagnostics of Cancers

William Kang[1], Valentina L. Kouznetsova[2,3] and Igor F. Tsigelny[2,3,4*]

[1]REHS program, San Diego Supercomputer Center, University of California at San Diego, CA, USA; [2]San Diego Supercomputer Center, University of California at San Diego, CA, USA; [3]BiAna, La Jolla, USA; [4]Department of Neurosciences, University of California at San Diego, CA, USA

Received: November 15, 2021 | **Revised:** December 13, 2021 | **Accepted:** December 24, 2021 | **Published:** February 18, 2022

Abstract

Background and objectives: In recent years, miRNAs have been shown to play an important role in many diseases, most notably cancers. Traditional experiments for cancer diagnostics are time-consuming and expensive, thus more attention is being diverted towards discovering computational methods for predicting miRNA–disease association and using them in cancer diagnostics.

Methods: miRNA numerical sequence information and genes targeted by miRNA were used for the construction of descriptors for machine-learning models. Next, we generated a table of miRNA descriptors using all of the miRNAs in a specific cancer dataset for disease classification. To show the effectiveness of the system, we constructed miRNA descriptor systems for pancreatic cancer, lung cancer, and breast cancer.

Results: With the Random Forest classifier, we obtained classification accuracies of 86.9%, 86.3%, and 85.1% for the above-mentioned cancers, respectively. Next, different disease datasets were tested on each model, including new miRNA sets for each cancer type from other studies. The models were able to classify the corresponding cancer miRNAs with >90% accuracy and other disease and cancer datasets with <60% accuracy.

Conclusions: With this information, we constructed a hard-voting scheme using the three cancer classification models that is able to perform cancer diagnostics. The results suggest that our method is effective in miRNA–disease association prediction and performing cancer diagnostics.

Introduction

MicroRNAs (miRNAs) are small noncoding RNAs that are typically 22 nucleotides in length.[1] miRNAs have been found to regulate approximately 30% of genes in humans at the post-transcriptional level.[2] In recent years, miRNA expression has been shown to correlate to diseases, most notably cancers.[3] An example of an miRNA that has been correlated with tumor suppression and cancer inhibition is the miRNA precursor lethal-7 (let-7).[4] As such, miRNAs are now being studied as promising biomarkers for various cancers including pancreatic cancer.[5]

There has been a lot of computational research surrounding miRNA–disease association prediction. For example, Shi and co-workers have proposed a calculation method for miRNA–disease relationship prediction based on random walk analysis.[6] This model uses the connection between miRNAs and disease genes in protein–protein interaction (PPI) networks to predict potential miRNA–disease associations. In addition, Chen and co-authors have proposed a bipartite network projection model for predicting potential associations between miRNAs and disease (BNPMDA) using miRNA functional similarity, disease semantic similarity, and the known human miRNA–disease associations.[7] This model constructs bias ratings between diseases and miRNAs

Keywords: Cancer; miRNA; Neoplasms; Machine learning; Cancer diagnosis.

Abbreviations: ACC, accuracy; AUC, area under the curve; BNPMDA, bipartite network projection for miRNA-disease association; FPR, false-positive rate; GraRep, graph representations; HMDD, human microRNA disease database; MCC, Matthews correlation coefficient; PBMDA, path-based miRNA-disease association; PPI, protein-protein interaction; PRC, precision-recall curve; PREC, precision; REC, recall; ROC (curve), receiver-operating characteristic (curve); TPR, true-positive rate.

***Correspondence to**: Igor F. Tsigelny, San Diego Supercomputer Center, University of California at San Diego, CA 92117, USA; BiAna, San Diego, CA 92117, USA; Department of Neurosciences, University of California at San Diego, CA 92117, USA. ORCID: https://orcid.org/0000-0002-7155-8947. Tel: +1-857-822-0953, Fax: +1-858-581-9073, E-mail: itsigeln@ucsd.edu

How to cite this article: Kang W, Kouznetsova VL, Tsigelny IF. miRNA in Machine Learning-based Diagnostics of Cancers. *Cancer Screen Prev* 2022;1(1):32–38. doi: 10.14218/CSP.2021.00001.

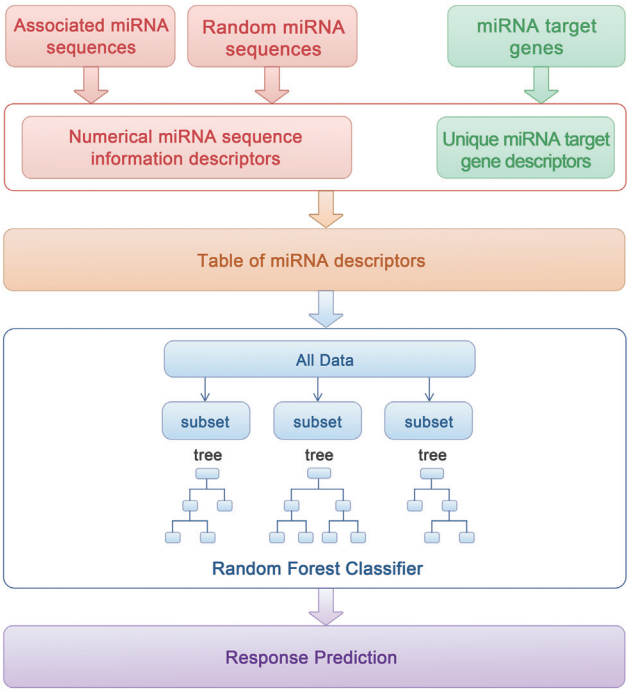
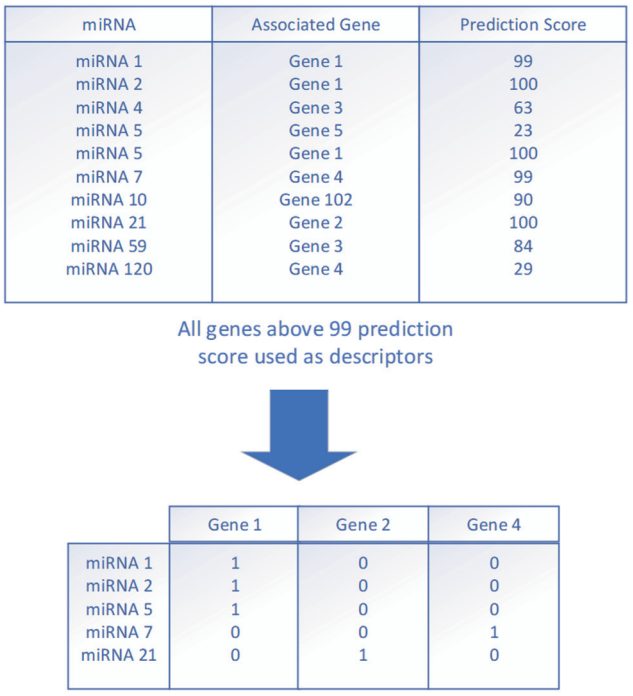


Fig. 1. Flowchart of the method.



miRNA	Associated Gene	Prediction Score
miRNA 1	Gene 1	99
miRNA 2	Gene 1	100
miRNA 4	Gene 3	63
miRNA 5	Gene 5	23
miRNA 5	Gene 1	100
miRNA 7	Gene 4	99
miRNA 10	Gene 102	90
miRNA 21	Gene 2	100
miRNA 59	Gene 3	84
miRNA 120	Gene 4	29

	Gene 1	Gene 2	Gene 4
miRNA 1	1	0	0
miRNA 2	1	0	0
miRNA 5	1	0	0
miRNA 7	0	0	1
miRNA 21	0	1	0

Fig. 2. miRNA descriptors based on target genes.

using miRNA function similarity and disease semantic similarity. Then, the bipartite network recommendation algorithm is applied to predict miRNA–disease association. Moreover, You and colleagues have proposed an miRNA–disease association prediction model called a path-based miRNA–disease association (PBMDA).[8] This model implements a personalized recommendation algorithm that recommends potential miRNA–disease pairs based on information of related miRNAs and diseases. Ji and co-authors also have proposed a network embedding-based heterogeneous information integration method to predict the potential associations between miRNA and disease.[9] This model first used a heterogeneous information network constructed using known associations between drugs, miRNA, protein, lncRNA, and disease and then applied the graph-representations (GraRep) method to learn and predict potential miRNA–disease associations.

Using the previously obtained results in miRNA–disease relation prediction, a set of databases of miRNA–disease relationships was created during the past several years. We conducted the next step—we used these databases to construct a set of descriptors for machine-learning diagnostics using miRNA. In our study, a novel miRNA descriptor system was proposed to predict potential associations between miRNAs and diseases. Based on our hypothesis that miRNA–disease association can be elucidated using sequence information of miRNAs and genes targeted by miRNA, we constructed our miRNA descriptor system using numerical sequence information of miRNAs and target genes.

Methods

Classification model

To show the effectiveness of our miRNA descriptor system, we constructed a classification model using known associations of miRNAs with various cancers. We illustrated the concept of the system in more detail using a pancreatic cancer model as an example. From the miRNA cancer association database miRCancer,[10] we extracted 107 miRNAs that are associated with breast cancer, and from the miRNA database miRBase,[11] we extracted 107 random miRNAs as training/testing data. The model was then constructed based on miRNA descriptors created using the training/testing data. Figure 1 shows the flowchart of our method. The model was then evaluated using the Random Forest machine-learning algorithm. The results reveal that our method performed with a high accuracy of 86.9%.

Developing the miRNA descriptor system

We developed a system of miRNA descriptors taking in consideration the known miRNA–cancer associations and miRNA target predictions (Fig. 2). The system was tested based on pancreatic cancer as an example as follows. A list of miRNAs that are known to be associated with pancreatic cancer was downloaded from the miRCancer[10] database, and miRNA target predictions were downloaded from the miRNA target prediction database miRDB.[12] A list of all known miRNAs was also downloaded from miRBase.[11] To extract the sequence information of the miRNAs that are associated with pancreatic cancer, a Python script was written to find the pancreatic cancer-associated miRNAs (name) in the miRDB and to extract the corresponding miRNA sequence. In total, 152 miRNA sequences (associated with pancreatic cancer) were extracted in this manner. An additional 152 human miRNA sequences with no known association to pancreatic cancer were randomly selected from miRBase, for a total of 304 miRNA sequences. The 152 miRNA sequences with a known association to pancreatic cancer were assigned "associated" labels and the 152 randomly selected miRNA sequences were assigned "non-associated" labels to create two categories for classification. These miRNA sequences were later used as inputs to

Table 1. miRNA descriptors based on the sequences

Name of descriptor	Description/Formula
Number of base pairs	N
Number of each base pair	x_A, x_U, x_C, x_G
Frequency of each base pair	x_A/N, x_U/N, x_C/N, x_G/N
Mean mass of each base pair	$(135.1(x_A) + 112.1x_U) + 111.1(x_C) + 151.1(x_G))/N$
Number of hydrogen bonds	$2(x_A + x_U) + 3(x_C + x_G)$
Symmetry score	If the first base pair is the same as the last base pair, add 1 to the symmetry score. If the second base pair is the same as the second-to-last base pair, add 1 to the symmetry score. Repeat until the middle of the miRNA ($N/2$ term) is reached.
2-base-pair motifs (i.e., AA, AU, AC) of the entire sequence	Each motif is a separate descriptor. If the miRNA has the motif, a "1" is assigned. Otherwise, a "0" is assigned.
3-base-pair motifs (i.e., AAA, AAU, AAC) of the entire sequence	Each motif is a separate descriptor. If the miRNA has the motif, a "1" is assigned. Otherwise, a "0" is assigned.
4-base-pair motifs (i.e., AAAA, AAAU) of the entire sequence	Each motif is a separate descriptor. If the miRNA has the motif, a "1" is assigned. Otherwise, a "0" is assigned.
Motifs (2-, 3-, 4-base pair) within the first 5 base pairs	Each motif is a separate descriptor. If the 5 first base pairs of the miRNA contains the motif, a "1" is assigned. Otherwise, a "0" is assigned.
Motifs (2-, 3-, 4-base pair) within the last 5 base pairs	Each motif is a separate descriptor. If the 5 first base pairs of the miRNA contains the motif, a "1" is assigned. Otherwise, a "0" is assigned.

create the miRNA descriptors.

Another Python program was developed to automatically extract the miRNA descriptors based on the miRNA sequences (associated with pancreatic cancer) as the input. One part of the miRNA descriptors consisted of numerical miRNA sequence information. The miRNA sequence information used in this study is more complete and comprehensive compared to previous studies.[12] The miRNA descriptors based on the sequence information consisted of the number of base pairs, the assigned number of each base pair, the frequency of each base pair, the mean mass of each base pair, the number of hydrogen bonds, the symmetry of the miRNA sequence, the motifs within the entire miRNA sequence (2, 3, 4 base pair motifs), and the motifs within the first five base pairs and within the last five base pairs. Each motif was a distinct descriptor and was assigned a score of "1" if the miRNA sequence had the motif and a score of "0" if it did not have the motif. Table 1 includes the names and formulas/descriptions for each of the numerical descriptors based on miRNA sequence information. In total, there were 996 miRNA descriptors based on the sequence information.

The other part of the miRNA descriptor set was based on miRNA target genes from the miRDB database.[12] These miRNA target genes were included as descriptors because we hypothesized that miRNAs that are associated with the same disease will share similar targets as well.[13] For this study, a target score threshold of 99 was used to make sure the miRNA and selected target genes were strongly correlated. A Python program was developed to automatically find all target genes with a target score of 99 or more for all miRNAs. Then, the program created a new descriptor for each unique gene selected. A target gene descriptor was assigned a score of "1" if the miRNA sequence had a target score of 99 or more with that target gene and a score of "0" otherwise. Figure 2 shows how the miRNA target gene descriptors were created. In this study, a total of 6,436 target gene descriptors were created from the 304 miRNA sequences.

The miRNA descriptor system was developed to take a list of miRNAs as the input and to issue a table with the sequence information and target gene descriptors for each miRNA as the output. This system could be applied to any disease with known miRNA–disease associations.

Machine learning

We describe the system performance based on pancreatic cancer as an example. The 6,436 target gene-based descriptors and the 996 numerical miRNA sequence-based descriptors from the 304 miRNA sequences were combined to create a single miRNA descriptor table with 304 miRNA sequences (rows) and 7,432 descriptors (columns). An additional column was added to the descriptor table to label the two classes of data for classification. The 304 miRNAs that are associated with pancreatic cancer were given the class of "associated" while the 304 randomly selected miRNAs were given the class of "non-associated." The descriptor table was then used as the input for multiple machine-learning classification algorithms. Out of all of the classification algorithms, Random Forest[14] with an 80%/20% training–testing split had the highest classification accuracy.

We first used Random Forest with an 80%/20% training-testing split to evaluate the performance of the model before any feature selection was done. An 80%/20% training-testing split ensures that there is no overfitting as 20% of the data is not used to build the model but is used for testing. Then, the InfoGainAttributeEval[15] algorithm was used to determine which descriptors contribute the most to information gain during classification. The descriptors that have no contribution to information gain were removed, thus leaving a list of descriptors that have a positive contribution to classification, ordered from the greatest contribution to the least contribution.

The reduced table of descriptors then went through more precise feature selection. A script removed descriptors one by one starting from the descriptors with the least information gain contribution, evaluated the performance of the model using Random Forest, and kept the deletion if the classification accuracy increased. Overall, the number of descriptors for our 304 miRNA sequences was reduced from 7,432 to 3,648 descriptors.

Table 2. Performance comparison of the different classifiers for the developed machine-learning models

Classifier	ACC	PREC	MCC	TPR	FPR	AUC	PRC area
LMT	81.97%	82.1%	80.4%	82.0%	18.5%	85.9%	84.9%
SVM	81.97%	82.1%	64.0%	82.0%	17.9%	82.1%	76.4%
Naïve Bayes	80.26%	82.6%	63.0%	80.3%	18.4%	84.9%	83.0%
Random Forest	86.88%	87.1%	73.9%	86.9%	12.8%	86.4%	86.1%

LMT, Logistic Model Tree; SVM, Support Vector Machine; ACC, accuracy; PREC, precision; MCC, Matthews correlation coefficient; TPR, true-positive rate; FPR, false-positive rate; AUC, area under the receiver-operating characteristic curve; PRC area, area under the precision-recall curve.

Results

The results of the proposed classification were evaluated using confusion matrices and their derivatives: the accuracy (ACC), precision (PREC), Matthews correlation coefficient (MCC), true-positive rate (TPR) or recall (REC), false-positive rate (FPR), as well as the area under the receiver-operating characteristic (ROC) curve (AUC), and the area under the precision-recall curve (PRC area). Comparison of different classifiers results (pancreatic cancer) is presented in Table 2. The best weighted averages for each of these metrics were as follows: ACC, 86.9%; PREC, 87.1%; MCC, 73.9%; TPR (REC), 86.9%; FPR, 12.8%; AUC, 86.4%; and PRC area, 86.1%.

The ROC curve compares the sensitivity and specificity across a range of values. Thus, the vertical axis is the TPR, that is, the sensitivity or recall; and the horizontal axis is the FPR or (1−specificity). The FPR is the probability of falsely classifying a positive class. The best performance showed the model based on the Random Forest classifier. The model's low FPR of 12.8% demonstrates a low probability of wrongly classifying an miRNA–breast cancer pair that is associated. The TPR (sensitivity) is the probability of correctly classifying a positive class. The model's high TPR of 86.9% indicates a high probability of correctly classifying an miRNA–breast cancer pair that is associated. The large average AUC value of 86.4% indicates that the Random Forest classifier is very robust. Another way to evaluate the performance of the proposed method is the PRC area, which shows precision values for the corresponding sensitivity (recall, i.e., TPR) values. The model's large PRC area value of 86.1% once again shows the good performance of our method.

Performance comparison of the different classifiers for the developed machine-learning models

To further test the significance of the classifier on our model, we compared the performance of the four classifiers Random Forest,[14] Naïve Bayes,[16] Logistic Model Tree,[17] and Support Vector Machine[18] using the 80%/20% training–testing split. In the comparison, the environment and training/testing set were kept the same and only the classifier engine was changed. Additionally, the same statistic metrics of ACC, PREC, MCC, TPR (REC), FPR, AUC, and PRC area were used. Table 3 shows the comparison of the performance of all of the classifiers. The comparisons show that the Random Forest classifier had a better performance, robustness, accuracy, and sensitivity than the other classifiers for our system.

miRNA-based diagnostics of various cancers

To prove the robustness of our system of descriptors for miRNA–disease prediction, we conducted case studies using pancreatic cancer, lung cancer, and breast cancer. Previously, we tested our method on pancreatic cancer using target gene descriptors based on target gene prediction scores of 99 or higher. To explore whether the number of parameters (descriptors) has a significant impact on the prediction performance (statistic metrics), we conducted each case study using two different target gene prediction thresholds, 90 and 99. Each case study was conducted using Random Forest with an 80%/20% training-testing data split, and we evaluated the models using the same statistic metrics of ACC, PREC, MCC, TPR (REC), FPR, AUC, and PRC area. Additionally, the same method was used to create the miRNA target gene-based descriptors and to perform feature selection on each study. Table 3 shows the average prediction statistic metrics of performance for each case study.

The results show that the accuracies of the case studies are both consistent and high, ranging from 85.1% to 88.5%, proving the robustness of the method for miRNA–disease association prediction. Additionally, the prediction accuracies are consistently high for both a target gene prediction threshold of 99 and a target gene prediction threshold of 90 for each case study, showing that the

Table 3. Comparison of the miRNA-based diagnostics on various cancers and different target gene thresholds

Cancer type and prediction threshold	ACC	PREC	MCC	TPR	FPR	AUC	PRC area
Breast cancer (target gene threshold of 90)	87.7%	88.0%	76.4%	87.7%	12.9%	88.7%	86.4%
Breast cancer (target gene threshold of 99)	85.1%	85.6%	70.1%	85.1%	14.3%	88.3%	86.5%
Lung cancer (target gene threshold of 90)	86.3%	86.9%	73.2%	86.3%	13.2%	88.5%	85.7%
Lung cancer (target gene threshold of 99)	86.3%	86.7%	72.8%	86.3%	13.0%	88.9%	87.9%
Pancreatic cancer (target gene threshold of 90)	88.5%	88.7%	78.4%	88.5%	11.5%	88.9%	87.5%
Pancreatic cancer (target gene threshold of 99)	86.9%	87.1%	73.9%	86.9%	12.8%	86.4%	86.1%

ACC, accuracy; PREC, precision; MCC, Matthews correlation coefficient; TPR, true-positive rate; FPR, false-positive rate; AUC, area under the receiver-operating characteristic curve; PRC area, area under the precision-recall curve.

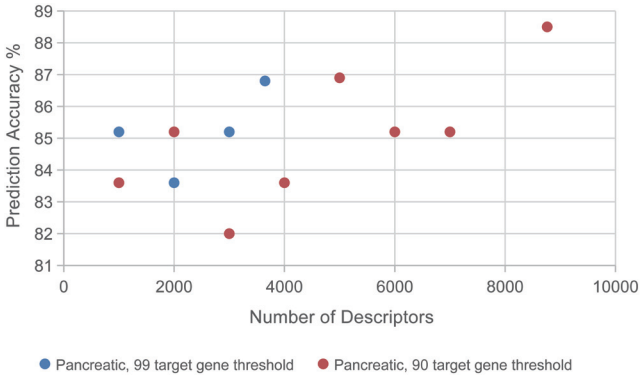


Fig. 3. Comparison of the prediction accuracy for the pancreatic cancer model with different numbers of descriptors.

method operates robustly even with small numbers of descriptors.

To further explore the relationship between the number of descriptors and the prediction accuracies, the number of descriptors and the prediction accuracy for each case study are compared in Figures 3–5. For each case study, descriptors were removed based on their information gain contribution (the descriptors with the least information gain were removed first). While there were fluctuations in the prediction accuracies when the number of descriptors was reduced, the prediction accuracies in each case study were still consistently high across all numbers of descriptors for all case studies. This finding proves that although the results were the highest when there was a high number of descriptors, high accuracies could still be achieved across all diseases with lower numbers of descriptors.

Testing of outside datasets on the developed models

To ensure that our models are able to differentiate between different diseases, various datasets were tested on each model. From each disease dataset, approximately 50 randomly selected associated miRNAs were paired with the same number of randomly selected non-associated miRNAs. Then, the selected data were used to test the model.

First, randomly selected data from the lung cancer and breast cancer datasets were tested on the pancreatic cancer model with a target gene threshold of 99. The model classified the lung cancer data with 57.8% accuracy and the breast cancer data with 56.4%

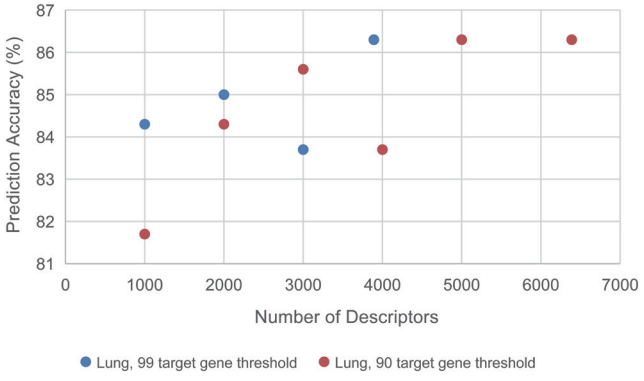


Fig. 4. Comparison of the prediction accuracy for the lung cancer model with different numbers of descriptors.

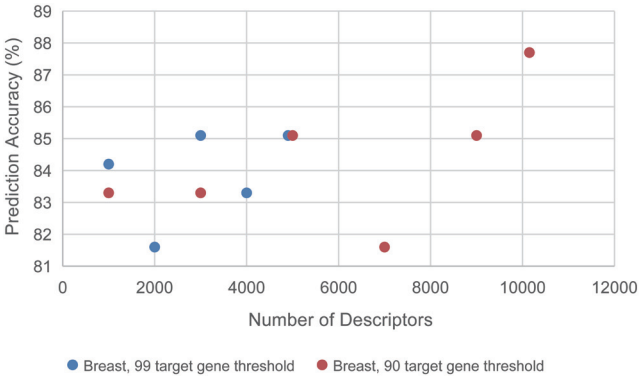


Fig. 5. Comparison of the prediction accuracy for the breast cancer model with different numbers of descriptors.

accuracy. Next, randomly selected data from the pancreatic cancer and breast cancer datasets were tested on the lung cancer model with a target gene threshold of 99. The model classified the pancreatic cancer data with 56.7% accuracy and the breast cancer data with 58.3% accuracy. Finally, randomly selected data from the lung cancer and pancreatic cancer datasets were tested on the breast cancer model with a target gene threshold of 99. The model classified the lung cancer data with 56.7% accuracy and the pancreatic cancer data with 55.9% accuracy. The accuracies are presented in Table 4. The accuracies were all higher than 50% because of some overlapping miRNAs between the three datasets. However, each model performed consistently worse when classifying datasets from other diseases. Thus, we conclude that the models are able to differentiate between different cancers.

We also tested a noncancer disease on the system to provide further verification. From the HMDD database,[19] we extracted 86 miRNAs associated with Alzheimer's disease and also extracted 86 miRNAs not associated with Alzheimer's disease. The selected data were tested on the pancreatic cancer, lung cancer, and breast cancer association models with 90 target genes. The pancreatic cancer model classified the Alzheimer's data with 48.1% accuracy, the lung cancer model classified the Alzheimer's data with 53.0% accuracy, and the breast cancer model classified the Alzheimer's data with 56.7% accuracy. The low classification accuracies of the Alzheimer's data further demonstrate that the models are able to differentiate between different diseases. The accuracies are shown in Table 5.

Finally, we tested pancreatic, breast, and lung cancer data from other studies[5,20-22] on our corresponding models. From the other studies, we were able to extract 12 lung cancer-associated miRNAs, 13 pancreatic cancer-associated miRNAs, and 30 breast cancer-associated miRNAs that were not presented in our model. Then, the same number of unassociated miRNAs was paired with the cancer data and tested on each model.

The lung cancer data from other studies[21] yielded a 91.7% accuracy when tested on our lung cancer model; the pancreatic cancer data from other studies[5,20] yielded a 92.3% accuracy when tested on our pancreatic cancer model; and the breast cancer data from other studies[22] yielded a 95.0% accuracy when tested on our breast cancer model. These results verify the validity of our models.

After establishing that our individual classifiers for breast, pancreatic, and lung cancer were able to differentiate between different diseases, we used a hard-voting scheme to recognize different cancers from a single input dataset. A hard-voting scheme uses

Table 4. Comparison of the diagnostic accuracies using different cancer datasets for tests on the developed models

	Pancreatic cancer model (99 target genes)	Lung cancer model (99 target genes)	Breast cancer model (99 target genes)
Pancreatic cancer dataset accuracy	86.9%	57.8%	56.4%
Lung cancer dataset accuracy	56.7%	86.3%	58.3%
Breast cancer dataset accuracy	55.9%	56.7%	85.1%

Table 5. Comparison of the diagnostic accuracies of Alzheimer's disease tested on the developed models

	Pancreatic cancer model (99 target genes)	Lung cancer model (99 target genes)	Breast cancer model (99 target genes)
Alzheimer's disease dataset accuracy	48.1%	53.0%	56.7%

majority voting for classification. The hard-voting scheme was applied to each of the three individual models. Table 6 shows examples of the results from the hard-voting scheme.

Discussion

Predicting the associations between miRNA and disease not only can greatly help to understand the role of miRNA in the development of diseases, but it can also significantly improve the early diagnosis of the specific disease. In this study, we proposed a new systematic method to predict the association between miRNA and disease using miRNA descriptors that consist of miRNA sequence information and target gene information. To demonstrate the effectiveness of the method, we used the system to create a machine-learning model for the diagnosis of several cancers using the miRNA profiles of breast, pancreatic, and lung cancer patients. The model's good performance shows that an miRNA's association with these cancers is highly related to patterns within the miRNA sequence information and target genes. Additionally, the InfoGainAttributeEval algorithm provides further insights on the specific properties that have the greatest effect on the information gain aspect of the classification. Deeper analysis could be done regarding the specific properties of miRNA that are the most important in determining their association with disease. We can conclude that the developed system of miRNA descriptors is effective in cancer diagnostics.

Conclusions

In this article, we proposed a new miRNA descriptor system created using miRNA sequence information and target gene information to develop machine-learning models for pancreatic, lung, and breast cancer. The models were trained and evaluated using the Random Forest classifier. Then, the models were tested using different disease datasets from other studies. Each model was able to classify its corresponding cancer with >90% accuracy and other diseases with <60% accuracy. Finally, a hard-voting scheme was created using the relative classification accuracies of each model to perform cancer diagnosis. The final experimental results show that our method performs well and is effective for the classification of cancers. In addition, the hard-voting scheme proves that our method is able to perform cancer diagnosis. Therefore, we believe that our proposed method will be a useful tool for performing cancer diagnosis in the future.

Acknowledgments

None.

Funding

No funding obtained.

Conflict of interest

Prof. Igor F. Tsigelny is the president of BiAna, and Prof. Valentina L. Kouznetsova is the CEO of BiAna. The authors have no other conflicts of interest related to this publication.

Author contributions

VLK and IFT proposed the strategy of using miRNA sequences and related genes for the creation of a machine-learning system for cancer diagnostics. WK developed the machine-learning set of descriptors, trained the system, and wrote the article. VLK and IFT supervised the development of the machine-learning system and edited the article.

Table 6. Hard-voting scheme for different cancer diagnostics

Breast cancer	Lung cancer	Pancreatic cancer	Hard-voting consensus
Accuracy >80	Accuracy <60	Accuracy <60	Breast cancer
Accuracy <60	Accuracy >80	Accuracy <60	Lung cancer
Accuracy <60	Accuracy <60	Accuracy >80	Pancreatic cancer
Accuracy <60	Accuracy <60	Accuracy <60	None

Data sharing statement

No additional information is available.

References

[1] Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer 2006;6(4):259–269. doi:10.1038/nrc1840, PMID:16557279.

[2] Felekkis K, Touvana E, Stefanou Ch, Deltas C. microRNAs: a newly described class of encoded molecules that play a role in health and disease. Hippokratia 2010;14(4):236–240. PMID:21311629.

[3] Peng Y, Croce CM. The role of MicroRNAs in human cancer. Signal Transduct Target Ther 2016;1:15004. doi:10.1038/sigtrans.2015.4, PMID:29263891.

[4] Zhang B, Pan X, Cobb GP, Anderson TA. microRNAs as oncogenes and tumor suppressors. Dev Biol 2007;302(1):1–12. doi:10.1016/j.ydbio.2006.08.028, PMID:16989803.

[5] Daoud AZ, Mulholland EJ, Cole G, McCarthy HO. MicroRNAs in Pancreatic Cancer: biomarkers, prognostic, and therapeutic modulators. BMC Cancer 2019;19(1):1130. doi:10.1186/s12885-019-6284-y, PMID:31752758.

[6] Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, *et al*. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst Biol 2013;7:101. doi:10.1186/1752-0509-7-101, PMID:24103777.

[7] Chen X, Xie D, Wang L, Zhao Q, You ZH, Liu H. BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. Bioinformatics 2018;34(18):3178–3186. doi:10.1093/bioinformatics/bty333, PMID:29701758.

[8] You ZH, Wang LP, Chen X, Zhang S, Li XF, Yan GY, *et al*. PRMDA: Personalized recommendation-based miRNA-disease association prediction. Oncotarget 2017;8(49):85568–85583. doi:10.18632/oncotarget.20996, PMID:29156742.

[9] Ji B, You Z, Cheng L, Zhou J, Alghazzawi D, Li L. Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. Sci Rep 2020;10(1):6658. doi:10.1038/s41598-020-63735-9, PMID:32313121.

[10] Xie B, Ding Q, Han H, Wu D. miRCancer: A microRNA-cancer association database constructed by text mining on literature. Bioinformatics 2013;29(5):638–644. doi:10.1093/bioinformatics/btt014, PMID:23325619.

[11] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: From microRNA sequences to function. Nucleic Acids Res 2019;47(D1):D155–D162. doi:10.1093/nar/gky1141, PMID:30423142.

[12] Chen Y, Wang X. miRDB: An online database for prediction of functional microRNA targets. Nucleic Acids Res 2019;48(D1):D127–D131. doi:10.1093/nar/gkz757, PMID:31504780.

[13] Kehl T, Backes C, Kern F, Fehlmann T, Ludwig N, Meese E, *et al*. About miRNAs, miRNA seeds, target genes and target pathways. Oncotarget 2017;8(63):107167–107175. doi:10.18632/oncotarget.22363, PMID:29291020.

[14] Svetnik V, Liaw A, Tong C, Culberson J, Sheridan R, Feuston B. Random Forest: A classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43(6):1947–1958. doi:10.1021/ci034160g, PMID:14632445.

[15] Alhaj TA, Siraj MM, Zainal A, Elshoush HT, Elhaj F. Feature selection using information gain for improved structural-based alert correlation. PLoS ONE 2016;11(11):e0166017. doi:10.1371/journal.pone.0166017, PMID:27893821.

[16] Kumar KUP, Kalimuthu M. Performance analysis of Naïve Bayes correlation models in machine learning. Int J Psychosoc Rehabil 2020;24(04):1153–1157. doi:10.37200/ijpr/v24i4/pr201088.

[17] Landwehr N, Hall M, Fran E. Logistic model trees. Mach Learn 2005;59(1–2):161–205. doi:10.1007/s10994-005-0466-3.

[18] Noble W. What is a support vector machine? Nat Biotechnol 2006;24(12):1565–1567. doi:10.1038/nbt1206-1565, PMID:17160063.

[19] Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, *et al*. HMDD v3.0: A database for experimentally supported human microRNA–disease associations. Nucleic Acids Res 2018;47(D1):D1013–D1017. doi:10.1093/nar/gky1010, PMID:30364956.

[20] Shams R, Saberi S, Zali M, Sadeghi A, Ghafouri-Fard S, Aghdaei HA. Identification of potential microRNA panels for pancreatic cancer diagnosis using microarray datasets and bioinformatics methods. Sci Rep 2020;10(1):7559. doi:10.1038/s41598-020-64569-1, PMID:32371926.

[21] Han Y, Li H. miRNAs as biomarkers and for the early detection of non-small cell lung cancer (NSCLC). J Thorac Dis 2018;10(5):3119–3131. doi:10.21037/jtd.2018.05.32, PMID:29997981.

[22] Jang JY, Kim YS, Kang KN, Kim KH, Park YJ, Kim CW. Multiple microRNAs as biomarkers for early breast cancer diagnosis. Mol Clin Oncol 2021;14(2):31. doi:10.3892/mco.2020.2193, PMID:33414912.