International Journal on Emerging Technologies 11(2): 453-460(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Discovery of Significant miRNA-biomarkers for Breast Cancer using Decision Tree Classifier

Apurva A. Mehta[1] and Himanshu S. Mazumdar[2]

[1]Assistant Professor, Department of Computer Engineering, Dharmsinh Desai Univeristy, Nadiad (Gujarat), India.
[2]Head, R & D Center, Faculty of Technology, Dharmsinh Desai University, Nadiad (Gujarat), India.

(Corresponding author: Apurva A. Mehta)
(Received 31 December 2019, Revised 15 February 2020, Accepted 18 February 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Breast cancer is the most common cancer in women around the world, making biomarker discovery for breast cancer very important. Machine learning based methods are adopted hugely in almost all computational biology tasks due to advancements in computing facility. Due to The Cancer Genome Atlas (TCGA) project, it is possible to analyze genomic and molecular cancer data using machine learning algorithms. Breast cancer data available from TCGA is processed and analyzed with decision tree algorithm. Tree classifier uses 465 miRNAs as attributes after removal of 581 miRNAs with 50% or more sparseness. The decision tree classifier identifies hsa-mir-139, hsa-mir-10b, hsa-mir-3677 and hsa-mir-21 as the significant miRNAs for distinguishing breast cancer and non-breast cancer sample. The functional enrichment results suggest hsa-mir-205, hsa-mir-21 and hsa-mir-33b are associated with breast cancer.

Keywords: breast cancer, decision tree classifier, miRNA biomarker, The Cancer Genome Atlas.

Abbreviations: TCGA, the cancer genome atlas; GDC, genomic data commons; RNA, ribonucleic acid; miRNA, micro ribonucleic acid; SVM, support vector machine.

I. INTRODUCTION

Breast cancer is reported as the most common diagnosed malignancy in the women [1-3]. Mortality can be controlled by finding out improved diagnostic strategies. The microRNA (i.e. miRNA) are small (19-24 nt), non-coding RNA [4]. miRNA are responsible for managing several activities in the human immune system [5]. Defects in the working of miRNA are related with cancers and other diseases [6, 7]. The emergence of high throughput technologies made it possible to collect large amount of biological data for cancer through The Cancer Genome Atlas (i.e. TCGA) project [8].

There are three main challenges while leveraging advantages due to availability of biological data using computational techniques and gaps found in research work of this field. The proposed approach must be capable of handling hundreds of parameters present within data. The obtained prediction results must be validated by results captured using experimental methods. The networked relation must be captured between predicted parameter and targeted entity.

Differential expression is the widely used statistical method for biomarker discovery. Machine learning algorithms can be used to identify key biomarkers in terms of miRNA. Recently, decision tree classifier is used to predict lung cancer status and sub-type [9] using TCGA dataset. The work used 5 miRNAs in the diagnosis of lung cancer status and sub-typing. The researchers use SVM based classifier to classify cancer patients of TCGA into early and advanced stages [10]. Their work identifies 34 significant miRNAs that achieves mean accuracy of 80.38% during a 10-fold cross validation. The TCGA dataset is also used to find biomarkers for soft tissue sarcomas using Random forest algorithm [11]. The colorectal cancer data from TCGA is used to find prognostic miRNA using expression analysis [12].

It is observed during literature survey, research works on identifying miRNA-biomarkers from TCGA dataset has not yet obtained desired results for breast cancer. While breast cancer is the malignancy reported widely all over the world in the women population. Even, recent research works are not fully establishing relation between reported miRNA with gene product for understanding impact due to malignancy. Thus, in the current study we aim to identify biomarkers for breast cancer using decision tree classifier. Biomarkers used for deciding thresholds in tree classifier are linked with other experimental works related to breast cancer. This step is essential in validating our research results. Then, network of significant miRNA and gene product is prepared for functional enrichments. This gives understanding from other dimension in terms of miRNA-gene relation targeting disease non-disease situation. The functional enrichment results assert decision tree classifier claims. Decision tree classifier is used to establish a machine learning model which is sound and interpretable. Decision tree algorithm also implicitly perform feature screening which has additive advantage where thousands of attributes are present.

The classifier predicts important miRNA-biomarker for breast cancer with accuracy of more than 97% consistently surpassing other research works [9, 10, 12] while handling hundreds of attributes during machine learning. The reported miRNA-biomarkers are verified against experimentally found evidences for their role in breast cancer. Finally, a networked relation is established for showing clear association between predicted parameter and targeted entity. This would become a starting point for research in protein-protein interaction and targeted drug discovery.

II. MATERIALS AND METHODS

A. Dataset

The Cancer Genome Atlas (i.e. TCGA) [8] is a project for indexing various cancer genomics data. Through comprehensive and coordinated efforts TCGA program has catalogued genomics data for 33 cancer types. The TCGA data is available through Genomics Data Commons (i.e. GDC) [13]. There are different ways available for downloading TCGA data from GDC ranging from manual selection to programmable way. In this work, TCGA biolinks [14, 15] an R/Bioconductor package is used to download TCGA data from GDC. TCGA biolinks allows users to query, download and perform analysis on genomics data.

Pseudocode-I

```
query_brca ← GDCquery(project = "TCGA-BRCA",
data.category = "Gene expression", data.type =
"miRNA gene quantification", legacy = TRUE )
GDCdownload(query_brca)
```

The queries mentioned in Pseudocode-I are used to query and download miRNA sequencing of TCGA BRCA (i.e. Breast Invasive Carcinoma) legacy data. The legacy archive holds the TCGA data submitted by original submitter.

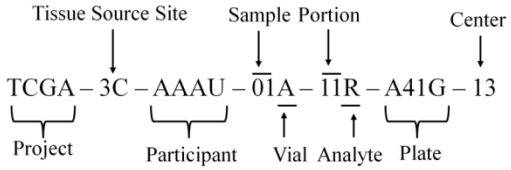

Fig. 1. TCGA Barcode Structure.

There are 1054 legacy miRNA quantification files available. Each file is recognized uniquely through the barcode. The structure of barcode is shown in Fig. 1. The meaning of each label of barcode is summarized in Table 1.

Table 1: TCGA Barcode Label Description.

Label	Identifier for	Value	Value Description
Project	Project name	TCGA	TCGA project
Tissue Source Site (TSS)	Tissue Source Site	3C	Columbia University (Breast invasive carcinoma)
Participant	Study participant	AAAU	AAAU participant from BRCA study at Columbia University
Sample	Sample type	01	Primary solid tumor (TP)
Vial	Order of sample in a sequence of samples	A	The first vial
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	11	The eleventh portion of the sample
Analyte	Molecular type of analyte for analysis	R	RNA
Plate	Order of plate in a sequence of 96-well plates	A41G	The A41Gth plate
Center	Sequencing or characterization center that will receive the aliquot for analysis	13	Canada's Michael Smith Genome Sciences Centre (BCGSC)

The BRCA project has 951 solid tumor samples (TP, Sample=01) and 103 normal tissue samples (NT, Sample=11). TP samples represent breast cancer patient records and NT samples represent non-breast cancer patient records. All the miRNA quantification samples of TP and NT contain information as given in Table 2. In this work miRNA_ID is considered as feature and reads_per_million_miRNA_mapped is considered as feature value during machine learning pipeline.

Table 2: miRNA quantification file description.

Header	Header description
miRNA_ID	miRNA name
read_count	raw read count
reads_per_million_miRNA_mapped	The read counts are normalized to counts per million by dividing the total read counts of a miRNA by the total read counts of the sample and multiplying this number by 10^6
cross-mapped	cross-mapped to other miRNA forms (Yes or No)

B. Data-preprocessing

The miRNA quantification files contain read counts for 1046 miRNAs. These records are heavily sparse. There are large number of miRNAs with no read count (i.e. 0)

in each record. Statistical analysis is performed to identify level of sparseness across TP and NT samples. Details as shown in Table 3 confirm the higher degree of sparseness present within data.

Table 3: Sparseness statistics for miRNA quantification records across TP and NT samples.

Rule for measuring sparseness	Count of sparse attributes in TP samples	Count of sparse attributes in NT samples
80% and more sparse data	506	516
60% and more sparse data	581	590
50% and more sparse data	609	614

These information will be helpful during machine learning for deciding whether an attribute to include or not. Sparse data can introduce bias into learning and may lead to formulate incorrect/ incomplete inferences.

C. Machine learning algorithm

The machine learning algorithms are widely used to solve various computational biology problems. There are different machine learning algorithms like support vector machine, random forest, decision tree classifier used in past works to identify significant miRNAs for a target disease using TCGA miRNA quantification data. We propose to use decision tree classifier algorithm [16] using python's scikit-learn library [17]. The decision tree algorithm is given in pseudocode-II. Python's scikit-learn library uses Gini impurity as attribute selection measure.

Pseudocode-II

(1) Select an attribute with attribute selection measure (i.e. Gini impurity, Information gain) for splitting the dataset records.

(2) Set the selected attribute as a decision node and continue splitting datasets further.

(3) Repeat steps (1) and (2) for each child node generated before one of the following termination criteria is met.
There are no more attributes remaining to apply on node.
There are no more samples.
All the samples belong to same decision class.

Gini impurity uses three steps process for attribute selection [18]. First, the impurity at node n is calculated using Eqn. 1. Then, each attribute from the randomly selected set is evaluated for each of its value with the focus on reducing entropy after a split at a node n using Eqn. 2. Then, reduction in the impurity in form of the entropy after a split using an attribute is evaluated using Eqn. 3. An attribute that gives the minimum value (i.e. maximum drop in entropy after split) for Eqn. 3 is finally selected for the split at node n.

$$Gini(n) = 1 - \sum_{m=0}^{c-1} [p(m \mid n)]^2 \quad (1)$$

where m represents possible class labels, p is a function calculating conditional probability

$$Gini_f(n) = \frac{|D1|}{|D|} Gini(D1) + \frac{|D2|}{|D|} Gini(D2) \quad (2)$$

where f is a feature from random feature subset at node n, D indicates samples before node split, D1 and D2 samples created after split.

$$\Delta Gini_f(n) = Gini(n) - Gini_f(n) \quad (3)$$

The machine learning process followed for training and validation of identifying significant miRNAs using decision tree algorithm is shown in Fig. 2. It initiates with miRNA quantification data downloaded from GDC. The sparse records are removed before further processing. We remove features (i.e. attributes) with 50% sparseness across TP and NT samples.

There are 609 and 614 features with 50% and more sparse data across TP and NT samples respectively. There are common 581 features present within TP and NT samples and having 50% and more sparseness, further removed from datasets. The training and validation is performed with dataset of 465 features. Datasets of features and binary classes is divided exclusively into training and testing set with proportion of 70% and 30% respectively. The thresholds for building machine learning model are found from the trained model and utilized for performance measurement during validation (i.e. testing).

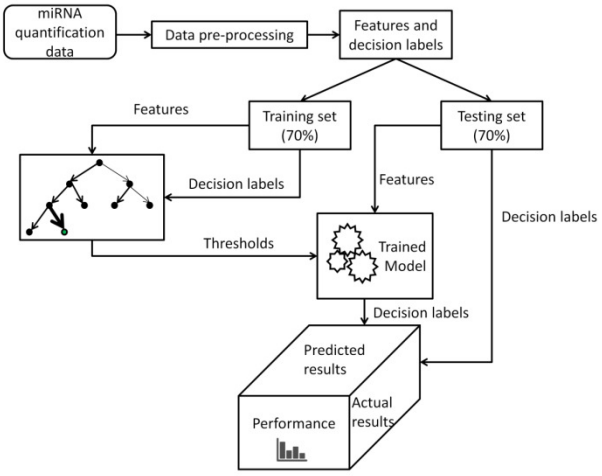

Fig. 2. Training and validation process for retrieving significant miRNAs using decision tree algorithm.

III. RESULTS AND DISCUSSION

The decision tree classifier is used to predict between a breast cancer positive sample (TP) and breast cancer negative (NT) sample. The initial classifier is trained with 465 features due to removal of 581 sparse attributes common to both classes as per 50% sparseness rule. The Fig. 3 shows the performance achieved on the test data. The Fig. 3 highlights the performance in terms of accuracy, precision, recall and specificity. The accuracy is a widely used parameter for performance evaluation designating the correctly predicted cases against all cases. The precision assists in deciding how precise is a model in correctly predicting actual positive cases out of all cases predicted as positive. The recall expresses how many actual positive cases are predicted from all available actual positive cases. The specificity shows out of all the negative cases, how many samples are predicted as negative [18]. It is observed from Fig. 3, there are majority true positive and true negative samples in comparison to false positive and false negative samples. The precision parameter has lesser value (≈81%) for NT status as out of predicted NT (breast cancer negative) samples some are actually TP (breast cancer positive) samples. This may have serious consequences.

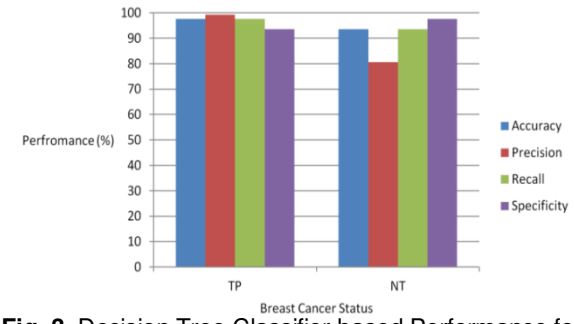

Fig. 3. Decision Tree Classifier based Performance for Breast Cancer Status using 465 attributes.

The tree generated for this prediction is shown in Fig. 4. The Fig. 4 displays potentially significant biomarkers (i.e miRNAs) used for deciding thresholds for prediction. Significance of biomarkers based on importance for a particular biomarker in a tree generated against dataset given. The Table 4 lists each biomarker with respective quantitative importance recorded. Table 4 also includes recent references validating miRNAs role as biomarkers in breast cancer cases. Biomarkers hsa-mir-139 and hsa-mir-21 are selected as dysregulated diagnostic biomarker for breast cancer during an integrated study of 1100 cases [19].

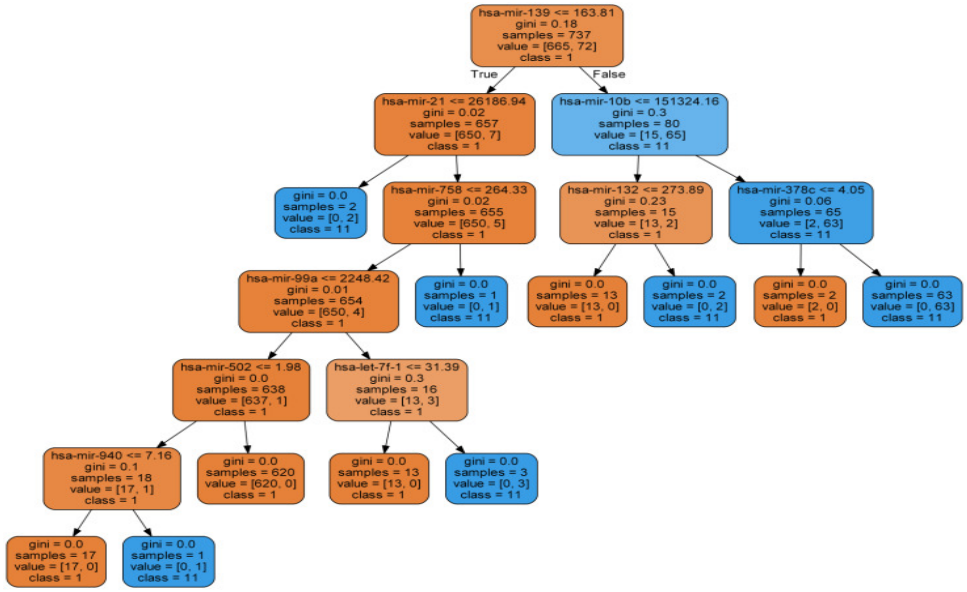

Fig. 4. Descriptive Decision Tree for Classifier of Breast Cancer Status using 465 attributes.

The study confirms that experimental (i.e PCR, microarray) evidences suggests that hsa-mir-21 function for regulating several tumour suppressor directly and hsa-mir-139 is part of metastatis-related pathways. The work of [20] suggests that hsa-mir-139 is dysregulated in different types of cancer including breast cancer. Tumour growth can be restricted by loss of hsa-mir-139-5p and serve as biomarker. The hsa-mir-10b is considered as one of the first miRNA to identity as having abnormal expression in many cancers including breast cancer [21].

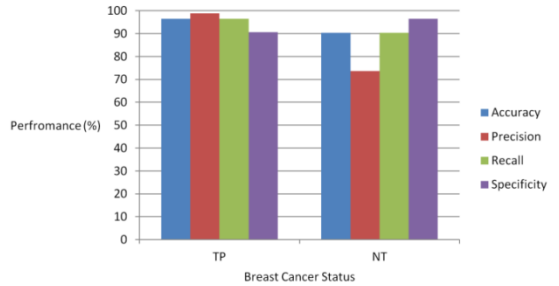

Fig. 5. Decision Tree Classifier based Performance for Breast Cancer Status using 1046 attributes.

Table 4: Quantitative Importance of Biomarkers predicted using Decision Tree Classifier trained on 465 attributes and references validating role of miRNA as Biomarker.

Biomarker(miRNA)	Quantitative Importance (%)	References
hsa-mir-139	70.58	[19, 20, 25]
hsa-mir-10b	13.10	[21-23]
hsa-mir-3677	3.75	[24]
hsa-mir-21	3.02	[19, 24]
hsa-mir-378c	2.98	—
hsa-mir-432	2.66	—
hsa-mir-758	1.51	—
hsa-mir-3074	1.45	—
hsa-mir-99a	0.8	[24]
hsa-mir-502	0.08	—

Biomarker hsa-mir-10b can contribute for metastatis-related pathways for stopping spread of the tumour [22]. Zhang *et al.,* suggests, hsa-mir-10b may be a biomarker for breast cancer and its potential target for clinical treatment. Their analysis is based on clinical data processed using Real time reverse transcription-PCR. Their work concludes that expression level of hsa-mir-10b correlates with disease stage and tumour size [23]. miRNAs hsa-mir-3677, hsa-mir-21 and hsa-mir-99a are considered as differentially expressed miRNA based on

cancer sample and normal sample data [24]. A decision tree classifier is also built without considering sparseness present within data. This classifier has all 1046 miRNAs as its attributes. The Fig. 5 shows the performance achieved on test data. The Fig. 5 shows that when all attributes are included performance reduces in terms of all performance evaluation measures. This may be due to sparseness present within data.

The tree generated for this prediction is shown in Fig. 6. It displays potentially important biomarkers (i.e miRNAs) used while deciding thresholds for prediction.

Table 5 lists each biomarker with respective quantitative importance. Table 5 also includes recent references validating miRNAs' role as biomarkers in breast cancer. Classification model of 1046 attributes also uses miRNAs hsa-mir-139, hsa-mir-10b and hsa-mir-21 for deciding thresholds. The study of [26] identifies hsa-mir-184 as a candidate breast cancer suppressor. It may help in suppressing cell proliferation and delaying the formation of metastatic lesions in vivo. Various works have established that mir-184 is lowly expressed in various malignancies. The work of [27] suggests that mir-184 along with some other miRNAs play crucial role in breast cancer tissues and small-cell lung cancer tissues. These results are based on microarray and quantitative real-time PCR experiments. It is reported that hsa-mir-92a is part of cluster which is reported earlier for links with cancers establishments. The studies based on *in situ* hybridization and related to clinico-pathological data associates down regulation of hsa-mir-92a with aggressive breast cancer features [28, 29]. It is identified, hsa-mir-33b restricts breast cancer metastatis and is down regulated in samples of breast cancers malignancy [30]. The experimental data suggested hsa-mir-33b works as an onco-suppressive miRNA in the progression of breast cancer. The miRNA hsa-mir-205 is also reported to work as breast cancer suppressor [31]. It is also identified, that hsa-mir-205 targets onco-genes which lead to resistance for targeted therapy [32].

There are miRNAs (hsa-mir-378c, hsa-mir-432, hsa-mir-758, hsa-mir-3074, hsa-mir-502, hsa-mir-1289-1, hsa-mir-1273c, hsa-mir-337, hsa-mir-483 and hsa-mir-454) whose involvement with breast cancer is not reported so far but used by tree classifier.

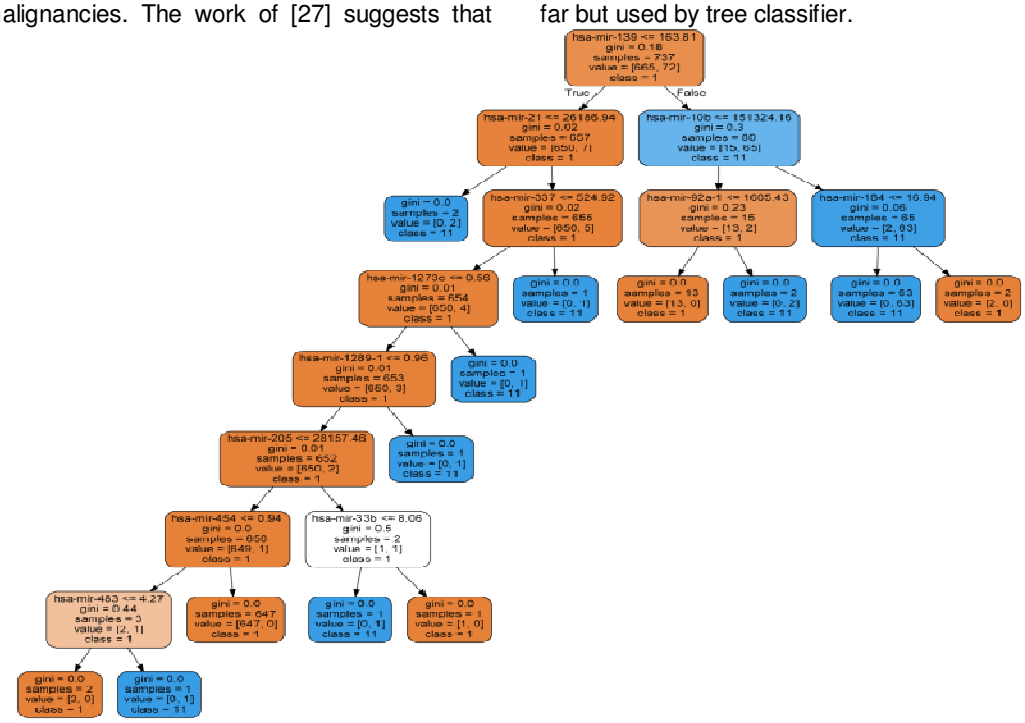

Fig. 6. Descriptive Decision Tree for Classifier of Breast Cancer Status using 1046 attributes.

Table 5: Quantitative Importance of Biomarkers predicted using Decision Tree Classifier trained on 1046 attributes and references validating role of miRNA as Biomarker.

Biomarker(miRNA)	Quantitative Importance (%)	References
hsa-mir-139	70.58	[19, 20, 25]
hsa-mir-10b	13.10	[21-23]
hsa-mir-21	3.02	[19, 24]
hsa-mir-184	2.98	[26, 27]
hsa-mir-92a-1	2.66	[28, 29]
hsa-mir-1289-1	1.53	—
hsa-mir-1273c	1.52	—
hsa-mir-337	1.51	—
hsa-mir-483	1.02	—
hsa-mir-33b	0.77	[30]
hsa-mir-205	0.76	[31, 32]
hsa-mir-454	0.51	—

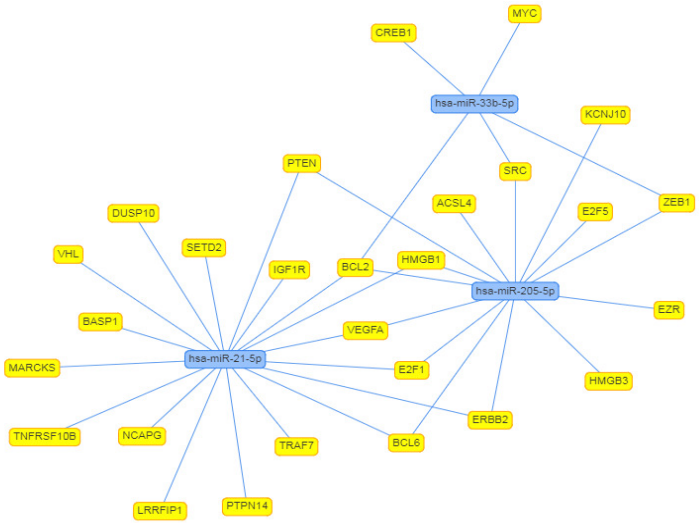

Fig. 7. Network of miRNA and genes prepared using MIENTURNET.

It is known that one miRNA may be able to target multiple gene products. Also, one gene product can be targeted by multiples miRNAs. Thus, miRNAs are involved in regulating the expression of many genes. Current work studies miRNA-target interactions using MIENTURNET [33]. A network of miRNA and target is created using only strong evidences from miRTarBase [34] in MIENTURNET and visualized in Fig. 7. It shows gene products targeted by miRNAs. It is observed that hsa-mir-205-5p and hsa-mir-21-5p has maximum number of interactions.

Further, functional enrichment for miRNA target genes is performed for hsa-mir-205-5p, hsa-mir-21-5p and hsa-mir-33b-5p using Disease Ontology [35] in MIENTURNET. It shows significant correlation with breast cancer and other diseases based on adjusted p-value in Fig. 8.

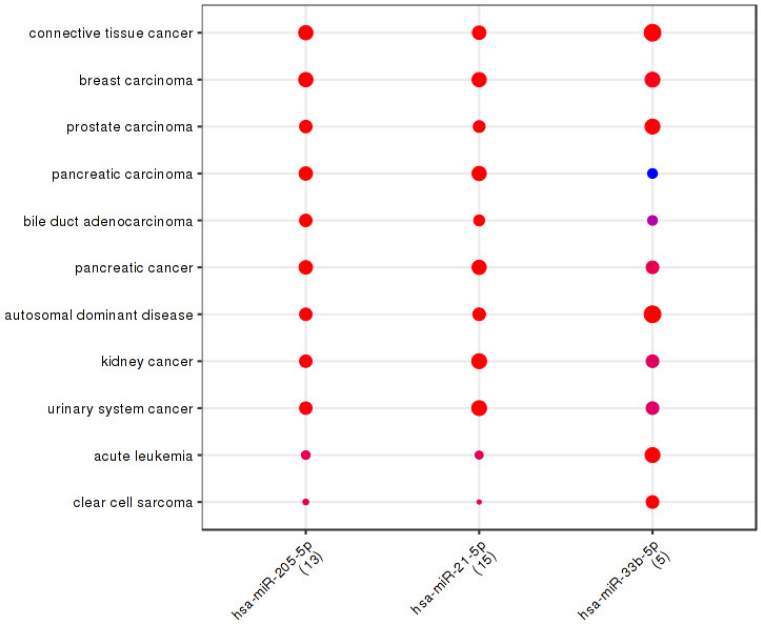

Fig. 8. Functional enrichment results from Disease Ontology in MIENTURNET.

The biomarkers identified as miRNA can help perform target prediction. Through the integrative modelling of miRNA binding and target network functional enrichment can be performed. This can potentially help to computationally find drug-disease relationship based on miRNA data.

IV. CONCLUSION

In this work, the decision tree classifier is used for discovering important miRNA based biomarkers for breast cancer. The necessary (i.e. miRNA) data for experiment is taken from TCGA using TCGAbiolinks package of R through GDC.

The experiment is performed on 1054 legacy miRNA quantification samples. Each miRNA sample contains 1046 miRNAs (i.e features). The reads_per_million record from each miRNA sample is taken as a value for each input attribute. The barcode structure is utilized to distinguish among positive and negative breast cancer samples. Sparseness within data is handled by removing attributes with more than 50% sparse data. Python's scikit-learn library is used for decision tree classifier based on gini index. Accuracy of better than 97% is achieved during binary classification of breast cancer sample and non-breast cancer sample. miRNAs playing pivotal role in binary classification are identified and validated using peer-reviewed research articles. The common miRNAs are used for functional enrichment using MIENTURNET. The functional enrichment results from disease ontology assert that hsa-mir-205, hsa-mir-21 and hsa-mir-33b are associated with breast cancer.

V. FUTURE SCOPE

The current work can be extended in terms of short term goals and long term goals. Following are the immediate extension possible.

– To find miRNA biomarkers for other diseases data available at TCGA.

– To find mutational and other genomic biomarkers for breast cancer based on TCGA data.

Following are the research works which will need more resources and expertise for extending current work.

– Establishing a recommender system, predicting and suggesting a possible malignancy based on miRNA quantification data.

– Finding drug which can be used to target breast cancer based on miRNA and target relation found.

ACKNOWLEDGEMENTS

We are grateful to Research and Development Center and Computer Engineering department of Faculty of Technology, Dharmsinh Desai University for supporting us while conducting our research.

Conflict of Interest. The authors declare no conflict of interest.

REFERENCES

[1]. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394-424.

[2]. Ghoncheh, M., Pournamdar, Z., & Salehiniya, H. (2016). Incidence and mortality and epidemiology of breast cancer in the world. *Asian Pacific Journal of Cancer Prevention*, *17*(S3), 43-46.

[3]. Tao, Z., Shi, A., Lu, C., Song, T., Zhang, Z., & Zhao, J. (2015). Breast cancer: epidemiology and etiology. *Cell biochemistry and biophysics*, *72*(2), 333-338.

[4]. Krol, J., Loedige, I., & Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews Genetics*, *11*(9), 597-610.

[5]. Ha, T. Y. (2011). The role of microRNAs in regulatory T cells and in the immune response. *Immune network*, *11*(1), 11-41.

[6]. Rupaimoole, R., & Slack, F. J. (2017). MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nature reviews Drug discovery*, *16*(3), 203.

[7]. Nelson, K. M., & Weiss, G. J. (2008). MicroRNAs and cancer: past, present, and potential future. *Molecular Cancer Therapeutics*, *7*(12), 3655-3660.

[8]. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, *19*(1A), A68.

[9]. Sherafatian, M., & Arjmand, F. (2019). Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data. *Oncology letters*, *18*(2), 2125-2131.

[10]. Sathipati, S. Y., & Ho, S. Y. (2018). Identifying a miRNA signature for predicting the stage of breast cancer. *Scientific reports*, *8*(1), 1-11.

[11]. van IJzendoorn, D. G., Szuhai, K., Briaire-de Bruijn, I. H., Kostine, M., Kuijjer, M. L., & Bovée, J. V. (2019). Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS computational biology*, *15*(2), e1006826.

[12]. Yang, J., Ma, D., Fesler, A., Zhai, H., Leamniramit, A., Li, W., ... & Ju, J. (2017). Expression analysis of microRNA as prognostic biomarkers in colorectal cancer. *Oncotarget*, *8*(32), 52403.

[13]. Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, *375*(12), 1109-1112.

[14]. Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Garolini, D., Cava, C., ... & Ceccarelli, M. (2019). Package 'TCGAbiolinks'.

[15]. Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., ... & Ceccarelli, M. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, *44*(8), e71-e71.

[16]. Breiman, L., Friedman, J., Olshen, R., & Storne, C. (1984). Classification and regression trees. Belmont, CA: Wadsworth International Group; 1984. *Google Scholar*.

[17]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

[18]. Apurva, M., & Mazumdar, H. (2020). Predicting structural class for protein sequences of 40% identity based on features of primary and secondary structure using Random Forest algorithm. *Computational Biology and Chemistry*, *84*, 107164.

[19]. Xiong, D. D., Lv, J., Wei, K. L., Feng, Z. B., Chen, J. T., Liu, K. C., ... & Luo, D. Z. (2017). A nine-miRNA signature as a potential diagnostic marker for breast carcinoma: An integrated study of 1,110 cases. *Oncology reports*, *37*(6), 3297-3304.

[20]. Huang, L. L., Huang, L. W., Wang, L., Tong, B. D., Wei, Q., & Ding, X. S. (2017). Potential role of miR-139-5p in cancer diagnosis, prognosis and therapy. *Oncology letters*, *14*(2), 1215-1222.

[21]. Sheedy, P., & Medarova, Z. (2018). The fundamental role of miR-10b in metastatic cancer. *American journal of cancer research*, *8*(9), 1674.
[22]. Ma, L. (2010). Role of miR-10b in breast cancer metastasis. *Breast cancer research*, *12*(5), 210.
[23]. Zhang, J., Yang, J., Zhang, X., Xu, J., Sun, Y., & Zhang, P. (2018). MicroRNA-10b expression in breast cancer and its clinical association. *PloS one*, *13*(2).
[24]. Cheng, D., He, H., & Liang, B. (2018). A three-microRNA signature predicts clinical outcome in breast cancer patients. *Eur. Rev. Med. Pharmacol. Sci*, *22*, 6386-6395.
[25]. Pajic, M., Froio, D., Daly, S., Doculara, L., Millar, E., Graham, P. H., ... & Zaratzian, A. (2018). miR-139-5p modulates radiotherapy resistance in breast cancer by repressing multiple gene networks of DNA repair and ROS defense. *Cancer research*, *78*(2), 501-515.
[26]. Phua, Y. W., Nguyen, A., Roden, D. L., Elsworth, B., Deng, N., Nikolic, I., ... & Cowley, M. J. (2015). MicroRNA profiling of the pubertal mouse mammary gland identifies miR-184 as a candidate breast tumour suppressor gene. *Breast Cancer Research*, *17*(1), 83.
[27]. Zhou, R., Zhou, X., Yin, Z., Guo, J., Hu, T., Jiang, S., ... & Wu, G. (2015). Tumor invasion and metastasis regulated by microRNA-184 and microRNA-574-5p in small-cell lung cancer. *Oncotarget*, *6*(42), 44609.
[28]. Nilsson, S., Möller, C., Jirström, K., Lee, A., Busch, S., Lamb, R., & Landberg, G. (2012). Downregulation of miR-92a is associated with aggressive breast cancer features and increased tumour macrophage infiltration. *PloS one*, *7*(4).
[29]. Smith, L., Baxter, E. W., Chambers, P. A., Green, C. A., Hanby, A. M., Hughes, T. A., ... & Speirs, V. (2015). Down-regulation of miR-92 in breast epithelial cells and in normal but not tumour fibroblasts contributes to breast carcinogenesis. *PLoS One*, *10*(10).
[30]. Lin, Y., Liu, A. Y., Fan, C., Zheng, H., Li, Y., Zhang, C., ... & Luo, Q. (2015). MicroRNA-33b inhibits breast cancer metastasis by targeting HMGA2, SALL4 and Twist1. *Scientific reports*, *5*, 9995.
[31]. Elgamal, O. A., Park, J. K., Gusev, Y., Azevedo-Pouly, A. C. P., Jiang, J., Roopra, A., & Schmittgen, T. D. (2013). Tumor suppressive function of mir-205 in breast cancer is linked to HMGB3 regulation. *PloS one*, *8*(10).
[32]. De Cola, A., Volpe, S., Budani, M. C., Ferracin, M., Lattanzio, R., Turdo, A., & Di Ilio, C. (2015). miR-205-5p-mediated downregulation of ErbB/HER receptors in breast cancer stem cells results in targeted therapy resistance. *Cell Death Dis., 6*(7), e1823--e1823.
[33]. Licursi, V., Conte, F., Fiscon, G., & Paci, P. (2019). MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis. *BMC bioinformatics*, *20*(1), 1-10.
[34]. Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., ... & Chiew, M. Y. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic acids research*, *46*(D1), D296-D302.
[35]. Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., ... & Bisordi, K. (2019). Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, *47*(D1), D955-D962.

How to cite this article: Mehta, A. A. and Mazumdar, H. S. (2020). Discovery of Significant miRNA-biomarkers for Breast Cancer using Decision Tree Classifier. *International Journal on Emerging Technologies*, *11*(2): 453–460.