Exploring microRNA Regulation of Cancer with Context-Aware Deep Cancer Classifier

Blake Pyman†, Alireza Sedghi, MSc, Shekoofeh Azizi, PhD, Kathrin Tyryshkin, PhD, Neil Renwick, MD, PhD, Parvin Mousavi, PhD

School of Computing, Queen's University,
Kingston, Ontario K7L 3N6, Canada
†*E-mail: pyman@cs.queensu.ca*
http://www.queensu.ca/

Background: MicroRNAs (miRNAs) are small, non-coding RNA that regulate gene expression through post-transcriptional silencing. Differential expression observed in miRNAs, combined with advancements in deep learning (DL), have the potential to improve cancer classification by modelling non-linear miRNA-phenotype associations. We propose a novel miRNA-based deep cancer classifier (DCC) incorporating genomic and hierarchical tissue annotation, capable of accurately predicting the presence of cancer in wide range of human tissues.

Methods: miRNA expression profiles were analyzed for 1746 neoplastic and 3871 normal samples, across 26 types of cancer involving six organ sub-structures and 68 cell types. miRNAs were ranked and filtered using a specificity score representing their information content in relation to neoplasticity, incorporating 3 levels of hierarchical biological annotation. A DL architecture composed of stacked autoencoders (AE) and a multi-layer perceptron (MLP) was trained to predict neoplasticity using 497 abundant and informative miRNAs. Additional DCCs were trained using expression of miRNA cistrons and sequence families, and combined as a diagnostic ensemble. Important miRNAs were identified using backpropagation, and analyzed in Cytoscape using iCTNet and BiNGO.

Results: Nested four-fold cross-validation was used to assess the performance of the DL model. The model achieved an accuracy, AUC/ROC, sensitivity, and specificity of 94.73%, 98.6%, 95.1%, and 94.3%, respectively.

Conclusion: Deep autoencoder networks are a powerful tool for modelling complex miRNA-phenotype associations in cancer. The proposed DCC improves classification accuracy by learning from the biological context of both samples and miRNAs, using anatomical and genomic annotation. Analyzing the deep structure of DCCs with backpropagation can also facilitate biological discovery, by performing gene ontology searches on the most highly significant features.

Keywords: Deep learning; miRNA; Autoencoder; Cancer classification; PSB

1. Introduction

Following rapid advances in biotechnology (RNA-Seq) and machine learning, mining of high-resolution transcriptomic data has become a promising tool for the discovery of potential RNA cancer biomarkers.[1] However, the ability to use this high-dimensional data to predict cancer is limited by the tendency of large models to overfit available data, known as the curse of dimensionality.[2] This problem can be mitigated by filtering variables, and reducing the dimensionality of input, techniques that can be incorporated in machine learning algorithms.

Deep learning (DL) describes a family of machine learning algorithms designed to model non-linear features at various levels of abstraction, by processing training data over multiple connected layers. DL models are a type of artificial neural network (ANN) which learn by calibrating the weights of connections between nodes by backpropagation of the error gradient. Applying backpropagation to deeper networks can be ineffective due to the problem of "vanishing gradients", but this problem was solved in 2006 by Hinton and Salakhutdinov, who devised a procedure for pre-training hidden layers.[3] Hinton's original formulation used stochastic, binary networks with one hidden layer and symmetrical weights, known as Restricted Boltzmann Machines (RBMs). RBMs were pre-trained such that the hidden layer of one RBM formed the input of the next. After pre-training, the entire model (named a Deep Belief Network, or DBN) could be fine-tuned with supervised learning. Because each layer encodes features based on the previous layer, the higher layers contain increasingly abstract feature sets. In addition, the non-linearity of deep learning results in highly generalizable models that are less dependent on preprocessing and normalization. Both of these characteristics - complex internal structure, and insensitivity to input variance - makes DL models well-suited to transcriptomic applications. DBNs have been used with microarray data to cluster breast cancers[4] and glioblastomas into prognostically relevant subtypes.[5] Deep Boltzmann Machines, a related architecture, have also been used to classify human colorectal carcinomas by subtype.[6] In each of these studies, training data was limited to a single cancer type, permitting subtype discrimination but limiting the scope for transfer learning between cancers. Deep neural nets have also been used to predict cancer type based on genetic data (somatic point mutations) albeit with poorer accuracy than RNA-based models.[7]

The pre-training method applied to DBNs can be generalized for layers with continuous outputs,[8] known as autoencoders, which recreate their input using a single real-valued hidden layer and non-symmetrical weights. Autoencoders can be stacked and pre-trained in a manner analogous to DBNs, and the resulting stacked (or deep) autoencoders (SAE/DAE) can be fine-tuned using supervised learning. It is possible to limit overfitting by imposing a constraint on the sparsity (number of active nodes) of autoencoders in an SAE. Recently, stacked sparse autoencoders have shown promising results classifying cancer sub-types[9,10]

Most studies applying deep learning to RNA-based cancer prediction have focused on the familiar protein-coding variety, mRNA. However, at least 15 types of non-coding RNA are also produced (accounting for approx. 98% of nuclear output), including potentially valuable biomarkers such as microRNAs.[11] microRNA (or miRNA) are a small non-coding class of RNA responsible for post-transcriptional repression of mRNAs. In contrast to over 22,000 mRNA in the human genome, the high-confidence set of miRNAs is limited to just over one thousand.

This relatively smaller input space mitigates the effects of the curse of dimensionality in DL models. In addition, miRNAs individually display much greater tissue- and tumour-specificity than mRNAs, perhaps due to their role as upstream regulators of RNA activity.[12]

To date, most miRNA DL applications have focused on diagnostic or prognostic classification of tumour subtypes. In one study, DBNs were used to select miRNAs to classify six tumour types.[13] Other projects have examined so-called multimodal architectures integrating miRNA expression with other data sources. One study fed a combination of miRNA, mRNA and gene methylation data to a DBN to cluster ovarian and breast cancer samples.[14] Another study combined the same inputs using a 3-layer stacked autoencoder to predict survival time in liver cancer.[15] Combining data from multiple sources may improve results, at the cost of increasing complexity and the risk of overfitting. Instead, the proposed DCC is supplied with concise data situating both samples and miRNAs in a biological context enabling comparisons between related samples and miRNAs.

In this paper, we propose a DL model to predict the presence of cancer based on miRNA sequencing across over 30 human tissues, from approximately 3600 patients, using an ensemble of deep autoencoders. The proposed model leverages the biological context of both samples and sequences. First, hierarchical anatomical annotation was used to score and filter miRNAs based on their information content. In addition, annotation of miRNA cistrons and sequence families were used to create variants of input data, used to train ensembles of classifiers with superior performance than any single component. The DCC also goes beyond cancer classification, by identifying significant miRNAs with backpropagation, and exploring with network visualization. Finally, targets of selected miRNAs were analyzed using gene ontology, to provide insight into the biological nature of the selected miRNAs' association with cancer.

2. Data

Samples were sequenced between September 2008 and December 2015 at Rockefeller University using the Illumina HiSeq. Samples were richly annotated by expert clinicians using over 30 features, including the type of biological material, disease state and anatomical site, as well as expression of 1187 miRNAs. Our original data included samples from body tissues, body fluids and cultured or sorted cells, from a wide array of anatomical sites (Fig. 2). All subsequent analyses were confined to tissue samples, due to greater average sequencing depth and balanced subclasses. Of tissue samples, 2026 (56%) were neoplastic, and 1606 (44%) were either normal or affected by an unrelated diseased. Site-of-origin was described at three different levels, namely organ, organ substructure, and cell type, organized into a 3-level anatomical hierarchy. The dataset includes samples from 26 organs, 6 organ sub-structures and 68 cell types. *Sequence families*: are sets of miRNAs defined on the basis of sequence similarities and represent miRNAs which are likely to share targets - due to the pleiotropy of miRNA targeting, they are likely to have overlapping sets of targeted mRNAs. *Precursor clusters*: are defined to include miRNAs that either share an identical mature form, or are clustered closely in the genome, and are likely to be co-expressed due to shared promoters. Because of this fact, they may be up- or down-regulated in concert, which may indicate involvement in shared (patho-)physiological pathways.[16]

Table 1. Organs represented by the largest number of samples, with number of samples from each.

Source organ	Number
C, SC and other soft tissues	913
Breast	762
Thyroid gland	333
Brain	292
Skin	269
Kidney	244
Hematopoietic and RC system	173
Heart, mediastinum, and pleura	141

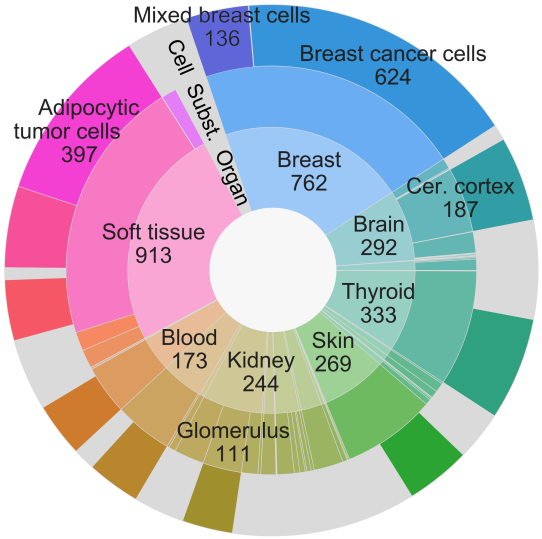


Fig. 2. Left: Eight organs representing the greatest number of tissue samples. Right: Sunburst diagram depicting tissue samples at three anatomical levels: (from the centre outwards) organ, organ substructure and cell type. Only cell types with greater than 100 samples are shown. C = connective, SC = subcutaneous, RC = reticuloenothelial.

2.1. *Preprocessing*

Outlier Removal: We used the inter-quartile range (IQR) to label and subsequently remove outliers and batch effects .[17] Upper and lower bounds were established at a certain distance below the first quartile and above the third quartile of the data and we measured the distance of outlier points beyond the bulk of the data. The distance is usually set at a multiple of the IQR; 1.5×IQR was suggested in previous work and that is the value used herein.[17]

Batch effects were identified using median Spearman coefficient and the bounds established by the IQR method. Batches were removed if at least half their samples were flagged by the IQR method. Following the removal of batches, the Spearman correlations of the remaining points were recalculated. Removal of samples with extreme Spearman values results in tighter bounds, which may enable the detection of further outliers and batch effects. This process was performed iteratively until batch effects could no longer be identified.

Filtering: The initial set of 1187 human miRNAs was filtered based on abundance of expression, eliminating the lowest-expressed miRNAs. An expression threshold was set at 1.41 x 10^{-5}, which corresponds to 14 counts in one million. Any miRNA that was not expressed at or above this level in at least 1% of samples was removed. 397 miRNAs were filtered out on account of low-expression. The remaining miRNAs were also filtered on the basis of information content or cancer specificity.

The specificity of an miRNA, m, for a given tissue, t, ($G_{m,t}$) is a measure of the relative expression level of m in t, compared to other tissues.$G_{m,t}$ can be understood as the proportion of total m expression in all samples that would occur in t, if all classes were sampled from

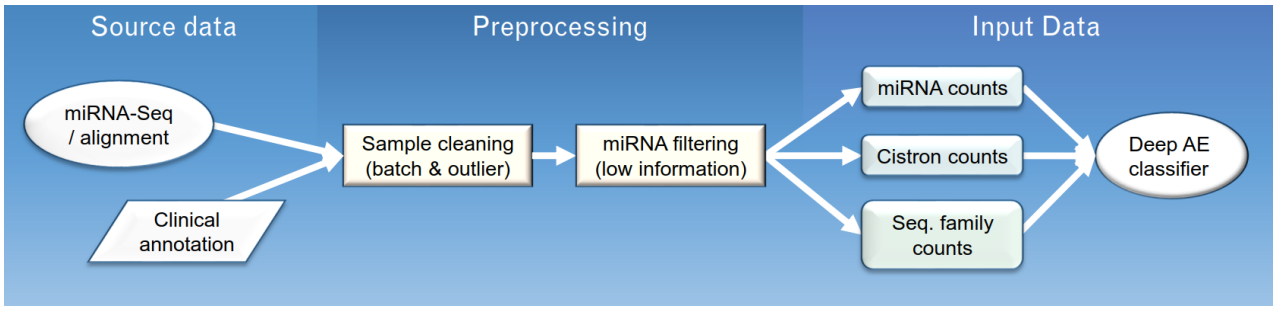


Fig. 1. Flowchart illustrating the datasets used (rounded boxes) and the various transformations and preprocessing steps (rectangular boxes) applied to it, before it is fed as input in the the deep autoencoder model.

equally. $G_{m,t}$ is a value ranging from 0 to 1, representing the specificity of miRNA m for tissue t. If $G_{m,t}$ is close to 1, this indicates that m is expressed much more in t than in other tissues. The specificity of m is determined by the distribution of $G_{m,t}$ for different tissues. If $G_{m,t}$ were the same for all values of t, then the specificity of m would be 0. Once $G_{m,t}$ is determined for all m and t, the total information content of a miRNA, s_m, can be calculated using all $G_{m,t*}$ terms of m for all tissues $t*$.

$$s_m = \log_2(\#oftissues) + \sum_t (G_{m,t} * \log_2(G_{m,t})) \quad (1)$$

Therefore, the maximum possible specificity (i.e. a miRNA expressed solely in one class) is $\log_2(\#ofclasses)$, and cancer specificity of miRNAs ranges from 0 to 1. miRNAs that did not meet a minimum information content threshold ($s_m > 0.01$) were excluded. The remaining 497 miRNAs were inputs to the DCC to predict the presence of cancer (Fig. 1).

Normalization: We used Total Counts Scaling (TCS), in which read counts are divided by the total number of sequenced counts, known as the sequencing depth for normalization.[18] TCS was preferred to more complex methods due to its widespread use and ease of interpretation.[18]

3. Deep Cancer Classifier

The proposed deep cancer classifier (DCC) merges stacked autoencoders with a multilayer feedforward network to accurately classify cancer using miRNA in a range of human tissues (Fig. 2). Data (miRNA, cistron and sequence family expression) was presented to DCC via the input layer. Each successive autoencoder layer is smaller than the last, the layer sizes forming a geometric series. By training each autoencoder in the usual unsupervised manner (minimizing mean squared error with respect to input) it is possible to represent abstract, latent features in the data. These latent features were repeatedly transformed and compressed to 20 in the third AE layer. Following pre-training, the weights of each AE layer were initialized with the weights of the corresponding hidden layers, the AE layers were joined together, and a feed-forward MLP was added. After this step the DCC undergoes supervised learning to boost its classification performance. The model now uses the complex latent features learned in the first stage to predict the presence of cancer. Weights throughout the entire network were fine-tuned through backpropagation to minimize cross-entropy loss of predictions. The proposed

multi-modal architecture of DCC allows for learning multiple layers of latent features from miRNA expression while integrating expert clinical annotation. After developing the model using miRNA profiles using training data, the DCC's performance was tested on left-out samples, and benchmarked against other popular ML methods.

3.1. *Training & testing*

Of the initial 2518 tissue samples, 40% were used as a development set to tune model parameters, while the other 60% were set aside to provide an unbiased measure of the model's performance. This selection was stratified with respect to both cancer status and organ type. Five-fold cross-validation (CV) was performed on the development set, requiring each fifth of the data to provide validation for a model trained on the other four fifths. Once the model's hyperparameters were tuned, the previously unseen test set was used to assess its performance.

An ensemble of development models was used to maximize test set performance (Fig. 2). Each model predicts cancer status as a probability and the output probabilities were averaged over models from different cross-validations. Model variants based on the three input sources (miRNA, cistron and sequence family) were combined in the same way, so each test set prediction was based on an ensemble of $3 \times 5 = 15$ classifiers. These values were finally rounded to 0 (non-neoplastic) or 1 (neoplastic) to establish the number of true and false predictions for each type, and by extension the model's accuracy, sensitivity and specificity.

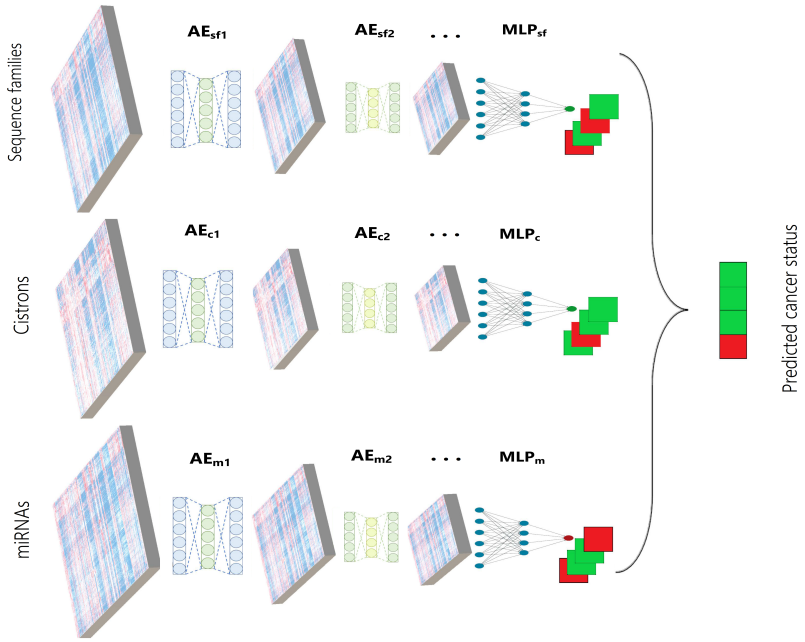


Fig. 2. Simplified schematic of the network topology with three input types (expression of individual miRNAs; miRNA cistron profiles; miRNA sequence family profiles) and main phases in the model's operation. First the AEs undergo unsupervised pre-training, each layer recreating the hidden layer activations of the last, following which the entire model is fine-tuned to classify cancer.

3.2. *Parameter tuning*

Deep learning models are characterized by a large number of configurable parameters, including the overall layout (topology) of the network, initial weights for connections between nodes, and various functions determining the way the network learns (i.e. how weights are updated). By using CV on the development set, different model configurations can be compared, allowing select parameters to be optimized. Even when working exclusively on the development set, there is a risk of overfitting the model if too many parameters are tuned to the development set, reducing the test performance. Therefore, a small number of significant parameters were selected for optimization. Namely, the size the latent features produced by the deep autoencoders, the number of stacked AE layers, the optimizer function to training the autoencoder, and an analogous optimizer for training the classifier.

The size of the smallest AE layer, also called the encoding size, determines the level of compression the input must undergo, and the amount of information available to the classifier. The size of the other AE layers were chosen to form a geometric series, the size of each layer decreasing by a constant factor between the input and final compressed form. The number of AE layers affects the amount of information learned in a different way. Each layer tends to represent different latent features in the data, so deeper networks can capture a greater number of more complex features. The downsides to increased network depth include the risk of overfitting, as well as potentially long training times.[19] Optimizers are algorithms which control the way weights are updated during training, and may control parameters such as (initial) learning rate, momentum and others. It is typically easier and more effective to use these thoroughly tested configurations, instead of varying these parameters independently. Because the performance of any single run is affected by random occurrences (e.g. the splitting of samples, random initialization of weights), the CV optimization procedure was repeated 30 times for each value of each parameter. Based on the distribution of scores for each configuration, the *Kruskal-Wallis (K-W)* test was applied to detect significant differences between groups. If a difference was detected, pairs of samples were compared using the *Mann-Whitney U* test to determine which samples were involved. As the *K-W* and *Mann-Whitney U* tests are non-parametric, no assumptions were made about the normality of the underlying distribution.

3.3. *Feature importance*

The importance of individual miRNAs (or cistrons, etc.) for cancer classification can be estimated using backpropagation, the same algorithm used to train models in supervised learning. However, rather than the gradient of error, we calculated the gradient of activation across input nodes.[20] Signed activation gradients can be computed for every edge between nodes. By taking the sum of the absolute values of activation gradients for all edges connected to a given node, the "contribution" of input features to the activation of higher nodes was determined. A distinct activation is produced in response to each batch of samples presented as input. To calculate the average activation gradient across input features, five DCC variants were trained using 5-fold cross validation. Then, the test set was presented to each variant in 16-sample batches, and the input activation gradients were recorded for each batch. Finally, the gradients were averaged across batches and CV variants, producing a single score for each input feature

(e.g. miRNA) representing its relative importance for classifying cancer.

The miRNAs with the greatest putative cancer association were validated using a network analysis tool called the Integrated Complex Traits Network (iCTNet2).[21] iCTNet2 links numerous biological databases (miRNA, gene, protein, disease, etc.) allowing visualization of indirect associations. Having made a list of miRNAs known to regulate cancer-related genes, one may expect a degree of overlap between this set of cancer-related miRNAs and those returned by the backpropagation method described above. Comparing the number of miRNAs found in both sets to the number expected by chance alone will provide an estimate of level of "cancer enrichment" in the miRNA set produced by the analysis of feature contribution.

Targets of selected miRNAs can also be investigated using gene ontology (GO). BiNGO is a cytoscape app that illustrates gene oncologies as hierarchical networks, with nodes (representing processes) coloured to illustrate their level of enrichment.[22] Enrichment is calculated as over-representation relative to entire GO annotation. This is measured by a p-value, adjusted using the Benjamini & Hochberg correction.

4. Results and Discussion

4.1. *Model selection*

Classification accuracy of the DCC was strongly associated with minimum AE size at the lower end of the tested range. Increasing the encoded size from 5 to 20 caused a clear benefit, although further increases had a null or negative effect (Fig. 3). While compressing miRNA profiles to just five features was clearly sub-optimal, it was still sufficient to classify samples with 93.7% accuracy. A similar trend was observed in relation to layer number. Validation accuracy was greatest when using three stacked autoencoders. Additional AE layers seemed to increase training times, without any significant performance increases. The choice of model optimizers had a strong effect on performance for supervised learning, but pre-training appeared to be relatively insensitive to optimizer choice (at least between the 5 tested configurations). Adagrad exhibited marginally superior performance for pre-training, while Adam was the most effective algorithm for supervised classification via backpropagation (Accuracy = 0.948).

Table 3. Summary of key parameters, with optimal values

Parameter	Optimal value (Range)	Accuracy range
Encoding size	20 (5-60)	0.937 - 0.948
AE layers	3 (1-5)	0.933 - 0.949
AE Optimizer	Adagrad (*)	0.948 - 0.949
MLP Optimizer	Adam (*)	0.844 - 0.948

* Tested optimizers: SGD, RMSprop, Adagrad, Adadelta, Adam

4.2. *Classifier performance*

The most common and intuitive way of assessing the performance of a classifier is its accuracy, given by the number of true predictions over the total number of predictions. The true pre-

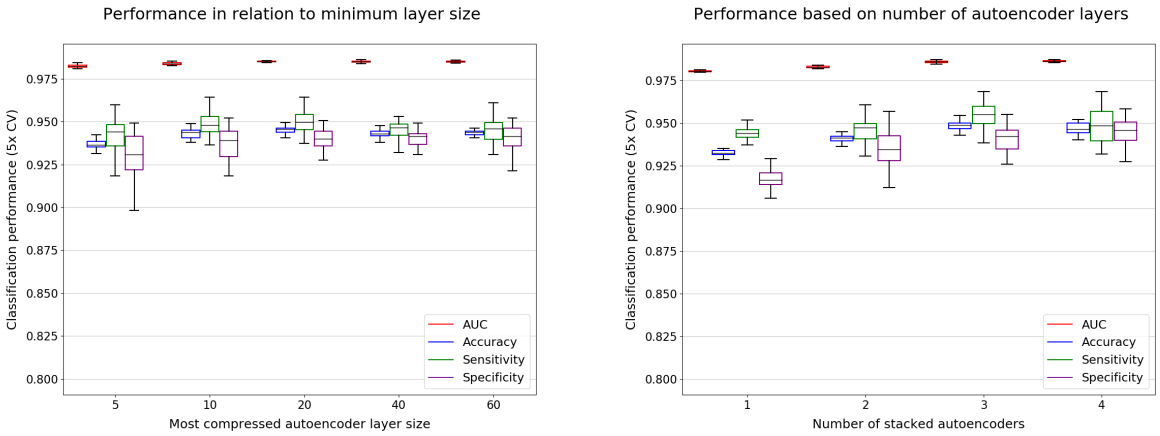


Fig. 3. Left: Test set AUC/ROC, accuracy, sensitivity and specificity in relation to the size of the latent feature representation produced by the stacked autoencoder (with 3 layers). Right: Test set performance (min. encoding size 20), with various numbers of stacked autoencoder layers.

dictions are the sum of the true positive (TP) and true negative (TN) predictions. Accuracy is a suitable metric for problems with similarly-sized target classes, but for highly imbalanced datasets, the success rate for positive and negative samples can be measured using sensitivity or specificity, respectively. Out of 1511 samples, the DCC was able to correctly classify 1421 of them (94.8%). The model had slightly better sensitivity (0.95) than specificity (0.94). The Receiver Operating Characteristic (ROC) illustrates the trade-off between Type I and Type II errors. The Area Under the Curve (AUC) was 0.985.

4.3. *Comparison with other methods*

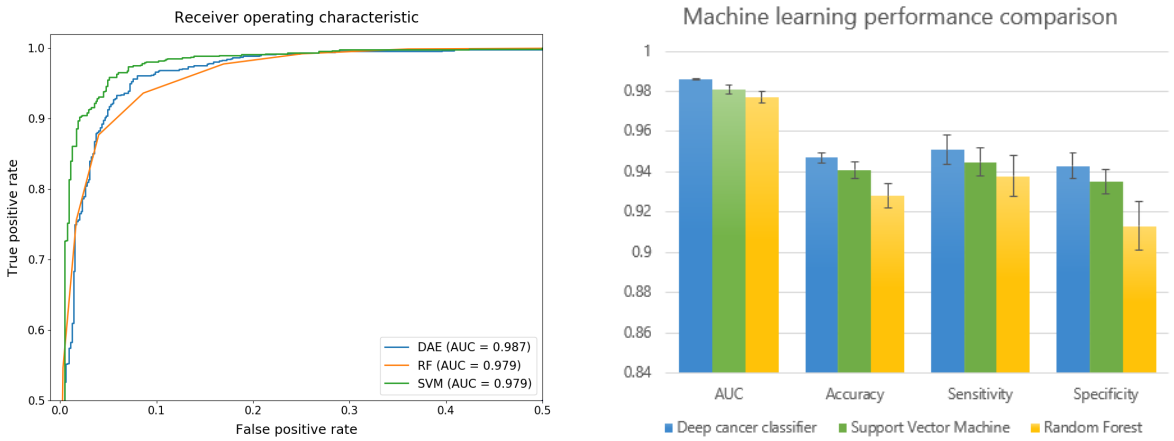


Fig. 4. Left: Receiver Operating Characteristic (ROC) graph showing performance at different thresholds, compared to random forests (RF) and support vector machines (SVM). Right: DCC performance compared with RF and SVM. Error bars show std. dev. for 30 trials, with 5xCV.

The model's performance on test set of 1510 samples was compared to the that of two well-known machine learning methods, support vector machines (SVM) and random forests (RF). The proposed DL model significantly surpassed the performance of a SVM with a linear kernal, and C=1. The example ROC on the following page shows that DCC outperforms SVM

and RF at every error threshold, with an average Area Under the Curve (AUC) of 0.9854.

4.4. *Feature importance*

Backpropagating the activation of the output node to the input nodes enables an estimation of the contribution each feature makes to the model's output. The top 20 miRNAs by output contribution are shown below (Fig. 5). The distribution of activation gradients across miRNAs was highly skewed; while the maximum gradient (for miR-21) was 22.9, only 26 of the 1187 features had an average activation gradient greater than 1. It would appear most of the information required to classify samples is concentrated in a small number of sequences.

iCTNet was used to link miRNAs, genes(/proteins) and human cancers, outputing a list of 61 miRNAs linked to cancer-associated genes (Fig. 5). iCTNet uses a different miRNA reference library (miRCat); after converting miRNAs to a common form and collapsing duplicates, the iCTNet network was reduced to 46 miRNAs, of which 44 were present in our reduced set of 582 miRNAs. Of the top 20 miRNAs by average output activation gradient, 8 were found in the list of cancer-linked miRNAs. Since 7.6% of the miRNAs in the larger list are present in the cancer-associated network, the expected number of matches (based on the binomial distribution) is just 1.5. Therefore, the backpropagation method returned a set of miRNAs with a cancer enrichment of $8/1.5 = 5.3$.

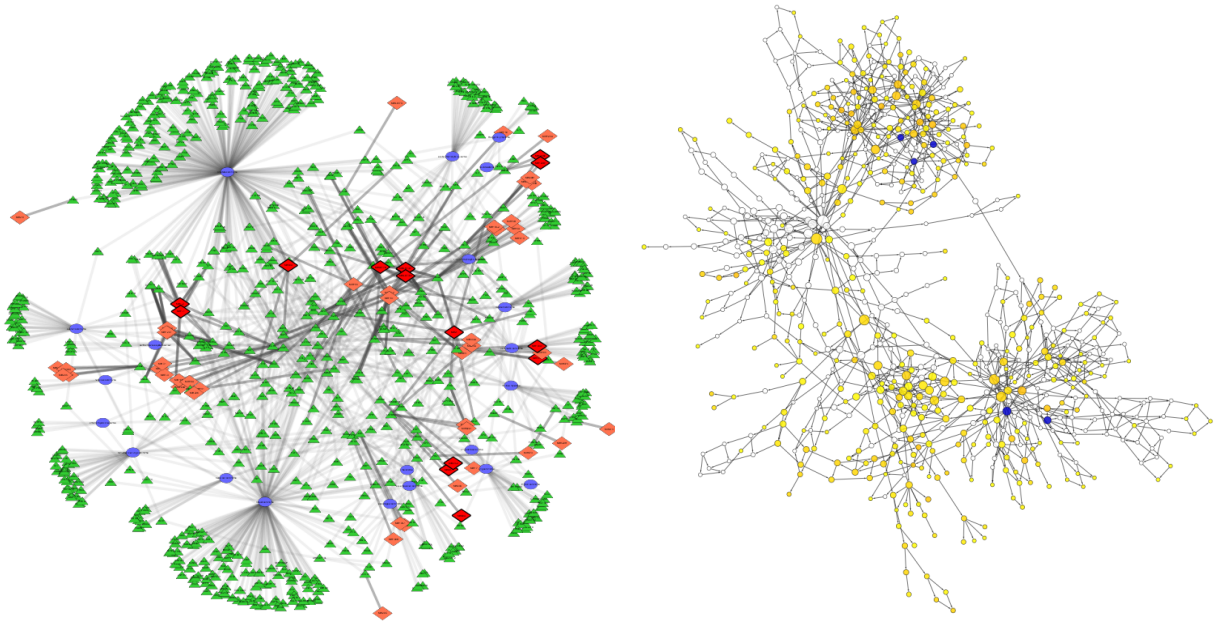


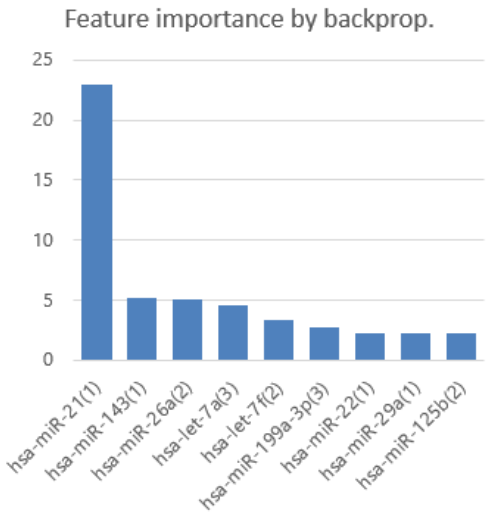
Fig. 5. Left: Network graph showing links between cancers (blue), genes/proteins (green) and miRNAs (red). Of the 44 cancer-linked miRNAs shown, 8 ranked in the top 20 by output activation gradient (bright red, outlined). Right: Gene ontology graph for genes targeted by selected miRNAs. Node color represents p-value of category over-representation. Five most over-represented categories highlighted in blue.

Sixteen miRNAs from the ICTNet graph are among the top 50 miRNAs by activation gradient. Gene ontology analysis was performed on 49 genes linked to these miRNAs (Fig. 5).

The graph displays a hierarchical GO network, containing at least two major clusters of related enriched biological processes. The bottom-right cluster stems from "Regulation of cellular process" and "Regulation of biological process", and contains "Positive regulation of cellular process" and "Regulation of cell proliferation". The other 3 of the top 5 most highly-enriched GO categories (Table 4) are found in the uppermost cluster, descended from "Developmental process", and "Multicellular organismal development".

Table 4. Most over-represented gene ontology categories linked to selected miRNAs

Gene ontology category	Adj. p-value	# of genes
Positive regul'n of cellular process	4.42E-5	22/49
Regulation of cell proliferation	3.60E-5	15/49
Epithelium development	4.42E-5	10/49
Tissue morphogenesis	1.84E-5	10/49
Morphogenesis of an epithelium	1.84E-5	9/49



5. Conclusion

The proposed deep cancer classifier is capable of diagnosing cancer in a wide range of human samples with almost 95% accuracy, which represents an improvement on conventional machine learning algorithms random forests and support vector machines. The model's performance is enhanced by exploiting two forms of contextual information, namely anatomical annotation of samples, and sequence annotation linking miRNAs to cistrons and sequence families. Once trained, the deep structure of the DCC can be interrogated for insights into the links between miRNAs and cancer. In particular, this enables the identification of miRNAs that may play serve as biomarkers or mediate the effects of cancers across diverse tissue types. The absolute activation gradient reveals a highly skewed distribution of feature importance, led by miR-21, a ubiquitous miRNA known to be disregulated in cancer.[23] This highly skewed feature importance distribution suggests the possibility of creating diagnostic arrays using small numbers of miRNAs. Gene ontology analysis of cancer-linked miRNAs identified multiple highly-enriched processes, some of which bare an obvious relationship to cancer (e.g. regulation of cellular proliferation) while others may indicate possible directions for future research.

References

1. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, *Machine Learning* (2002).
2. Y. Bengio, *Foundations and Trends® in Machine Learning* (2009).
3. G. E. Hinton and R. R. Salakhutdinov, *Science* (2006).
4. M. Khademi and N. S. Nedialkov, *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* , 727 (2015).
5. J. D. Young, C. Cai and X. Lu, *BMC Bioinformatics* **18** (2017).
6. A. F. Syafiandini, I. Wasito, S. Yazid, A. Fitriawan and M. Amien, *2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA)* , 108 (2016).
7. Y. Yuan, Y. Shi, C. Li, J. Kim, W. Cai, Z. Han and D. D. Feng, *BMC Bioinformatics* **17** (2016).
8. Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, Greedy Layer-Wise Training of Deep Networks, in *Proceedings of Neural Information Processing Systems, NIPS 2006*, (1)2006.
9. R. Fakoor, F. Ladhak, A. Nazi and M. Huber, *Proceeding of the 30th international conference on machine learning Atlanta, Georgia,USA* **28** (2013).
10. V. Singh, N. Baranwal, R. K. Sevakula, N. K. Verma and Y. Cui, *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016* , 1542 (2016).
11. J. S. Mattick, *EMBO Reports* **2**, 986 (2001).
12. N. Rosenfeld, R. Aharonov, E. Meiri, S. Rosenwald, Y. Spector, M. Zepeniuk, H. Benjamin, N. Shabes, S. Tabak, A. Levy, D. Lebanony, Y. Goren, E. Silberschein, N. Targan, A. Ben-Ari, S. Gilad, N. Sion-Vardy, A. Tobar, M. Feinmesser, O. Kharenko, O. Nativ, D. Nass, M. Perelman, A. Yosepovich, B. Shalmon, S. Polak-Charcon, E. Fridman, A. Avniel, I. Bentwich, Z. Bentwich, D. Cohen, A. Chajut and I. Barshack, *Nature Biotechnology* **26**, 462 (2008).
13. R. Ibrahim, N. A. Yousri, M. A. Ismail and N. M. El-Makky, *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* , 3957 (2014).
14. M. Liang, Z. Li, T. Chen and J. Zeng, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12**, 928 (2015).
15. K. Chaudhary, O. B. Poirion, L. Lu and L. X. Garmire, *Clinical Cancer Research* , p. clincanres.0853.2017 (2017).
16. P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foà, J. Schliwka, U. Fuchs, A. Novosel, R. U. Müller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H. I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan and T. Tuschl, *Cell* **129**, 1401 (2007).
17. J. W. Tukey, *Exploratory Data Analysis* 1977.
18. S. Tam, M. S. Tsao and J. D. McPherson, *Briefings in Bioinformatics* **16**, 950 (2015).
19. C. Angermueller, T. Pärnamaa, L. Parts and O. Stegle, *Molecular Systems Biology* **12**, p. 878 (2016).
20. S. Azizi, F. Imani, S. Ghavidel, A. Tahmasebi, J. T. Kwak, S. Xu, B. Turkbey, P. Choyke, P. Pinto, B. Wood, P. Mousavi and P. Abolmaesumi, *International Journal of Computer Assisted Radiology and Surgery* **11**, 947 (2016).
21. L. Wang, D. S. Himmelstein, A. Santaniello, M. Parvin and S. E. Baranzini, *F1000Research* , 1 (2015).
22. S. Maere, K. Heymans and M. Kuiper, *Bioinformatics* **21**, 3448 (2005).
23. T. A. Farazi, J. I. Hoell, P. Morozov and T. Tuschl, *Advances in Experimental Medicine and Biology* (2013).