Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers

Abdullah Al Mamun
School of Computing and Information Sciences
Florida International University
Miami, FL 33199
mmamu009@fiu.edu

Ananda Mohan Mondal
School of Computing and Information Sciences
Florida International University
Miami, FL 33199
amondal@fiu.edu

***Abstract*—Long noncoding RNA (lncRNA) plays key roles in tumorigenesis. Misexpression of lncRNA can lead to changes in expression profiles of various target genes, which are involved in cancer initiation and progression. So, identifying key lncRNAs for a cancer would help develop the cancer therapy. Usually, to identify key lncRNAs for a cancer, expression profiles of lncRNAs for normal and cancer samples are required. But, this kind of data are not available for all cancers. In the present study, a computational framework is developed to identify cancer specific key lncRNAs using the lncRNA expression of cancer patients only. The framework consists of two state-of-the-art feature selection techniques - Recursive Feature Elimination (RFE) and Least Absolute Shrinkage and Selection Operator (LASSO); and five machine learning models - Naive Bayes, K-Nearest Neighbor, Random Forest, Support Vector Machine, and Deep Neural Network. For experiment, expression values of lncRNAs for 8 cancers - BLCA, CESC, COAD, HNSC, KIRP, LGG, LIHC, and LUAD - from TCGA are used. The combined dataset consists of 3,656 patients with expression values of 12,309 lncRNAs. Important features or key lncRNAs are identified by using feature selection algorithms RFE and LASSO. Capability of these key lncRNAs in classifying 8 different cancers is checked by the performance of five classification models. This study identified 37 key lncRNAs that can classify 8 different cancer types with an accuracy ranging from 94% to 97%. Finally, survival analysis supports that the discovered key lncRNAs are capable of differentiating between high-risk and low-risk patients.**

***Index Terms*—Cancer Classification; Feature Selection; lncRNA expression; Machine Learning; Survival Analysis.**

I. Introduction

Recent studies indicate that several cancer risk loci are transcribed into lncRNAs and these transcripts play key roles in tumorigenesis [1], [2]. In their review paper, Cheetam at al. [1] enumerated that lncRNAs play key roles in cancer progression through a variety of mechanisms such as lncRNA ANRIL for remodeling of chromatin [3], H19 for transcriptional co-activation and co-repression [4], TERRA for protein inhibition [5], MALAT1 for post-transcriptional modifications [6] and PTENP1 for decoy [7]. LncRNA ANRIL, which causes PCR1-mediated repression of tumor suppressor locus INK4A-ARF-INK4b, is up-regulated in prostate cancer [3]. Similarly, H19 plays significant role in proliferation of gastric cancer cell due to its up-regulation [4]. TERRA facilitates telomeric heterochromatin formation [5]. MALAT1 induces migration and tumor growth in lung cancer [6]. PTENP1 controls the expression level of tumor supressor gene PTEN [7]. Also, lncRNAs have key functions in transcriptional, post-transcriptional, and epigenetic gene regulation [8]. Schmitt et al. discussed the impact of lncRNA in cancer pathway [9]. They described the involvement of lncRNAs in six hallmarks of cancer [10] such as proliferation, growth suppression, motility, immortality, angiogenesis, and viability.

While some researchers detailed the role of lncRNAs in cancer progression, other discovered a number of lncRNA biomarkers in several cancers by creating lncRNA-miRNA co-expression networks [11], [12], [13]. Wang et al. on the other hand identified 6 key lncRNAs for metastatic melanoma from a competing endogenous RNA (ceRNA) network analysis using mRNA, miRNA and lncRNA expression [14]. By constructing the similar network, Sui et al. found 41 lncRNAs biomarkers in human lung adenocarcinoma [15]. Also, Chen et al. identified 24 hub lncRNAs in smoking-associated lung cancer by forming protein-protein interaction (PPI) networks [16]. Similarly, Lanzos et al. identified cancer driver lncRNAs as new candidates and distinguishing features by analyzing the mutational patterns in tumor DNA [17]. Another model, CRlncRC used machine learning algorithms including RF, NB, SVM, LR and KNN to classify cancer related lncRNAs from cancer-unrelated lncRNAs [18]. For this classification, authors used combination of genomic, epigenetic, network and expression features. In the present study, cancer related key lncRNAs are identified using lncRNA expression values of cancer patients applying feature selection algorithms. Then the capability of identified lncRNAs in classifying 8 different cancers is checked by the performance of five classification models. Finally, survival analysis is conducted to check whether the discovered lncRNAs are really capable of differentiating between high-risk and low-risk patients.

Hoadley et al. showed that cell of origin patterns dominate the molecular classification of tumors available in TCGA [19]. For their analysis, they used copy number, mutation, DNA methylation, RPPA protein, mRNA and miRNA expression. But, they did not consider another important molecular signature of cancer, which is lncRNA expression. While their work motivates us to classify multiple cancers using lncRNA expression, the main objective of this study is to find the key lncRNAs related to specific cancer. However, research on such classification is rarely found due to the high dimensionality of the data [20]. Though RNAseq data from TCGA contains

978-1-7281-1867-3/19/$31.00 © 2019 IEEE

reasonable number of samples, even it poses challenges for classification task due to large number of features (mRNA, miRNA, or lncRNA) with respect to the number of samples. Many computational methods fail to identify a small number of important features, rather increase learning costs and deteriorates performance [21]. To overcome this issue, researchers used feature selection algorithm for dimension reduction such as RFE (Recursive Feature Elimination) is used in [22], [23] and LASSO is used in [24] as a feature selection method.

It is clear from the literature that lncRNAs, play a key role in causing a cancer and its development. More research is needed to identify cancer specific lncRNAs. Existing methods used co-expression network such as lncRNA-mRNA or lncRNA-miRNA-mRNA. As per our knowledge, there is no study that uses lncRNA expression only to find the cancer specific lncRNAs except our previous work [25] where feature extraction did not consider. Thus, we proposed a computational framework using feature selection and classification methods that can identify key lncRNAs and classify different cancers based on the expression value of those key lncRNAs. Important features or lncRNAs are selected in two steps: First, number of feature is reduced using a cutoff on expression values and then using a combination of two feature selection algorithms RFE and LASSO. This study discovered 37 key lncRNAs for eight different cancers.

The paper is organized as follows: Section 1 introduces the role of lncRNAs as a key factor in tumorigenesis, different methods to identify key lncRNAs, rationale of present study in identifying cancer specific lncRNA. Section 2 describes the data preparation. Section 3 presents the methodology by enumerating feature selection, machine learning model selection, configuration, parameter tuning and evaluation. The results and discussion, model performance and validation are presented in section 4. Finally, section 5 concluded the research question and future directions.

TABLE I: Summary of TCGA RNA-seq data sets used in this study.

Tumor Types	#Tumor Samples	#Expressed lncRNAs
Bladder Cancer (BLCA)	430	2501
Cervical Cancer (CESC)	309	2327
Colon Cancer (COAD)	512	2178
Head and Neck Cancer (HNSC)	546	1831
Kidney Papillary Cell Carcinoma (KIRP)	321	2651
Lower Grade Glioma (LGG)	529	2941
Liver Cancer (LIHC)	424	1771
Lung Adenocarcinoma (LUAD)	585	2854
Total (Unique)	3656	4786

(Initial dataset contains 12,309 common lncRNAs with expression data for 8 cancers. Third column represents the number of expressed lncRNAs using threshold, mean lncRNA expression $\geq$ 0.3.)

II. Data Preparation

To validate the idea, RNAseq FPKM normalized expression data for 8 cancers - Bladder Cancer (BLCA), Cervical Cancer (CESC), Colon Cancer (COAD), Head and Neck Cancer (HNSC), Kidney Papillary Cell Carcinoma (KIRP), Lower Grade Glioma (LGG), Liver Cancer (LIHC) and Lung Adenocarcinoma (LUAD) - are downloaded (April, 2019) from UCSC xena [26]. Selection of this eight cancers is based on the number of samples (ranges from 309 to 585) to have a balanced dataset as shown in Table I. Combined dataset consists of 3656 patients with 60483 RNA (mRNA, miRNA, lncRNA) expression representing 8 tumor types. The row and column headings represent the RNAs and sample IDs respectively. Values of each cell represents the normalized read counts of an RNA for a specific sample. Since this study focuses on identification of key lncRNAs for a cancer, expression values of lncRNAs are isolated from the combined dataset using lncRNA IDs available in TANRIC (The Atlas of non-coding RNA in Cancer)[27]. This mapping resulted in 12,309 common lncRNAs with expression data for 8 cancers. In the present study, we used a cutoff, mean lncRNA expression $\geq$ 0.3 as used in [24], to determine expressed lncRNAs. The number of expressed lncRNAs for different cancer are shown in Table I. The combined number of expressed lncRNAs applying this threshold is 4786. The total number of cancer patient analyzed is 3656.

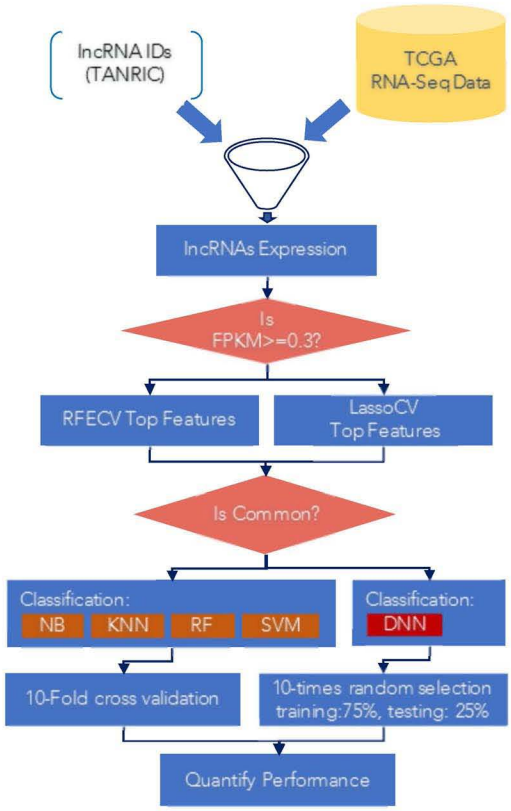

Fig. 1: Overall Process for Data Preparation and Methodology

III. Methodology

Fig. 1 shows the over all process for data preparation and methodology. After reducing features (lncRNA) using cutoff, mean expression $\geq$0.3, two feature selection methods RFE and LASSO are used to reduce the dimension further. To validate the capability of selected lncRNAs in classifying different cancers, 5 different learning algorithms - Nave Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Deep Neural Network (DNN) - are used.

A. Feature selection

The lncRNAs those have more contribution towards the classification of cancer types are more likely to be the key lncRNA for cancer diagnosis and prognosis. Feature selection methods can reduce the number of irrelevant and noisy lncRNAs and select the most related lncRNAs to improve the classification results, which decrease the computational costs and improve the cancer classification performance [21]. To achieve this goal, two widely applied wrapper based feature selection methods: Recursive Feature Elimination (RFE) [28] and Least Absolute Shrinkage and Selection Operator (LASSO) [29] are used. These algorithms have better classification efficiency and do not have limit on data types and can effectively deal with nominal or continuous features, missing data and noisy tolerance [30].

1) LASSO: The Least Absolute Shrinkage and Selection Operator method applies a regularization (shrinking) process where it penalizes the coefficients of the regression variables and shrinks these to zero. The variables that still have a non-zero coefficient are selected as the top features. The tuning parameter λ controls the strength of the penalty. The larger is the parameter λ the more number of coefficients are shrinked to zero, less number of features are selected. In this experiment, the optimized $\lambda = 0.0036$ is calculated by 5-fold cross validation which is able to picked 765 important features in 62 secs with 96% accuracy.

2) RFECV: Similarly, the Recursive Feature Elimination RFE algorithm constructs a ranking coefficient according to the weight vector w generated by an estimator e.g. linear regression during training. It removes a set of features with the smallest ranking coefficient in each iteration, and finally obtains an optimized number of significant features.

Scikit-learn feature selection [31], a python package, has been used for feature selection procedure. For a given number of features, both LASSO and RFE can produce an optimum number of features. The number of features produced by LASSO and RFE are 765 and 786 respectively from 4786 features.

B. Classification

We used scikit-learn [31], a python library, for machine learning models. For KNN model, k was set to 7. In SVM, linear kernel is used. For the RF model, the number of estimator is 10 with entropy ensembling. Finally, Gaussian NB algorithm is used for Naive Bayes model. In DNN, the number of hidden layer is one. The number of node in input layer is equal to the number of features (4786 lncRNAs). The hidden layer consists of 20 nodes which are identified by parameter tuning. The output layer has 8 nodes corresponding to 8 different cancer types. After tuning hyper-parameters and optimizing model parameters, a good convergence is found with learning rate 0.1 and epoch size 100. These parameters adjust the network for appropriate weights to prevent over-fitting. XAVIER is used as weight initializer in the model, which is a Gaussian distribution with mean 0, variance $2.0/(fanIn+fanOut)$. The function that learns the weight vector is called the optimizer function which is the stochastic gradient descent (*SGD*) in this experiment. In training a deep learning model, the selection of the optimizer, number of epochs, and batch size are important for achieving a good performance. The activation function allows the model to learn the complex data set. The activation function *ReLU* is used in all layers and negative log likelihood is used as the loss function.

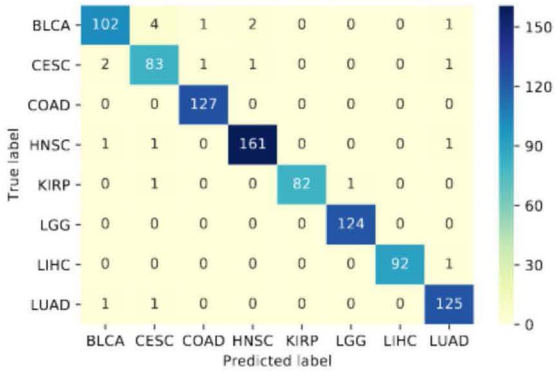

Fig. 2: Confusion Matrix of DNN Model (Accuracy = 98%, Number of Features = 37)

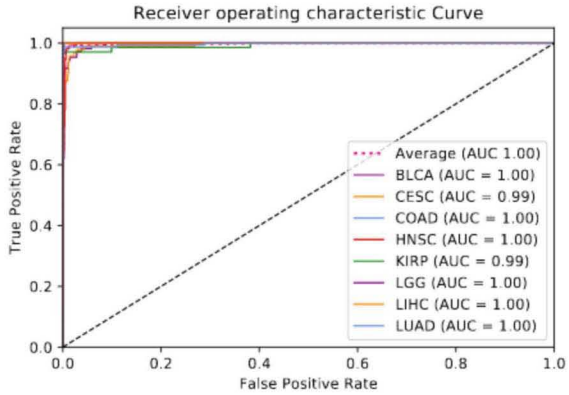

Fig. 3: ROC curve and AUC scores of different classes from SVM classifier

C. Parameter Tuning

The grid search method is used for ML to find the optimized parameter for machine learning algorithms. The hyper-parameters for deep neural network such as epoch, learning rate, number of hidden layers, etc. also need to be tuned to achieve high accuracy or precision. First, tuning is started with learning rate and epoch. One hyper-parameter is fixed to a certain value and observed the performance by changing the other. For example, epoch is fixed to 30 and the value of learning rate is changed in a range of 0.001 to 1.0. It is noticed that accuracy increases with the increase of learning rate then it stops increasing at a certain point and starts decreasing. The learning rate at which accuracy reached to its highest value is selected for experiment. Finally, learning rate 0.1 and epoch 100 produce a convergent result. Other hyper-parameters such as the number of hidden layers and seed are tuned in similar fashion. Deeplearning4J [32], a java machine learning package, is used for DNN model development. All models are executed on a CPU Intel core i7 with 16GB RAM. For training 75% of each cancer type is selected randomly using seed 123 for random number generation. The remaining 25% is used for testing in DNN. This training and testing procedure has been repeated 10 times. The average of these 10 results are used as the performance of the model. On the other hand, performance for the machine learning algorithms are measured by 10-fold cross validation. The ROC curve with AUC score for 8 different classes is shown in Fig. 3.

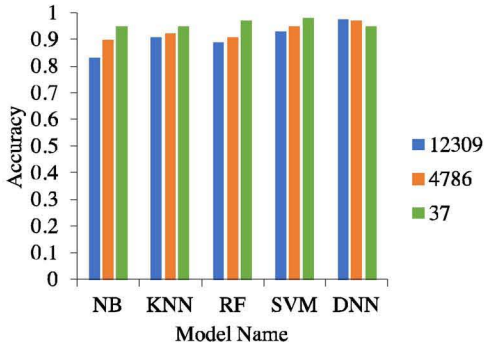

Fig. 4: Accuracy of different models with different number of features

D. Evaluation

In this study, five classification models - NB, KNN, RF, SVM and DNN - are used to classify 8 cancers - BLCA, CESC, COAD, HNSC, KIRP, LGG, LIHC, and LUAD. To compare the model performance, first, a confusion matrix is generated and then three different performance metrics - accuracy, precision, recall - are evaluated. Fig. 2 shows one of the confusion matrices obtained from DNN model with accuracy of 98%. Row labels represent the actual labels and column labels represent the predicted labels. Accuracy is the number of correct predictions made by the model over

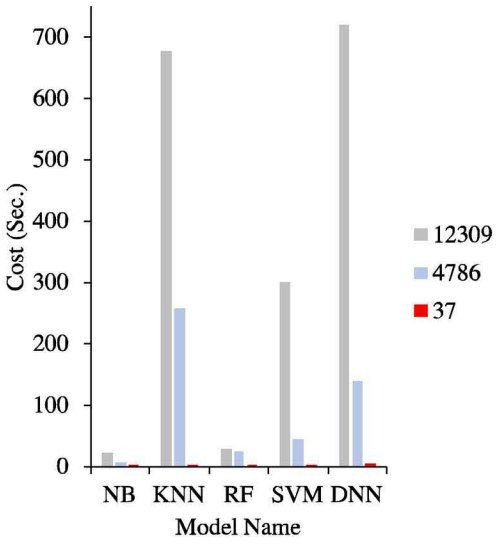

Fig. 5: Cost of different models with different number of features

all kinds of predictions made. True positives(TP) and True Negatives(TN) are the correct prediction.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is the number of correct positive results divided by the number of positive results predicted by the classifier. It indicates the predicted positive portion of the samples.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is the number of correct positive results divided by the number of all relevant samples.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

All scores are calculated from the test data.

IV. Results and Discussion

Table II shows the values of performance metrics for NB, KNN, RF, SVM and DNN models using three different number of feature sets or lncRNAs - 12,309, 4,786, and 37. First, all lncRNA expressions (12,309) are used to classify eight cancers. After initial reduction of feature size using cutoff, mean expression$\geq$0.3, 4,786 lncRNAs are left for classification. Then feature selection methods RFE and LASSO are used on 4,786 lncRNAs to find the optimum number of features. RFE and LASSO produced 786 and 765 features respectively. Classification is performed separately using RFE features and LASSO features and results show that RFE features performed better for most of the classifiers with accuracy ranging from 97% to 99%. Then common features, 344 lncRNAs between these two optimum feature sets are used to classify the tumor types which resulted in accuracy ranging from 96% to 99%. Since, RFE features performed better, further experiments

TABLE II: Performance comparison of different classifiers with different number of features

#Features	Model Name	Recall	Precision	Accuracy	Cost (sec.)
Block-a	NB	0.80±0.02	0.85± 0.02	0.83± 0.02	22.95
	KNN	0.90±0.01	0.91±0.01	0.91±0.01	676.23
	RF	0.89±0.02	0.89±0.02	0.89±0.01	28.39
	SVM	0.91±0.00	0.92±0.01	0.93±0.02	300.48
	DNN	0.97±0.01	0.97±0.01	0.97±0.01	720.33
Block-b	NB	0.89±0.01	0.89±0.01	0.90±0.01	7.41
	KNN	0.92±0.01	0.92±0.01	0.92±0.01	257.28
	RF	0.90±0.01	0.91±0.01	0.91±0.01	24.73
	SVM	0.95±0.01	0.95±0.01	0.95±0.01	42.79
	DNN	0.97±0.01	0.97±0.01	0.97±0.01	139.16
Block-c	NB	0.95±0.02	0.94±0.02	0.95±0.01	0.09
	KNN	0.94±0.01	0.95±0.01	0.95±0.01	1.44
	RF	0.97±0.01	0.97±0.01	0.97±0.01	2.28
	SVM	0.97±0.01	0.97±0.01	0.98±0.01	0.82
	DNN	0.95±0.01	0.95±0.01	0.95±0.01	4.07

(Block-a: Performance using all features (12309 lncRNAs)
Block-b: Performance using features obtained using mean FPKM $\geq$ 0.3 (4786 lncRNAs)
Block-c: Performance using features obtained from RFE and LASSO (37 lncRNAs)

TABLE III: 37 key lncRNAs identified in this study

lncRNA
AC000111.6, AC005082.12, AC005355.2, AC009299.3, AL450992.2, AP001626.1, BBOX1-AS1, CTA-384D8.31, EMX2OS, FAM182A, FENDRR, GATA3-AS1, H19, HAGLR, HOXA10-AS, HOXA11-AS, HOXD-AS2, KIZ, LINC00857, LINC00958, LINC01082, LINC01158, MIR205HG, NKX2-1-AS1, RP11-157J24.2, RP11-30K9.5, RP11-373D23.2, RP11-435O5.6, RP11-445O3.2, RP11-535M15.1, RP11-76C10.5, SFTA1P, TBX5-AS1, TMEM51-AS1, TP53TG1, UCA1, XIST

are conducted with reduced number of RFE features such as 200, 100, and 50 features while considering all LASSO features (765) that resulted in common features of 129, 68, and 37 lncRNAs respectively. With 37 lncRNAs, accuracy of classification ranges from 94% to 97% as shown in Fig. 4. Further reduction of features deteriorates the performance considerably. Thus, 37 lncRNAs as shown in Table III can be considered as the key lncRNAs related to eight cancers considered for analysis in this study.

It is clear from Fig. 4 and Fig. 5 that 37 key lncRNAs produces better performance in terms of both accuracy and cost compared to full (12,309 lncRNAs) and cutoff (4,786 lncRNAs) feature sets.

Validation: The results obtained, 37 key lncRNAs, are visually validated using t-SNE plot and survival analysis. Fig. 6 shows the t-SNE plot of eight cancer samples derived using expression values of 37 lncRNAs identified in the present study. It is clear from this figure that 37 lncRNAs are capable of differentiating eight different cancers. So, t-SNE plots provide a validation that 37 lncRNAs can be considered as the key features for diagnosis and prognosis of eight cancers.

Fig. 7 shows the validation of discovered lncRNAs using survival analysis. Fig.7a shows top 10 lncRNAs with importance score obtained by LASSO, six of these are positively co-related whereas four are negatively co-related. Fig. 7b shows the box plot of expression value of second negatively co-related lncRNA HOXD-AS2 for different cancers. It is clear from the box plot that this lncRNA has distinguishable 5-point statistics over the cancer types.

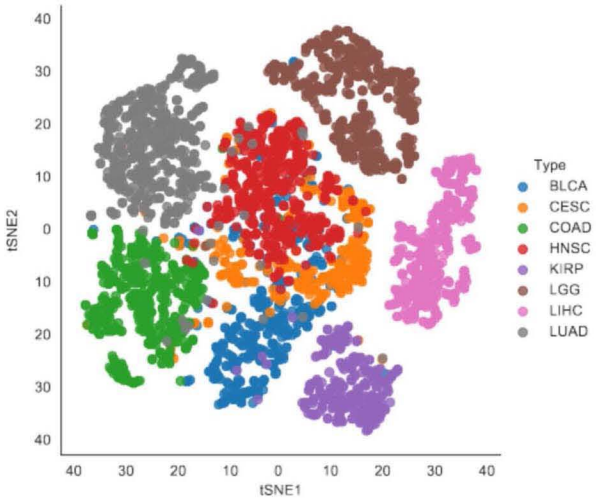

Fig. 6: tSNE representation of 37 lncRNA expressions for eight tumor types

Fig.7c shows the survival analysis using positively co-related lncRNA NKX2-1-AS1, which means a patient with high expression (red line) would have low probability of survival while a patient with low expression (Blue line) would have high probability of survival. Fig.7d shows survival analysis using negatively co-related lncRNA RP11-435O5.6, which means a patient with low expression (Blue line) would

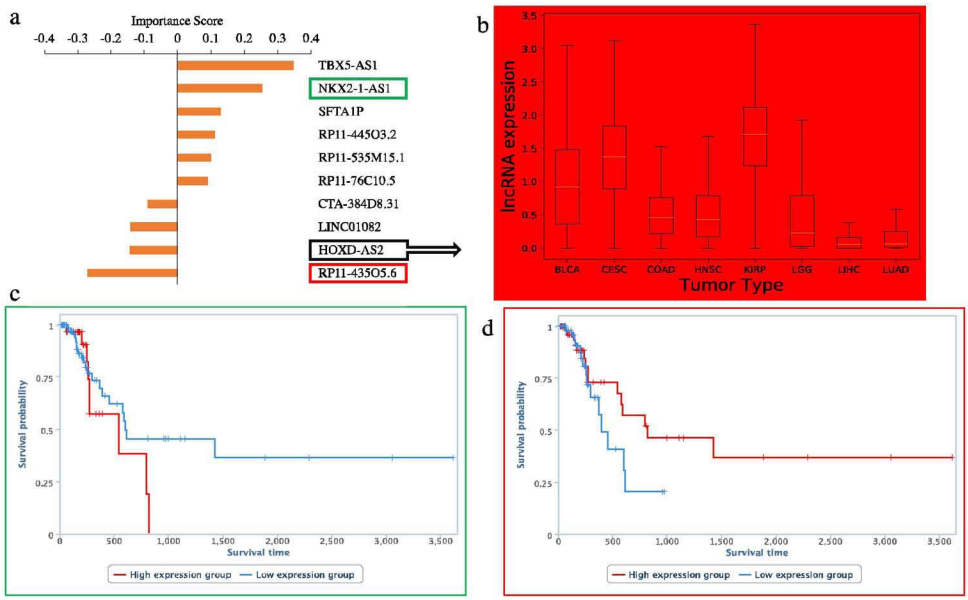

Fig. 7: Validation of discovered key lncRNAs. a) Top-10 lncRNAs with importance score by LASSO b) Box plot of expression values of lncRNA HOXD-AS2 for different cancers, c) Survival analysis using positively co-related lncRNA NKX2-1-AS1 in BLCA, and d) Survival analysis using negatively co-related lncRNA RP11-435O5.6 in BLCA. Survival Analysis is done using *TANRIC*.

have low probability of survival while a patient with high expression (red line) would have high probability of survival. These two survival analyses evidenced that first lncRNA is acting as positively co-related and second lncNRA is acting as negatively co-related lncRNA biomarker. Other lncRNAs also provided similar corelation which implicates that the discovered 37 lncRNAs can be considered as the key features in terms of diagnosis and prognosis of these 8 cancers. This approach is also used to identify the list of 214 key genes related to breast cancer which are used in constructing the trajectory of cancer development in [33].

V. Conclusion and Future Remarks

A computational framework is developed to identify key lncRNAs for multiple cancers employing two feature selection and five classification methods using lncRNA expression of cancer samples only. This study identified 37 key lncRNAs that can classify 8 cancers with an accuracy ranging from 94% to 97%. t-SNE plot and survival analysis support that the discovered 37 lncRNAs are capable of differentiating 8 cancers as well as differentiating between high-risk and low-risk patients. Thus, the discovered lncRNAs can be used as diagnostic and prognostic features for 8 cancers considered in this study. In the extended version of the paper, we plan to compare the discovered list of lncRNA with the existing literature if available. Another extension of this paper could be inclusion of lncRNA expression of corresponding normal samples.

Acknowledgment

This research is funded by NSF CAREER award #1651917 (transferred to #1901628) to AMM.

References

[1] SW Cheetham, F Gruhl, JS Mattick, and ME Dinger. Long noncoding rnas and the genetics of cancer. *British journal of cancer*, 108(12):2419, 2013.

[2] Yiwen Fang and Melissa J Fullwood. Roles, functions, and mechanisms of long non-coding rnas in cancer. *Genomics, proteomics & bioinformatics*, 14(1):42–54, 2016.

[3] Y Kotake, T Nakagawa, K Kitagawa, S Suzuki, N Liu, M Kitagawa, and Y Xiong. Long non-coding rna anril is required for the prc2 recruitment to and silencing of p15p15 ink4b tumor suppressor gene. *Oncogene*, 30(16):1956, 2011.

[4] Feng Yang, Jianwei Bi, Xuchao Xue, Luming Zheng, Kangkang Zhi, Jide Hua, and Guoen Fang. Up-regulated long non-coding rna h19 contributes to proliferation of gastric cancer cells. *The FEBS journal*, 279(17):3159–3165, 2012.

[5] Sophie Redon, Patrick Reichenbach, and Joachim Lingner. The non-coding rna terra is a natural ligand and direct inhibitor of human telomerase. *Nucleic acids research*, 38(17):5797–5806, 2010.

[6] Lars Henning Schmidt, Tilmann Spieker, Steffen Koschmieder, Julia Humberg, Dominik Jungen, Etmar Bulk, Antje Hascher, Danielle Wittmer, Alessandro Marra, Ludger Hillejan, et al. The long noncoding malat-1 rna indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. *Journal of thoracic oncology*, 6(12):1984–1992, 2011.

[7] Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. A coding-independent function of gene and pseudogene mrnas regulates tumour biology. *Nature*, 465(7301):1033, 2010.

[8] Hui Tao, Jing-Jing Yang, Xiao Zhou, Zi-Yu Deng, Kai-Hu Shi, and Jun Li. Emerging role of long noncoding rnas in lung cancer: Current status and future prospects. *Respiratory medicine*, 110:12–19, 2016.

[9] Adam M Schmitt and Howard Y Chang. Long noncoding rnas in cancer pathways. *Cancer cell*, 29(4):452–463, 2016.

[10] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[11] Chih-Hsun Wu, Chia-Lang Hsu, Pei-Chun Lu, Wen-Chang Lin, Hsueh-Fen Juan, and Hsuan-Cheng Huang. Identification of lncrna functions in lung cancer based on associated protein-protein interaction modules. *Scientific reports*, 6:35939, 2016.

[12] Guangle Zhang, Cong Pian, Zhi Chen, Jin Zhang, Mingmin Xu, Liangyun Zhang, and Yuanyuan Chen. Identification of cancer-related mirna-lncrna biomarkers using a basic mirna-lncrna network. *PloS one*, 13(5):e0196681, 2018.

[13] Weikang Xing, Zhenyu Qi, Cheng Huang, Nan Zhang, Wei Zhang, Yao Li, Minyan Qiu, Qi Fang, and Guozhen Hui. Genome-wide identification of lncrnas and mrnas differentially expressed in non-functioning pituitary adenoma and construction of an lncrna-mrna co-expression network. *Biology open*, 8(1):bio037127, 2019.

[14] Li-Xin Wang, Chuan Wan, Zheng-Bang Dong, Bai-He Wang, Hong-Ye Liu, and Yang Li. Integrative analysis of long noncoding rna (lncrna), microrna (mirna) and mrna expression and construction of a competing endogenous rna (cerna) network in metastatic melanoma. *Medical science monitor: international medical journal of experimental and clinical research*, 25:2896, 2019.

[15] Jing Sui, Yun-Hui Li, Yan-Qiu Zhang, Cheng-Yun Li, Xian Shen, Wen-Zhuo Yao, Hui Peng, Wei-Wei Hong, Li-Hong Yin, Yue-Pu Pu, et al. Integrated analysis of long non-coding rna-associated cerna network reveals potential lncrna biomarkers in human lung adenocarcinoma. *International journal of oncology*, 49(5):2023–2036, 2016.

[16] Ying Chen, Youmin Pan, Yongling Ji, Liming Sheng, and Xianghui Du. Network analysis of differentially expressed smoking-associated mrnas, lncrnas and mirnas reveals key regulators in smoking-associated lung cancer. *Experimental and therapeutic medicine*, 16(6):4991–5002, 2018.

[17] Andrés Lanzós, Joana Carlevaro-Fita, Loris Mularoni, Ferran Reverter, Emilio Palumbo, Roderic Guigó, and Rory Johnson. Discovery of cancer driver long noncoding rnas across 1112 tumour genomes: new candidates and distinguishing features. *Scientific reports*, 7:41544, 2017.

[18] Xuan Zhang, Jun Wang, Jing Li, Wen Chen, and Changning Liu. Crlncrc: a machine learning-based method for cancer-related long noncoding rna identification using integrated features. *BMC medical genomics*, 11(6):120, 2018.

[19] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

[20] Boyu Lyu and Anamul Haque. Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 89–96. ACM, 2018.

[21] Lin Sun, Xianglin Kong, Jiucheng Xu, Ruibing Zhai, Shiguang Zhang, et al. A hybrid gene selection method based on relieff and ant colony optimization algorithm for tumor classification. *Scientific Reports*, 9(1):8978, 2019.

[22] Fan Zhang, Howard L Kaufman, Youping Deng, and Renee Drabier. Recursive svm biomarker selection for early detection of breast cancer in peripheral blood. *BMC medical genomics*, 6(1):S4, 2013.

[23] Ying Zhang, Qingchun Deng, Wenbin Liang, and Xianchun Zou. An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. *BioMed research international*, 2018, 2018.

[24] Leng Han, Yuan Yuan, Siyuan Zheng, Yang Yang, Jun Li, Mary E Edgerton, Lixia Diao, Yanxun Xu, Roeland GW Verhaak, and Han Liang. The pan-cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature communications*, 5:3963, 2014.

[25] Abdullah Al Mamun and Ananda Mohan Mondal. Long non-coding rna based cancer classification using deep neural networks. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 541–541. ACM, 2019.

[26] Mary Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Akhil Kamath, Fran McDade, Dave Rogers, Angela N Brooks, Jingchun Zhu, and David Haussler. The ucsc xena platform for cancer genomics data visualization and interpretation. *BioRxiv*, page 326470, 2019.

[27] Jun Li, Leng Han, Paul Roebuck, Lixia Diao, Lingxiang Liu, Yuan Yuan, John N Weinstein, and Han Liang. Tanric: an interactive open platform to explore the function of lncrnas in cancer. *Cancer research*, 75(18):3728–3737, 2015.

[28] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[29] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[30] Raul-Jose Palma-Mendoza, Daniel Rodriguez, and Luis De-Marcos. Distributed relieff-based feature selection in spark. *Knowledge and Information Systems*, 57(1):1–20, 2018.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Eclipse Deeplearning4j Development Team. Deeplearning4j: Open-source distributed deep learning for the JVM. *Apache Software Foundation License 2.0. deeplearning4j dot org*.

[33] Tasmia Aqila, Abdullah Al Mamun, and Ananda M. Mondal. Pseudotime based discovery of breast cancer heterogeneity. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2019)*. IEEE, 2019.