3-23-2022

A Machine Learning Framework for Identifying Molecular Biomarkers from Transcriptomic Cancer Data

Md Abdullah Al Mamun
mmamu009@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

A MACHINE LEARNING FRAMEWORK FOR IDENTIFYING MOLECULAR BIOMARKERS FROM TRANSCRIPTOMIC CANCER DATA

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Md Abdullah Al Mamun

2022

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Md Abdullah Al Mamun, and entitled A Machine Learning Framework for Identifying Molecular Biomarkers from Transcriptomic Cancer Data, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Giri Narasimhan

_____
Fahad Saeed

_____
Leonardo Bobadilla

_____
Wenrui Duan

_____
Ananda M. Mondal, Major Professor

Date of Defense: March 23, 2022

The dissertation of Md Abdullah Al Mamun is approved.

_____
Dean John L. Volakis
College of Engineering and Computing

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

To my parents and my better half.

ACKNOWLEDGMENTS

I want to thank my advisor and committee members, everybody in the Machine Learning and Data Analytics Group (MLDAG), Bioinformatics Research Group (BioRG), my friends, and my family for all your support.

ABSTRACT OF THE DISSERTATION

A MACHINE LEARNING FRAMEWORK FOR IDENTIFYING MOLECULAR BIOMARKERS FROM TRANSCRIPTOMIC CANCER DATA

by

Md Abdullah Al Mamun

Florida International University, 2022

Miami, Florida

Professor Ananda M. Mondal, Major Professor

Cancer is a complex molecular process due to abnormal changes in the genome, such as mutation and copy number variation, and epigenetic aberrations such as dysregulations of long non-coding RNA (lncRNA). These abnormal changes are reflected in transcriptome by turning oncogenes on and tumor suppressor genes off, which are considered cancer biomarkers.

However, transcriptomic data is high dimensional, and finding the best subset of genes (features) related to causing cancer is computationally challenging and expensive. Thus, developing a feature selection framework to discover molecular biomarkers for cancer is critical.

Traditional approaches for biomarker discovery calculate the fold change for each gene, comparing expression profiles between tumor and healthy samples, thus failing to capture the combined effect of the whole gene set. Also, these approaches do not always investigate cancer-type prediction capabilities using discovered biomarkers.

In this work, we proposed a machine learning-based framework to address all of the above challenges in discovering lncRNA biomarkers. First, we developed a machine learning

pipeline that takes lncRNA expression profiles of cancer samples as input and outputs a small set of key lncRNAs that can accurately predict multiple cancer types. A significant innovation of our work is its ability to identify biomarkers without using healthy samples. However, this initial framework cannot identify cancer-specific lncRNAs. Second, we extended our framework to identify cancer type and subtype-specific lncRNAs. Third, we proposed to use a state-of-the-art deep learning algorithm concrete autoencoder (CAE) in an unsupervised setting, which efficiently identifies a subset of the most informative features. However, CAE does not identify reproducible features in different runs due to its stochastic nature. Thus, we proposed a multi-run CAE (mrCAE) to identify a stable set of features to address this issue. Our deep learning-based pipeline significantly extended the previous state-of-the-art feature selection techniques.

Finally, we showed that discovered biomarkers are biologically relevant using literature review and prognostically significant using survival analyses. The discovered novel biomarkers could be used as a screening tool for different cancer diagnoses and as therapeutic targets.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

LIST OF ABBREVIATIONS

AE	Autoencoder
AEFS	Autoencoder Feature Selection
AI	Artificial Intelligence
ANOVA	Analysis of variance
AUC	Area Under the Curve
BLCA	Bladder Cancer
BLCA	Bladder Urothelial Carcinoma
CAE	Concrete Autoencoder
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
COAD	Colon Adenocarcinoma
DEGs	Differentially Expressed Genes
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
FC	Fold Change
FIU	Florida International University
FN	False Negative
FP	False Positive
FPKM	Fragments Per Kilobase per Million mapped reads
GEM	Gene Expression Matrix
HER2	Human Epidermal growth factor Receptor 2
HNSC	Head and Neck Squamous Cell Carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma

KIRP	Kidney Papillary Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
KNN	K-Nearest Neighbor
L1MSVM	L1-Norm Multiclass Support Vector Machine
LASSO	Least Absolute Shrinkage And Selection Operator
LGG	Brain Lower Grade Glioma
LIHC	Liver Hepatocellular Carcinoma
lncRNA	Long Non-Coding RNA
LR	Linear Regression
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MCFS	Multi-Cluster Feature Selection
miRNA	Micro RNA
MNIST	Modified National Institute of Standards and Technology
mrCAE	multi-run CAE
mRNA	Messenger RNA
MSE	Mean Squared Error
NB	Naive Bayes
NIH	National Institutes of Health
PCA	Principal Component Analysis
PCC	Pearson correlation coefficient
RF	Random Forest
RFE	Recursive Feature Elimination

RL1MSVM	RFE and L1-Norm Multiclass Support Vector Machine
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristics
RPPA	Reverse Phase Protein Array
SGD	Stochastic Gradient Descent
SVM	Support Vector Machines
TANRIC	The Atlas of Noncoding RNAs in Cancer
TCGA	The Cancer Genome Atlas
TN	True Negative
TP	True Positive
t-SNE	t-Distributed Stochastic Neighbor Embedding
UCSC	University of California Santa Cruz
UDFS	Unsupervised Discriminative Feature Selection

LIST OF FREQUENTLY USED MATHEMATICAL SYMBOLS

k	Number of features to be selected
$f_r(.)$	CAE Reconstruction function
$f_r(x_S)$	CAE Reconstructed sample
X	Original data matrix
y_i	Actual value
y'_i	Predicted value
τ	Temperature
λ	Tuning parameter controls the strength of the penalty of SVM
C	Regularization parameter
j	j^{th} *feature*
x_i	i^{th} sample over a d-dimensional feature vector
y_i	Label vector of i^{th} sample
w_k	Weight vector
$p(x)$	Probability distribution
x_s	Data matrix consists of only selected feature

CHAPTER 1 INTRODUCTION

1.1 Motivation and Goals

Cancer is one of the most deadly diseases worldwide, as one in three people develop cancer during their lifetime [1]. Cancer is a complex multi-omics molecular process, and it normally happens due to over-expression of oncogenes and under-expression of tumor suppressor genes [2].

Epigenetics, meaning above genetics, which works on top of the genome without modifying the genetic material, controls the normal development of different types of cells and organs [3]. Any aberration in epigenetic processes such as DNA methylation at promoter regions [4], histone modifications (addition of methyl or acetyl group at histone proteins, for example) [5], and dysregulations of non-coding RNA (ncRNA) including both micro RNA (miRNA) [6] and long non-coding RNA (lncRNA) [7] could play critical roles in causing cancer by turning oncogenes on and tumor suppressor genes off. Thus, cancer is a multi-omics molecular process combining abnormal gene expression, DNA methylation, histone modifications, and miRNA and lncRNA dysregulations.

Dysregulation of multiple lncRNAs was reported to play major roles in many different cancers. The lncRNAs are a class of RNA transcripts with a length of >200 nucleotides that do not encode proteins. Studies have revealed that lncRNAs play an important role in cancer biology, and the expression of specific lncRNAs is implicated in the development and progression of cancer [8]. The lncRNAs also have key roles in transcriptional, post-transcriptional, and epigenetic gene regulation [9]. They also impact cancer pathways [10] and are involved in six cancer hallmarks such as proliferation, growth suppression, motility, immortality, angiogenesis, and viability [11].

One can find the molecular biomarkers (molecules involved in causing a disease) for cancer, considering each transcriptomic (RNA Type) data separately or combining all RNA types. The transcriptomic data are high-dimensional in nature. For example, human genome has about 20K (20,000) coding genes [12], and 40K non-coding genes (miRNA and lncRNA) [13]. Thus, to represent a human genome or an individual human being, we need 20K and 40K dimensions for coding genes and non-coding genes, respectively. A subset of 20K coding genes and 40K non-coding genes are responsible for cancer, called cancer biomarkers. Any dataset with an N-number of features has 2^N possible subset of features [14]. In the presence of such a large number of possible combinations, finding the best subset of N features, which are related to causing cancer, is computationally challenging and expensive [15]. Since transcriptomic data carry valuable information about cancers, RNA expression data is used for the early prediction of cancer in many studies [16]–[18]. However, there is a high risk of overfitting if a machine learning model is trained using such high dimensional data without reducing its dimension, meaning without identifying important features [19]. Therefore, there is a definite need for developing a feature selection framework capable of discovering salient features or molecular biomarkers for cancer from such high-dimensional transcriptomic data and accurately predicting cancer types and subtypes.

1.2 Significance and Research Purpose

Identifying the salient features or molecular biomarkers from each transcriptomic data and early prediction of cancer types or subtypes will help develop the screening tools and targeted therapy for cancer. More specifically, this will contribute towards the development of precision medicine.

The number of molecular biomarkers for a cancer type should be as few as possible (preferably less than 100) [20][21] so that it is easy to develop a wet lab experiment to check the feasibility of the discovered biomarkers as a possible screening tool and targeted therapy. The purpose of this research is to develop and implement a feature selection framework that can identify molecular biomarkers, a set of molecules in the range of 10 to 100 related to a cancer type or subtype, from the original feature space of 20K or 40K.

1.3 Specific Aims

The goals of the proposed study are to develop an intelligent feature selection framework for identifying molecular biomarkers for cancers and cancer subtypes. The goals were achieved through three specific aims as outlined below.

1.3.1 Specific Aim 1: Feature Selection and Cancer Type Classification

In this aim, a feature selection framework was developed to isolate a set of key features from a high-dimensional feature space [22]. The isolated features can differentiate multiple types of cancers, but the framework is not capable of providing information about which features contribute to which cancer.

1.3.2 Specific Aim 2: Class-Specific Feature Selection and Cancer Subtype Classification

In the second aim, we developed a feature selection framework to identify class-specific key features. For example, our proposed model can successfully discover molecular biomarkers associated with each breast cancer subtype.

1.3.3 Specific Aim 3: Building Deep Learning-based Feature Selection Framework

Traditional machine learning models are often linear, and these linear models may fail to capture the complex non-linear relationships of multivariate signals, resulting in an inferior

performance at the cost of its efficiency. Aim 3 integrated a deep learning-based feature selection algorithm called multi-run concrete autoencoder (mrCAE) to develop an enhanced framework for selecting the most meaningful features to predict cancer types/subtypes mentioned in Aim 1 and 2.

1.4 Research Contributions

The contributions of this dissertation include developing a feature selection framework that enables discovering molecular biomarkers and early cancer prediction. The following suite of feature selection methods was proposed in developing the framework.

1.4.1 Feature Selection and Cancer Type Classification

This study developed a computational framework 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 the experiment, expression values of lncRNAs for eight cancers – BLCA, CESC, COAD, HNSC, KIRP, LGG, LIHC, and LUAD – from TCGA were used. The combined dataset consists of 3,656 patients with expression values of 12,309 lncRNAs. Important features or key lncRNAs were identified using feature selection algorithms RFE and LASSO. The capability of these key lncRNAs in classifying eight different cancers is checked by the performance of five classification models. This study identified 37 key lncRNAs that can classify eight different cancer types with an accuracy ranging from 94% to 97%. Finally, survival analysis supports that the discovered key lncRNAs can differentiate between high-risk and low-risk patients.

1.4.2 Class-Specific Feature Selection and Cancer Subtype Prediction

Every cancer is stratified into multiple molecular subtypes such as Breast cancer has five subtypes: Basal, HER2, Luminal A, Luminal B, and Normal-like. Identifying subtype-specific key lncRNAs with clinical outcomes might help develop appropriate cancer therapy. We proposed an approach for simultaneous feature selection and classification for a multiclass problem combining recursive feature elimination (RFE) and *l*1-norm multiclass Support Vector Machine (L1MSVM), thus calling it RL1MSVM. The newly proposed model RL1MSVM performs better than two state-of-the-art models, L1MSVM and Random Forest (RF), with respect to four evaluation metrics, including accuracy, precision, recall, and f1 score. A total of 196 lncRNAs, which are the optimum number of features based on RL1MSVM, were selected using all three methods for comparison. Using these sets of features, the subtype prediction accuracies were 84%, 90%, and 92% for RF, L1MSVM, and RL1MSVM, respectively. Finally, a stable set of 91 lncRNAs was obtained using the union of the intersections of the two sets of lncRNAs selected by two approaches, which are considered key lncRNAs. Out of 91 lncRNAs, 53 were previously identified, and the remaining 38 are novel. The subtype-specific distribution of novel lncRNAs is – Basal: 7, HER2: 8, Luminal A: 11, Luminal B: 7, and Normal-like: 5, respectively. One of the lncRNAs found in two subtypes. The combined list of this novel and known lncRNAs can further be studied for developing breast cancer subtype-specific targeted therapy.

1.4.3 Multi-Run Concrete Autoencoder for Feature Selection

To discover the critical lncRNAs that can identify the origin of different cancers, we proposed to use the state-of-the-art deep learning algorithm Concreate Autoencoder (CAE) in an unsupervised setting, which efficiently identifies a subset of the most informative

features [23]. However, CAE does not identify reproducible features in different runs due to its stochastic nature [24]. To address this issue, we proposed a multi-run CAE (mrCAE) to identify a stable set of features [25]. The assumption is that a feature appearing in multiple runs carries more meaningful information about the data under consideration. The genome-wide lncRNA expression profiles of 12 different types of cancers, a total of 4,768 samples available in The Cancer Genome Atlas (TCGA), were analyzed to discover the key lncRNAs. To obtain a stable set of lncRNAs capable of identifying the origin of 12 different cancers, a lncRNA identified by CAE in multiple runs was added to the final list of key lncRNAs.

Our results showed that mrCAE performs better in feature selection compared to single-run CAE and other state-of-the-art feature selection techniques, including Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Support Vector Machine with Recursive Feature Elimination (SVM-RFE), Multi-Cluster Feature Selection (MCFS), and Unsupervised Discriminative Feature Selection (UDFS). This study discovered a set of top-ranking 128 lncRNAs that could identify the origin of 12 different cancers with an accuracy of 94%. Survival analysis showed that 101 of 128 lncRNAs have the prognostic capability in differentiating high- and low-risk groups of patients in different cancers.

The proposed computational framework can be used as a diagnostic tool by physicians to discover the origin of cancers using the expression profiles of lncRNAs. The discovered lncRNAs can be studied further by biologists or drug designers to identify possible targets for cancer therapy.

1.5 Roadmap for the Dissertation

After setting up the stage for the drive-in this introductory chapter, the rest of the journey is organized as follows.

Chapter 2 will introduce the reader to all notations, definitions, and necessary terminologies to understand different machine learning models used for feature selection, classification, and visualization tasks. It also discussed the different metrics used in measuring the model performance. Finally, it contains a literature survey on different machine learning techniques in biomarker discovery.

Chapter 3 contains a detailed explanation of how feature selection is important for high dimensional transcriptomic data and its use in identifying important cancer biomarkers. This chapter also discussed the development of the whole feature selection pipeline.

Chapter 4 provides the necessary information about running the same experiments on a large scale with a high volume of data. It contains necessary information regarding the extended experiment for 33 cancers using a high volume of transcriptomic data. It also contains a detailed explanation of the concrete autoencoder used for feature selection.

Chapter 5 provides a detailed explanation of how glycome genes performed an important role in cancer progression. It also discussed how the proposed feature selection framework could identify important glycome biomarkers for different cancers.

Chapter 6 describes a subtype-specific feature selection framework to identify biomarkers associated with each molecular subtype of breast cancer. It also provides the necessary information on classifying breast cancer subtypes using machine learning methods. In addition, it shows the prognostic evaluation and a detailed discussion on how novel biomarkers can be used in cancer prognosis and diagnosis.

Chapter 7 provides the limitations of concrete autoencoder in feature selection from high dimensional transcriptomic data. It also discussed how the multi-run approach could overcome its limitations. This chapter contains the necessary information on data preprocessing, model development, model training, hyperparameters tuning, and performance evaluation. It also contains the biological validations of identified biomarkers. We close the dissertation in Chapter 8 with a summary of the dissertation and conclusions and suggestions for future work.

CHAPTER 2 BACKGROUND AND REVIEW

This chapter provides the necessary background to understand different machine learning models used for visualization, feature selection, and classification tasks. It also discussed the different metrics used to measure the model performance. Finally, it contains a literature survey on different machine learning techniques in biomarker discovery.

2.1 Data visualization

2.1.1 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique used to explore and visualize high-dimensional data [26]. In simpler terms, t-SNE provides us a feel or intuition of how the data is arranged in a high-dimensional space. For example, the clustering of eight different cancer types is visualized by two t-SNE components derived from 12K dimensions of lncRNA expression profile data in Figure 2.1.

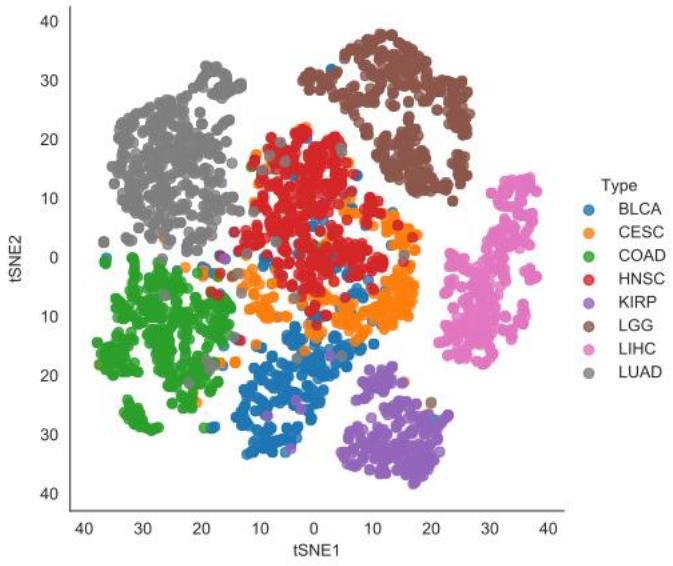

Figure 2.1: Visualization by t-SNE. The expression profile of 12K lncRNAs was reduced to 2 dimensions applying t-SNE. X-axis represents tSNE1, y-axis represents tSNE2, and each dot presents a patient of a particular cancer type.

2.2 Feature Selection Methods

Three general classes or types of feature selection techniques, filter, wrapper, and embedded methods, are discussed in the following subsections.

2.2.1 Filter Method

The filter method works by ranking the features using a statistical score assigned to each of them depending on their relevance to the class type. In both univariate and multivariate filter methods, the interactions among features are disregarded in the selection process. Studies like the ones in Pearson correlation coefficient (PCC), t-statistics (TS) [27], F-Test [28], and ANOVA [29] are examples where the filtering method is used. It is observed that these methods are effective for selecting features in high-dimensional data because of the reduced computation expenses. However, they fail to provide good accuracy, as discussed in [30].

2.2.2 Wrapper Method

As an enhancement, the researcher developed the wrapper-based feature selection method with a learning algorithm and a classifier to find a suitable subset of features. Initially, a random solution is generated, following which an objective function is maximized using black-box type optimization methods [31] like simulated annealing [32], particle swarm optimization [33], genetic algorithm [34], and ant colony optimization [35]. The iterative evaluation of every candidate subset of the features by a wrapper method leads to identifying a strong relationship between features, however, with an increase in the computational expense.

2.2.3 Embedded Method

Embedded feature selection methods, on the other hand, reduce computational costs because these are used as a part of the learning phase. Well-known embedded methods, which are considered as the state-of-the-art, are least absolute shrinkage and selection operator (LASSO) [36], recursive feature elimination with support vector machine estimator (SVM-RFE) [22], [37], [38], random forest [39], [40], Adaboost [41] , KNN [42] , and autoencoder [43].

2.3 Example Feature Selection Methods

2.3.1 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 the parameter λ, the more the number of coefficients shrunk to zero, the fewer features are selected.

2.3.2 Recursive Feature Elimination

A Recursive Feature Elimination (RFE) algorithm constructs a ranking coefficient according to the weight vector $\boldsymbol{w}$ generated by an estimator, e.g., linear regression during training. It removes features with the smallest ranking coefficient in each iteration and finally, obtains an optimized number of significant features. Recursive feature elimination is a recursive method in which less important features are eliminated in every iteration.

2.3.3 Random Forest

Random Forest (RF) works based on a tree structure that employs ensemble. RF consists of a number of decision trees. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two branches, so similar response values end up in the same set. The optimal condition is chosen based on impurity [44]. For classification, it is either *Gini* impurity or information gain/entropy. Thus, when the tree is fully developed, it can compute how much each feature decreases the weighted impurity on the tree. For forest, the impurity decrease from each feature can be measured as a feature rank. The feature importance is calculated as the sum over the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits [45].

2.3.4 Concrete Autoencoder (CAE)

Concrete autoencoder (CAE) proposed by Abid *et al.* [46] is a variation of the original autoencoder (AE) [47], which is used for dimension reduction. The motivation behind selecting CAE in the present study is that it takes advantage of both AE (which can achieve the highest classification accuracy) and concrete relaxation-based feature selection (capable of selecting actual features instead of latent features). An AE is a neural network that consists of two parts: (a) an encoder that selects latent features and (b) a decoder that uses selected latent features to reconstruct an output that matches the input with minimum error. In CAE, instead of using a sequence of fully connected layers in the encoder, a concrete relaxation-based feature selection layer is used where the user can define the number of nodes (features to be selected), k as shown in Figure 2.2. This layer selects a probabilistic linear arrangement of input features while training, which converges to a

discrete set of *k* features by the end of the training phase, which is subsequently used in the testing phase.

Let's *p(x)* is a probability distribution over a *d*-dimensional vector. The objective is to identify a subset of features, $S \equiv \{1 \ldots k\}$ of size $|S|=k$. Also, learning a reconstruction function $f_r(.): \mathbb{R}^{\mathrm{k}} \xrightarrow{\Delta} \mathbb{R}^{\mathrm{d}}$, such that the loss between original sample $\boldsymbol{x}$ and reconstructed sample $f_r(x_S)$ is minimized as stated in Eq. 1,

$$argmin_{S,r}\, E_{p(x)}[\|f_r\,(x_S\,) - x\|_2] \ldots\ldots\ldots\ldots\ldots (1)$$

where $x_s \in \mathbb{R}^k$ consists of only selected features x_i s.t. $i \in S$. Note that samples are represented in a *2D* matrix, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, and aim is to pick *k* columns of $\boldsymbol{X}$ such that sub-matrix $X_s \in \mathbb{R}^{n \times k}$.

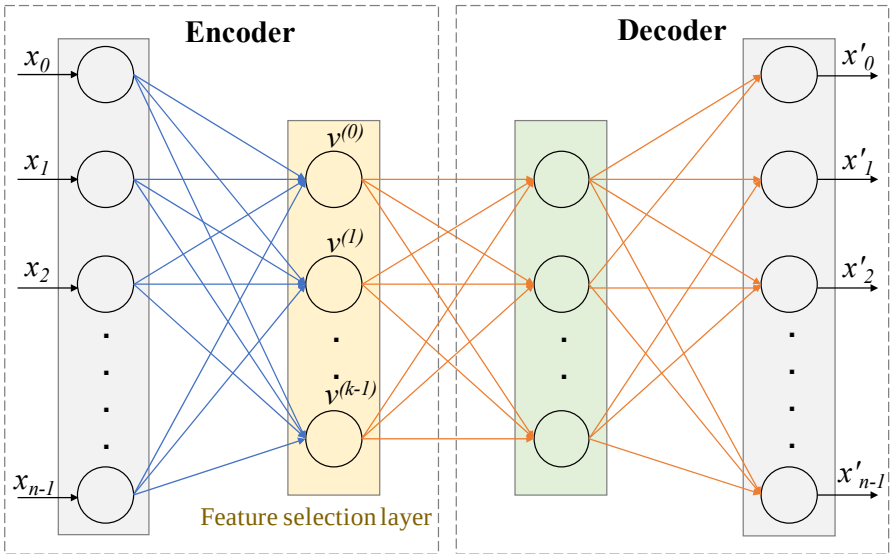

Figure 2.2 Architecture of Concrete Autoencoder. CAE architecture consists of an encoder and a decoder. The layer after the input layer of the encoder is called concrete feature selection layer shown in yellow. This layer has k number of the node where each node is for each feature to be selected. During the training stage, the $\boldsymbol{i^{th}}$ node $\boldsymbol{v^{(i)}}$ takes the value $\boldsymbol{X^T}$ f(i), where f(i) is the corresponding weight vector of node i. During the testing stage, these weights are fixed and the element with the highest value is selected by the corresponding $\boldsymbol{i^{th}}$ hidden node. The architecture of the decoder remains the same during training and testing.

Then, the selected feature set x_s can be used to reconstruct the original matrix X and classify the cancer types. In the feature selection layer of CAE (Figure 2.2), the original features are selected based on the temperature of this layer which is tuned using an annealing schedule. More specifically, the concrete selector layer identifies *k* important features as the temperature decreases to zero. For reconstructing the input, a simple decoder similar to the ones associated with a standard AE is used.

2.4 Classification Models

2.4.1 Support Vector Machine (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in *N*-dimensional space (*N* - the number of features) that distinctly classifies the data points [48]. Many possible hyperplanes could be chosen to separate the two classes of data points. SVM's objective is to find a plane with the maximum margin, *i.e.*, the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement to classify future data points with more confidence.

2.5 Performance Metrics

2.5.1 Confusion matrix

Confusion Matrix is the visual representation of the number of Actual vs. Predicted samples. It measures the performance of a supervised Machine Learning classification model on a set of test data for which the true values are known. It is useful for measuring Accuracy, Precision, Recall, and other important performance metrics.

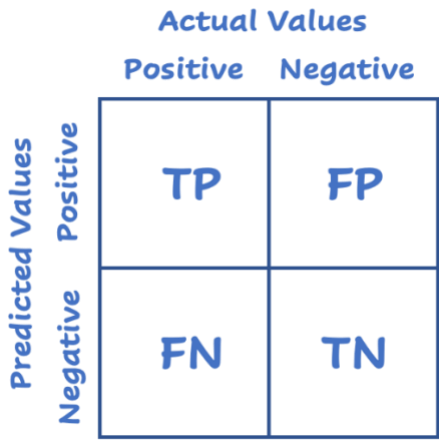

Figure 2.3 Confusion matrix of a binary classification problem. (TP: True Positive; FP: False Positive; FN: False Negative; TN: True Negative).

2.5.2 Accuracy

Accuracy is the number of correct predictions made by the model over all kinds of predictions made. True positives (TP) and True Negatives (TN) are the correct predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

2.5.3 Precision

The precision is the number of correct positive results divided by the number of positive results predicted by the model. It indicates the predicted positive portion of the samples.

$$Precision = \frac{TP}{TP+FP}$$

2.5.4 Recall

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

$$Recall = \frac{TP}{TP+FN}$$

2.5.5 F1-score

F1 score is the harmonic mean of precision and recall.

$$F1\ Score = 2\ \frac{Precision\ .\ Recall}{Precision + Recall}$$

2.5.6 Mean squared error (MSE)

Mean squared error (MSE) measures the amount of error in a regression problem. It assesses the average squared difference between the actual values (y_i) and predicted values (y'_i).

$$MSE = \frac{\Sigma(y_i - y'_i)^2}{n}$$

2.5.7 AUC - ROC Curve

Area Under the Curve (AUC) – Receiver Operating Characteristics (ROC) curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model predicts Positive classes as Positive and Negative classes as Negative. For example, the higher the AUC, the better the model can distinguish between patients with the disease and no disease.

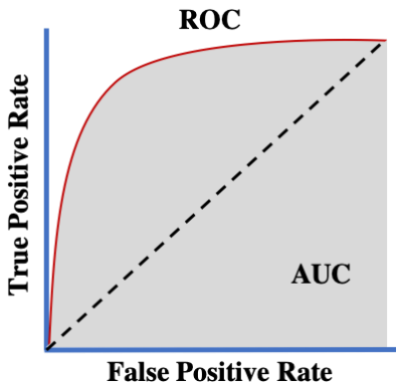

Figure 2.4 AUC-ROC Curve. AUC: Area Under the Curve; ROC: Receiver Operating Characteristic Curve. AUC-ROC curve is used to measure the performance of a binary classification system.

2.6 Data Domain

Data used in this experiment is transcriptomic cancer data. More specifically, we used the long non-coding part of RNA, which is more than 200 nucleotides long. The expression profiles of long non-coding RNA (lncRNA) of cancer and normal samples were used in this study. *It should be noted that normal samples refer to healthy samples in entire dissertation*.

2.7 Literature Review

Recent studies showed that long non-coding RNAs (lncRNAs) play key roles in tumorigenesis [49]–[51]. The lncRNAs also have key functions in transcriptional, post-transcriptional, and epigenetic gene regulation [9]. Schmitt and Chang discussed the impact of lncRNA in cancer pathways [10]. Hanahan and Weinberg described the involvement of lncRNAs in six hallmarks of cancer such as proliferation, growth suppression, motility, immortality, angiogenesis, and viability [11].

Hoadley *et al.* showed that cell of origin patterns dominate the molecular classification of tumors available in The Cancer Genome Atlas (TCGA) [52]. Their analysis used copy number, mutation, DNA methylation, RPPA protein, mRNA, and miRNA expression. However, they did not consider another important molecular signature of cancer, which is lncRNA expression. This work motivated us to investigate the importance of lncRNAs in identifying different types or subtypes of cancer.

However, research on such classification is rarely found due to the high dimensionality of the data [53]. Though RNAseq data from TCGA contains a reasonable number of samples, even it poses challenges for classification tasks due to a 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 increasing learning costs and deteriorating performance [54]. To overcome this issue, researchers used a feature selection algorithm for dimension reduction such as RFE (Recursive Feature Elimination) is used in [55], [56] and LASSO is used in [57] as a feature selection method. More research is needed to identify cancer type and subtype-specific lncRNAs.

In general, feature selection is worthwhile when the whole set of features is difficult to collect or expensive to generate [23]. For example, in TCGA, the lncRNA expression profile dataset contains more than 12,000 features (lncRNAs) for each of 33 different cancers, and it is expensive to generate this data.

On the other hand, standard dimension reduction methods, such as principal component analysis (PCA) [58] and autoencoders [47], can generate a greatly reduced set of ***latent features***. However, these latent features are not ***the original features*** but are combinations of the original features. Identifying original features increases the "explainability" of the results and allows us to perform biological interpretation in diagnosing various deadly diseases, such as cancers. Recently, few deep learning-based feature selection methods showed improvement in selecting original features in both supervised and unsupervised settings [23], [59]–[61].

CHAPTER 3 FEATURE SELECTION AND CANCER CLASSIFICATION (8 CANCERS)

This chapter contains a detailed explanation of how feature selection is important for high dimensional transcriptomic data and its use in identifying important cancer biomarkers. In this chapter, we also discussed the development of the whole pipeline of the initial feature selection framework from data preprocessing to classify the eight cancer types.

3.1 Introduction

Recent studies indicate that several cancer risk loci are transcribed into lncRNAs, and these transcripts play key roles in tumorigenesis [49], [50]. In their review paper, Cheetam et al. [49] enumerated that lncRNAs play key roles in cancer progression through a variety of mechanisms such as lncRNA *ANRIL* for remodeling of chromatin [62], *H19* for transcriptional co-activation and co-repression [63], *TERRA* for protein inhibition [64], *MALAT1* for post-transcriptional modifications [65] and *PTENP1* for decoy [66]. LncRNA *ANRIL*, which causes *PCR1*-mediated repression of tumor suppressor locus *INK4A-ARF-INK4b*, is up-regulated in prostate cancer [62]. Similarly, *H19* plays a significant role in the proliferation of gastric cancer cells due to its up-regulation [63]. The lncRNA *TERRA* facilitates telomeric heterochromatin formation [64], *MALAT1* induces migration and tumor growth in lung cancer [65],and *PTENP1* controls the expression level of the tumor suppressor gene *PTEN* [66]. Also, lncRNAs have key functions in transcriptional, post-transcriptional, and epigenetic gene regulation [9]. Schmitt et al. discussed the impact of lncRNA in the cancer pathway [10]. They described the involvement of lncRNAs in six hallmarks of cancer [67], such as proliferation, growth suppression, motility, immortality, angiogenesis, and viability. While some researchers detailed the role of lncRNAs in cancer

progression, others discovered a number of lncRNA biomarkers in several cancers by creating lncRNA-miRNA co-expression networks [68]–[70]. Wang et al., on the other hand, identified six key lncRNAs for metastatic melanoma from a competing endogenous RNA (ceRNA) network analysis using mRNA, miRNA, and lncRNA expression [71]. By constructing a similar network, Sui et al. found 41 lncRNAs biomarkers in human lung adenocarcinoma [72]. Also, Chen et al. identified 24 hub lncRNAs in smoking-associated lung cancer by forming protein-protein interaction (PPI) networks [73]. Similarly, Lanzos et al. identified cancer driver lncRNAs as new candidates and distinguishing features by analyzing the mutational patterns in tumor DNA [74]. Another model, CRlncRC, used machine learning algorithms including RF, NB, SVM, LR, and KNN to classify cancer-related lncRNAs from cancer-unrelated lncRNAs [75]. For this classification, the authors used a 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 eight different cancers is checked by the performance of five classification models. Finally, survival analysis is conducted to check whether the discovered lncRNAs can differentiate 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 [52]. They used copy number, mutation, DNA methylation, RPPA protein, mRNA, and miRNA expression for their analysis. 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 [53]. Though RNAseq data from TCGA contains a reasonable number of samples, even it poses challenges for classification tasks due to a 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 increasing learning costs and deteriorating performance [54]. To overcome this issue, researchers used a feature selection algorithm for dimension reduction such as RFE (Recursive Feature Elimination) is used in [55], [56] and LASSO is used in [57] as a feature selection method.

It is clear from the literature that lncRNAs play a key role in causing cancer and its development. More research is needed to identify cancer-specific lncRNAs. Existing methods used co-expression networks such as lncRNA-mRNA or lncRNA-miRNA-mRNA. As per our knowledge, no study uses lncRNA expression only to find the cancer-specific lncRNAs except our previous work [76], where we did not consider feature selection. Here, we proposed a computational framework using feature selection and classification methods to 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, the number of features 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.

3.2 Data Preparation

To validate the idea, RNAseq FPKM normalized expression data for eight 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 [77]. These eight cancers are selected based on the number of samples (ranges from 309 to 585) to have a balanced dataset, as shown in Table 3.1. The combined dataset consists of 3656 patients with 60483 RNA (mRNA, miRNA, lncRNA) expression profiles representing eight tumor types. The row and column headings represent the RNAs and sample IDs, respectively. The values of each cell represent the normalized read counts of an RNA for a specific sample. Since this study focuses on identifying key lncRNAs for a cancer type, expression values of lncRNAs are isolated from the combined dataset using lncRNA IDs available in the TANRIC (The Atlas of non-coding RNA in Cancer) repository [78]. This mapping resulted in 12,309 common lncRNAs with expression data for eight cancers. In the present study, we used a cutoff, mean lncRNA expression ≥ 0.3 as used in [57], to determine the expressed lncRNAs. The number of expressed lncRNAs for different cancer are shown in Table 3.1.

Table 3.1: Summary of TCGA RNA-seq data sets used in this study. The combined number of expressed lncRNAs is 4786. The total number of cancer patients analyzed is 3656.

Tumor Types	Short Name	#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**

3.3 Methodology

Figure 3.1 shows the overall process for data preparation and methodology. After reducing features (lncRNA) using cutoff, mean expression ≥ 0.3, two feature selection methods, RFE and LASSO, were used to select the key features related to different cancers. To validate the capability of selected lncRNAs in classifying different cancers, five different learning algorithms - Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Deep Neural Network (DNN) - were used.

3.3.1 Feature selection

The lncRNAs that 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, thus, decreasing computational costs and improving cancer classification performance [54]. To achieve this goal, two widely applied wrapper-based feature selection methods: Least Absolute Shrinkage and Selection Operator (LASSO) [79] and Recursive Feature Elimination (RFE) [37], were used. These algorithms have better classification efficiency and do not have a limit on data types and can effectively deal with nominal or continuous features, missing data, and noisy tolerance [80].

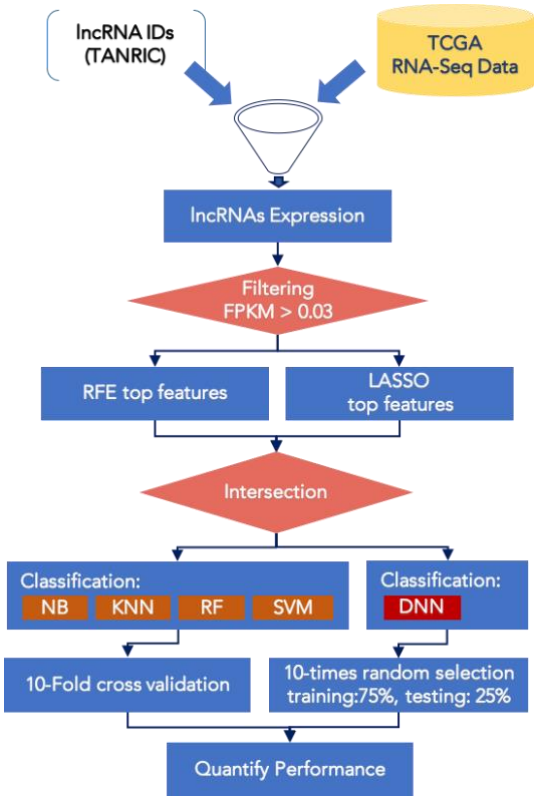

Figure 3.1 Overall Process for Data Preparation and Methodology. Initial dataset contains coding and noncoding genes, long non-coding part of data is extracted for further processing.

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 the parameter λ, the more the number of coefficients are shrunk to zero, fewer features are selected. In this experiment, the optimized λ = 0.0036 was calculated by 5-fold cross-validation, which was able to pick 765 important features in 62 secs with 96% accuracy.

RFE

Similarly, the Recursive Feature Elimination (RFE) algorithm constructs a features ranking 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 [81], a python package, was used for the feature selection procedures. Both LASSO and RFE can identify an optimum number of features from a given number of features. The numbers of features identified by LASSO and RFE were 765 and 786 respectively, from 4786 features.

3.3.2 Classification

We used *scikit-learn* [82], a python library, for machine learning models. For the KNN model, ***k*** was set to 7. In SVM, a linear kernel was used. For the RF model, the number of estimators is 10 with entropy ensembling. Finally, the Gaussian NB algorithm was used for the Naive Bayes model. In DNN, the number of the hidden layer was one. The number of nodes in the input layer was equal to the number of features (4786 lncRNAs).

The hidden layer consists of 20 nodes identified by parameter tuning. The output layer had eight nodes corresponding to eight different cancer types. After tuning hyperparameters and optimizing model parameters, a good convergence was found with a learning rate of 0.1 and an epoch size of 100. These parameters adjust the network for appropriate weights to prevent over-fitting. *XAVIER* is used as a 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 stochastic gradient descent (SGD) in this experiment. In training a deep learning model, selecting the optimizer, the

number of epochs, and batch size is important for achieving 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.

3.3.3 Parameter tuning

The grid search method is used for ML to find the optimized parameter for machine learning algorithms. The hyperparameters for the deep neural networks, such as epoch, learning rate, number of hidden layers, etc., must also be tuned to achieve high accuracy or precision. First, tuning is started with the learning rate and epoch. One hyperparameter is fixed to a certain value and observed the performance by changing the other. For example, an epoch is fixed to 30, and the learning rate value is changed in a range of 0.001 to 1.0. It is noticed that accuracy increases with the learning rate, then it stops increasing at a certain point and starts decreasing. The learning rate at which accuracy reached its highest value is selected for the experiment. Finally, the learning rate of 0.1 and an epoch of 100 produce a convergent result. Other hyperparameters such as the number of hidden layers and seed are tuned similarly.

Deeplearning4J [83], 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 was used as the performance of the model. On the other hand, performance for the machine learning algorithms is measured by 10-fold cross-validation.

3.3.4 Evaluation of Model Performance

This study used five classification models - NB, KNN, RF, SVM, and DNN - to classify eight cancers - BLCA, CESC, COAD, HNSC, KIRP LGG, LIHC, and LUAD. To compare the model performance, first, a confusion matrix was generated and then three different performance metrics - accuracy, precision, and recall - were evaluated. Accuracy is the number of correct predictions made by the model over all kinds of predictions made. True positives (TP) and True Negatives (TN) are the correct predictions. 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. The recall is the number of correct positive results divided by the number of all relevant samples. All scores are calculated from the test data.

3.4 Results and discussion

It is clear from Table 3.2 that 37 lncRNAs produced better classification accuracies ranging between 95% to 98% compared to that of 12,309 lncRNAs (accuracies: 83% to 97%) and 4,786 lncRNAs (accuracies: 89% to 97%).

Table 3.2 shows the values of performance metrics for NB, KNN, RF, SVM, and DNN models using three different sets of lncRNAs - 12,309, 4,786, and 37, respectively. First, the expression profiles of all lncRNA (12,309) were used to classify eight cancer types. After initial reduction of feature size using cutoff, mean expression ≥ 0.3, 4,786 lncRNAs were left for classification. Then feature selection methods RFE and LASSO were used on 4,786 lncRNAs to find the optimum number of features. RFE and LASSO produced 786 and 765 features, respectively. The classification was performed separately using features derived from RFE and LASSO, and results showed 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, were used to classify the tumor types, which resulted in accuracy ranging from 96% to 99%. Since the features derived from RFE performed better, further experiments were conducted to produce a reduced number of features using RFE, such as 200, 100, and 50. The intersection of these three feature sets with LASSO-derived features (765 lncRNAs) resulted in three sets of common features of 129, 68, and 37 lncRNAs, respectively. It is clear from Table 3.2 that 37 lncRNAs produced better classification accuracies ranging between 95% to 98% compared to that of 12,309 lncRNAs (accuracies: 83% to 97%) and 4,786 lncRNAs (accuracies: 89% to 97%).

Table 3.2: Performance comparison of different classifiers with three different sets of features consisting of 12,309, 4,786 and 37 lncRNAs, respectively.

#Features	Model Name	Recall	Precision	accuracy	cost (sec.)
12,309	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)	28.38
	SVM	0.91 (+/- 0.00)	0.92 (+/- 0.00)	0.93 (+/- 0.00)	300.48
	DNN	0.97(+/- 0.01)	0.97(+/- 0.01)	0.97(+/- 0.01)	720.33
4,786	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 (+/- 0.00)	0.95 (+/- 0.01)	0.95 (+/- 0.01)	42.78

	DNN	0.97 (+/- 0.01)	0.97 (+/- 0.01)	0.97 (+/- 0.01)	139.16
37	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

Figure 3.2 shows one of the confusion matrices obtained from the SVM model with a test accuracy of 98% using 37 lncRNAs. Row labels represent the actual labels, and column labels represent the predicted labels. It is clear from the confusion matrix that very few samples were misclassified in each of eight cancer types, which resulted in a high accuracy of 98%. The Receiver Operating Characteristics (ROC) curve with Area Under the Curve (AUC) score for eight different cancers is shown in Figure 3.3, which also supports the high accuracy by producing AUC score close to unity for each cancer type.

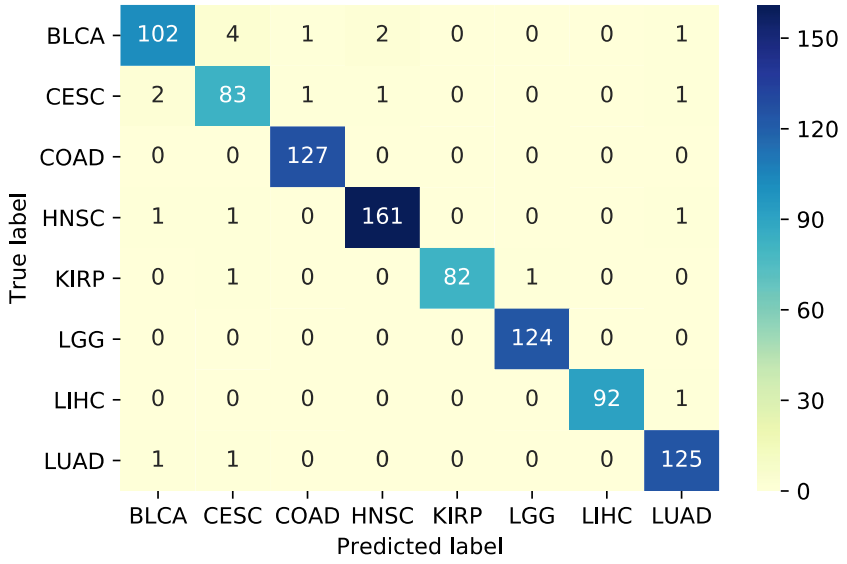

Figure 3.2 Confusion Matrix of SVM Model (Test Accuracy = 98%, Number of Features = 37). X-axis represents the predicted level and y-axis represents the True level.

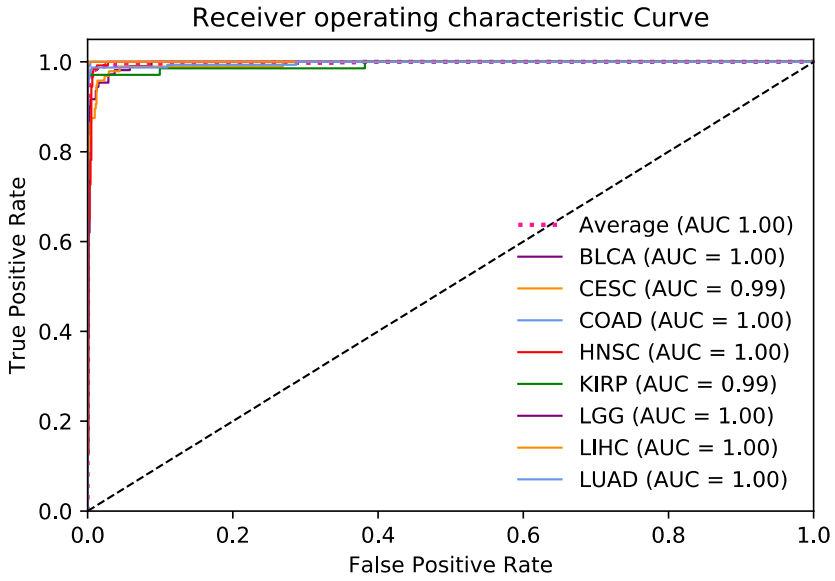

Figure 3.3 ROC curve and AUC scores of different classes from SVM classifier.

Further reduction of features deteriorates the performance considerably. Thus, 37 lncRNAs as shown in Table 3.3 can be considered as the key lncRNAs related to eight cancers considered for analysis in this study.

Table 3.3: 37 key lncRNAs identified in this study

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

Validation Using t-SNE Plot: The results obtained, 37 key lncRNAs, are visually validated using a t-SNE plot. Figure 3.6 shows the t-SNE plot of eight cancer types derived using expression values of 37 lncRNAs identified in the present study. It is clear from this figure that 37 lncRNAs can differentiate eight different cancers. So, the t-SNE plot validated that 37 lncRNAs can be considered the possible key features for the diagnosis and prognosis of eight different cancer types. In the following section, we shared the result of survival analyses.

Validation Using Survival Analyses: Figure 3.7a shows the top 10 lncRNAs with importance scores (six positively correlated and four negatively correlated) selected by LASSO. Figure 3.7b shows the box plot of samples of eight different types of cancer using the expression profile of lncRNA *HOXD-AS2*, one of the top 10 lncRNAs. It is clear that *HOXD-AS2* has distinctive characteristics in eight different cancer types, which makes it a potential biomarker for these cancer types.

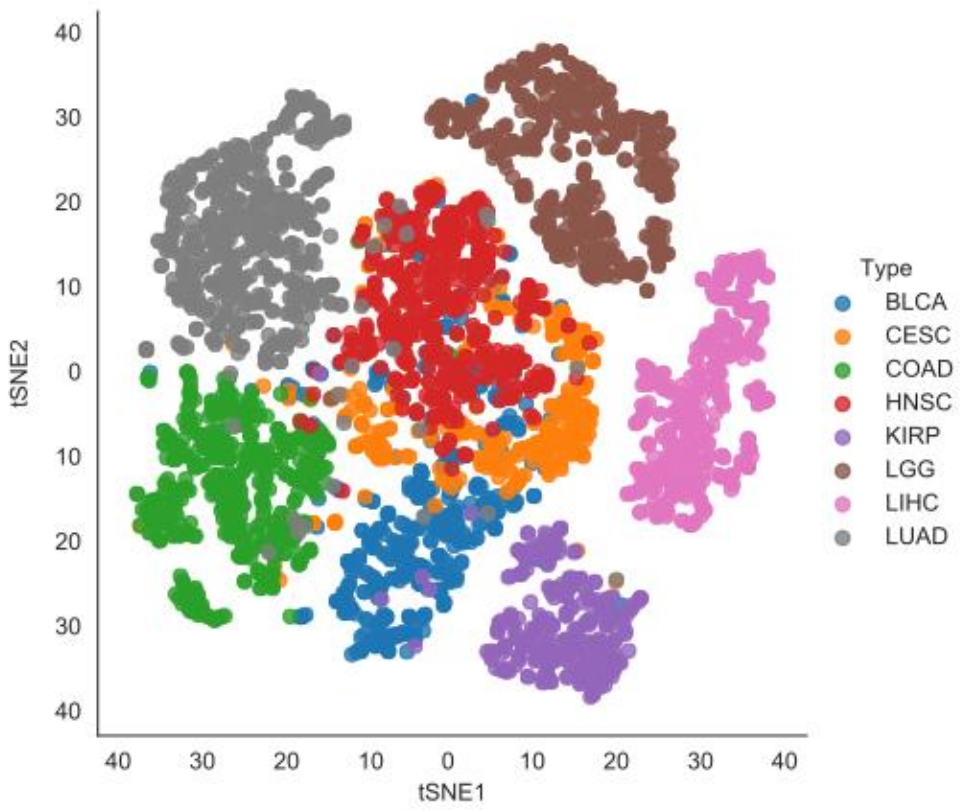

Figure 3.4: t-SNE plot of the samples of eight different cancer types using expression profiles of 37 key lncRNAs.

Figure 3.5(c) shows the survival analysis using positively co-related lncRNA NKX2-1-AS1, which means a patient with high expression (red line) would have a low probability of survival. In contrast, a patient with low expression (Blue line) would have a high probability of survival. Figure 3.5(d) shows survival analysis using negatively co-related lncRNA RP11-435O5.6, which means a patient with a low expression (Blue line) would have a low probability of survival. In contrast, a patient with a high expression (red line) would have a high probability of survival. These two survival analyses evidenced that the first lncRNA acts positively co-related, and the second lncRNA acts as a negatively co-related lncRNA biomarker. Other lncRNAs also provided a similar correlation, which implies that the discovered 37 lncRNAs can be considered the key features in the diagnosis and prognosis of these eight cancers.

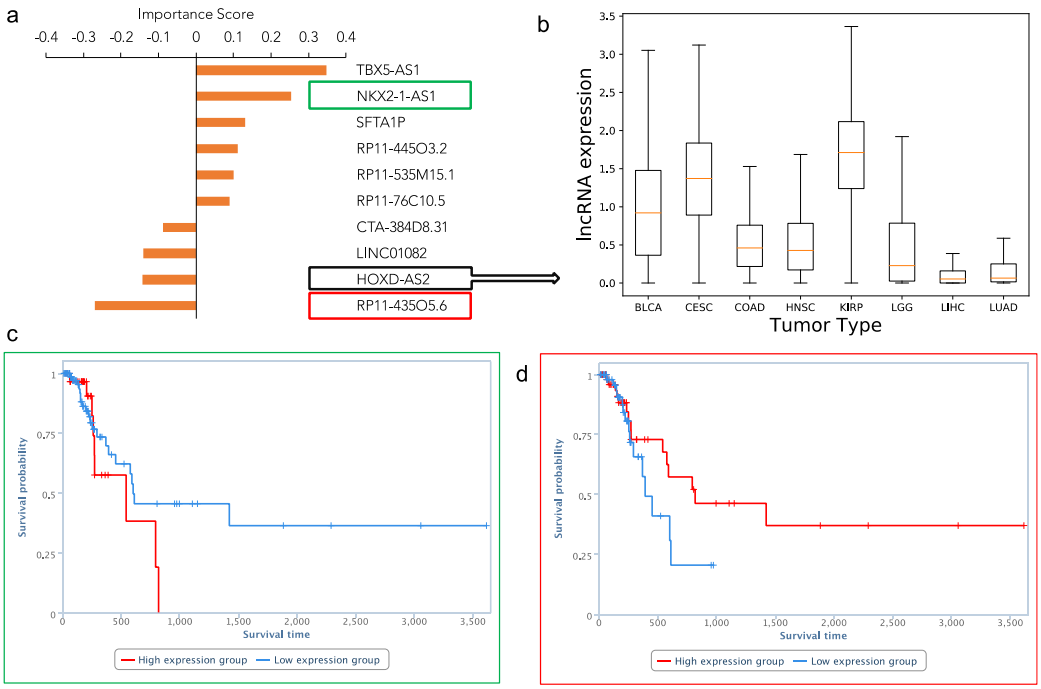

Figure 3.5: Validation of discovered key lncRNAs. a) Top-10 lncRNAs with importance score by LASSO, b) Box plot of expression values of lncRNA *HOXD-AS2* for different cancers, c) Survival analysis using positively co-related lncRNA *NKX2-1-AS1* in BLCA, and d) Survival analysis using negatively co-related lncRNA *RP11-435O5.6* in BLCA. Survival Analysis was done using TANRIC.

CHAPTER 4 FEATURE SELECTION AND CANCER CLASSIFICATION (33 CANCERS)

This chapter contains detailed information regarding the extended experiment for 33 cancers using a high volume of transcriptomic data. It also contains the necessary explanation of a deep learning-based feature selection technique, namely concrete autoencoder. In addition, the performance evaluation of feature selection and cancer prediction is shown in this chapter.

4.1 Introduction

Recent studies indicate that several cancer risk loci are transcribed into long non-coding RNAs (lncRNAs), and these transcripts play key roles in tumorigenesis [49], [50]. The lncRNAs also have key functions in transcriptional, post-transcriptional, and epigenetic gene regulation [9]. Schmitt *et al.* discussed the impact of lncRNA in cancer pathways [10]. They described the involvement of lncRNAs in six hallmarks of cancer such as proliferation, growth suppression, motility, immortality, angiogenesis, and viability [11]. Hoadley *et al.* showed that cell of origin patterns dominate the molecular classification of tumors available in The Cancer Genome Atlas (TCGA) [52]. They used copy number variation, mutation, DNA methylation, RPPA protein, mRNA, and miRNA expression for their analysis. But they did not consider another important molecular signature of cancer, which is lncRNA expression. This work motivated us to investigate the importance of lncRNAs in identifying cancer origins.

Though RNAseq data from TCGA contains a reasonable number of samples, even it poses challenges for classification tasks due to a large number of features (lncRNAs) with respect to the number of samples. Many computational methods fail to identify a small number of

relevant features, rather increasing learning costs and deteriorating performance [54]. It may be argued that the larger the feature set, the better the classification. However, not all of these features will be necessary for optimal classification in a general setting. Only a selected number of significant or relevant features can lead to optimal classification. A large part of the remaining features are not significant and could be either noise, irrelevant to the study, or even redundant [14]. The use of such insignificant features can lead to unwanted computational complexities and deteriorate the model's performance. This is more pronounced when working with high-dimensional data. Thus, it is essential to identify the significant features that can provide us with the optimal classification and clustering. To accomplish this objective, we need a robust method that can eliminate the redundant features and noise that do not carry any information about data labels, thus providing us with only relevant features [84].

Any dataset with an *N*-number of features has 2^N-possible subset of features [14]. In the presence of such a large number of possible combinations, finding the best subset of N features is computationally challenging and expensive [15]. An optimally selected set of features optimizes the performance of classification models and helps alleviate the effect of overfitting and high-dimensionality. Along with these benefits, selecting the appropriate features helps in the easier interpretation of the model and its predictions. On the other hand, the use of gratuitous features can significantly impact the training speeds and the accuracy of the learning models.

Filter, wrapper, embedded methods are the three general classes or types of feature selection techniques. The filter method works by ranking the features using a statistical score assigned to each of them depending on their relevance to the class type. In both

univariate and multivariate filter methods, the interactions among features are disregarded in the selection process. Studies like the ones in Pearson correlation coefficient(PCC), t-statistics(TS) [27], F-Test [28], and ANOVA [29] are examples where the filter method is used. It is observed that these methods are effective for selecting features in high-dimensional data because of the reduced computation expenses. However, they fail to provide good accuracy, as discussed in [30].

As an enhancement, the researcher developed the wrapper-based feature selection method with a learning algorithm and a classifier to find a suitable subset of features. Initially, a random solution is generated, following which an objective function is maximized using black-box type optimization methods [31] like simulated annealing [32], particle swarm optimization [33], genetic algorithm [34], and ant colony optimization [35]. The iterative evaluation of every candidate subset of the features by a wrapper method leads to the identification of a strong relationship between features, however, with an increase in the computational expense.

Embedded feature selection methods, on the other hand, reduce computational costs because these are used as a part of the learning phase. Well-known embedded methods, which are considered as the state-of-the-art, are least absolute shrinkage and selection operator (LASSO) [36], recursive feature elimination with support vector machine estimator (SVM-RFE) [22], [37], [38], random forest [39], [40], Adaboost [41] , KNN [42] , and autoencoder [43].

In general, the use of feature selection is worthwhile when the whole set of features is difficult to collect or expensive to generate [46]. For example, in TCGA, the lncRNA expression profile dataset contains more than 12 thousand features (lncRNAs) for each of

33 different cancers, and it is expensive to generate this data. Consequently, it is important to answer the question: *Is there a set of salient features (lncRNAs) capable of identifying the origin of 33 cancers?*

The distribution of the number of samples for 33 cancers in TCGA is highly imbalanced, ranging from 36 for CHOL cancer to 1089 for BRCA. Any supervised feature selection approach will be biased to heavy groups. To solve this problem, we need a robust unsupervised feature selection approach to find appropriate features related to 33 different cancers.

Feature selection works differently compared to the standard dimension reduction techniques such as principal component analysis (PCA) [58] and autoencoders [47]. The standard dimension reduction methods can preserve maximum variance with a highly reduced number of latent features. This means that PCA and standard autoencoder do not provide the original features in the reduced dimension, or these work as a black-box. For the real application of diagnosing the origin of cancer, a tool should tell what actual or measurable features are relevant. Recently, few deep learning-based feature selection methods showed little improvement in selecting original features in both settings, supervised and unsupervised [59]–[61].

In this research, we proposed to use concrete autoencoder (CAE) [46], a deep learning-based unsupervised feature selection algorithm, to discover the relevant lncRNAs capable of identifying the origin of different cancers. The CAE takes advantage of both (a) AE, which can achieve the highest classification accuracy, and (b) concrete relaxation-based feature selection [85], [86], which is capable of selecting actual features instead of latent features. The proposed model filtered the key lncRNAs from 12,309 lncRNAs related to

33 different cancers. The key lncRNAs discovered using the proposed CAE method produced higher classification accuracy and better diagnosis of cancer origin than the state-of-the-art embedded feature selection approaches – LASSO, RF, and SVM-RFE - while using a small number of lncRNAs.

4.1 Materials and Methods

The overall process flow diagram is illustrated in Figure 4.1. The following subsections describe the different aspects of the process flow diagram: (a) Data Preparation, (b) Feature Selection, (c) Reconstruction and Classification, and (d) Evaluation and Validation.

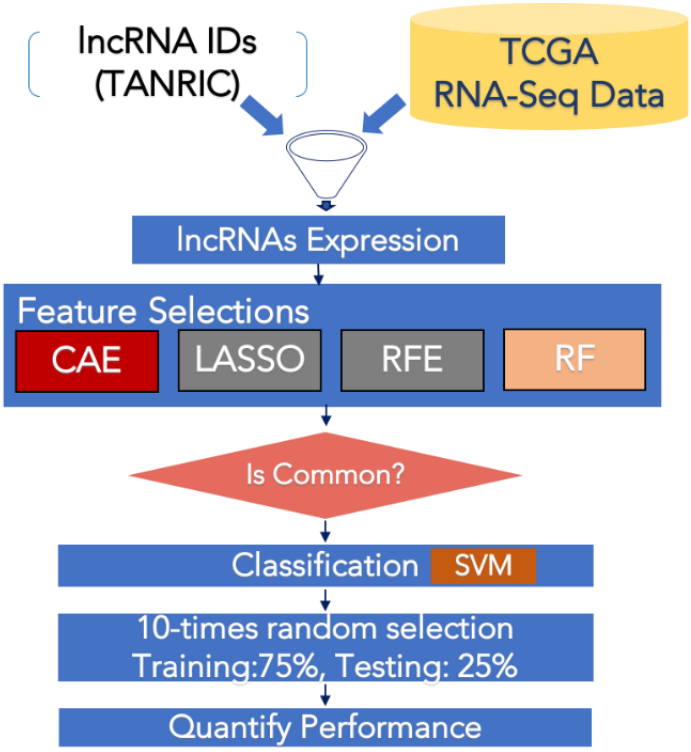

Figure 4.1: Process flow diagram. Data Preparation, Feature Selection, Classification, and Validation.

4.2 Data Preparation

We downloaded the expression profiles and clinical data for 33 different cancers from the UCSC Xena database [87] to identify the cancer-associated lncRNAs. This dataset contains

expression profiles of about 60 thousand RNAs, including coding genes (mRNAs) and non-coding genes (lncRNAs and miRNAs). In this study, only the expression profiles of lncRNA (n = 12,309) were considered for analysis and model evaluation. It should be noted that this study was based on cancer patients only. So, the normal samples available in the same cancer were removed. The final dataset contains 9,566 cancer patients. The cancer-specific distributions based on the 75/25 (training/testing) split are shown in Figure 4.2. To achieve good training performance, each lncRNA expression was processed using a min-max normalization method.

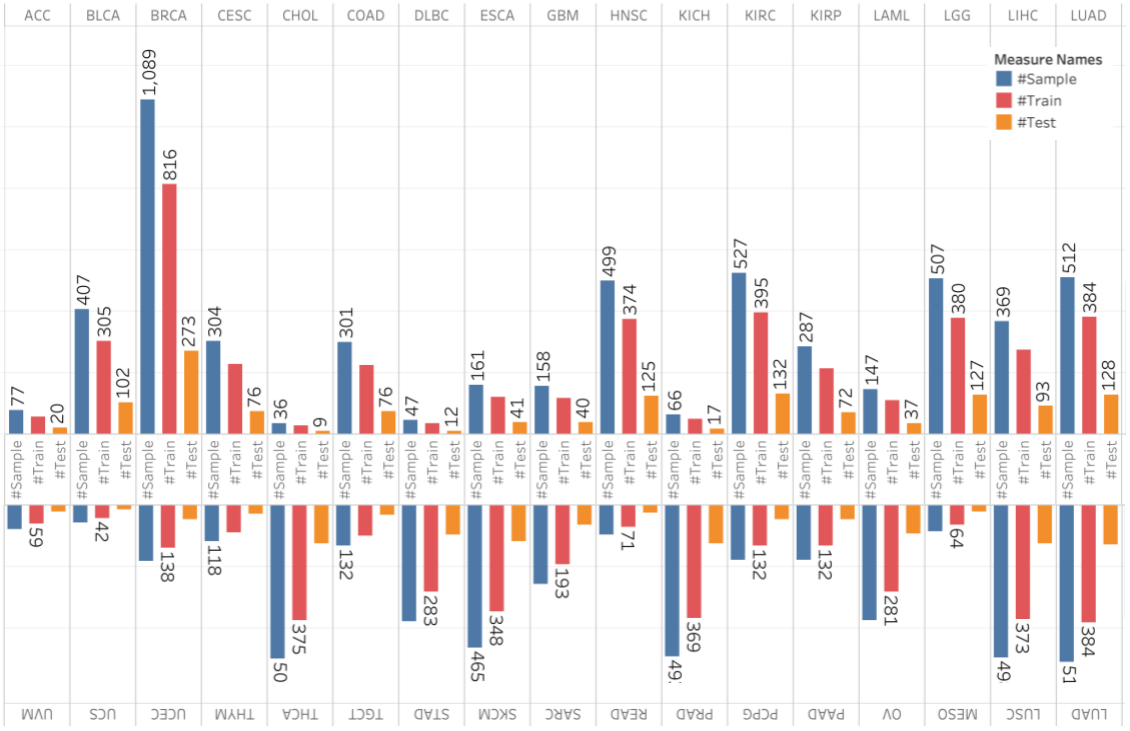

Figure 4.2: Sample distribution for 33 cancers along with 75-25 split for training and testing.

4.3 Feature Selection

For selecting important features (lncRNAs), a state-of-the-art deep learning-based unsupervised algorithm, Concrete Autoencoder (CAE), was used. To compare the results of CAE, three frequently used embedded feature selection models, including LASSO,

Random Forest (RF), and Support Vector Machine with Recursive Feature Elimination (SVM-RFE), were used. The following subsections briefly describe the implementation of feature selection algorithms.

4.3.1 Concrete Autoencoder (CAE)

Concrete autoencoder (CAE) proposed by Abid *et al.* [46] is a variation of the original autoencoder (AE) [47], which is used for dimension reduction. The motivation behind selecting CAE in the present study is that it takes advantage of both AE (which can achieve the highest classification accuracy) and concrete relaxation-based feature selection (capable of selecting actual features instead of latent features). An AE is a neural network that consists of two parts: (a) an encoder that selects latent features and (b) a decoder that uses selected latent features to reconstruct an output that matches the input with minimum error. In CAE, instead of using a sequence of fully connected layers in the encoder, a concrete relaxation-based feature selection layer is used where the user can define the number of nodes (features to be selected), *k* as shown in Figure 4.3. This layer selects a probabilistic linear arrangement of input features while training, which converges to a discrete set of *k* features by the end of the training phase, which is subsequently used in the testing phase.

Let's *p(x)* is a probability distribution over a *d*-dimensional vector. The objective is to identify a subset of features, S≡{1…k} of size |*S*|=*k*. Also, learning a reconstruction function $f_r(.): \mathbb{R}^{\mathrm{k}} \xrightarrow{\Delta} \mathbb{R}^{\mathrm{d}}$, such that the loss between original sample $\boldsymbol{x}$ and reconstructed sample $f_r(x_S)$ is minimized as stated in Eq. 1,

$$argmin_{S,r}\, E_{p(x)}[\|f_r\,(x_S\,) - x\|_2] \ldots\ldots\ldots\ldots\ldots (1)$$

where $x_s \in \mathbb{R}^k$ consists of only selected features x_i s.t. $i \in S$. Note that samples are represented in a *2D* matrix, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, and aim is to pick *k* columns of $\boldsymbol{X}$ such that sub-matrix $X_s \in \mathbb{R}^{n \times k}$.

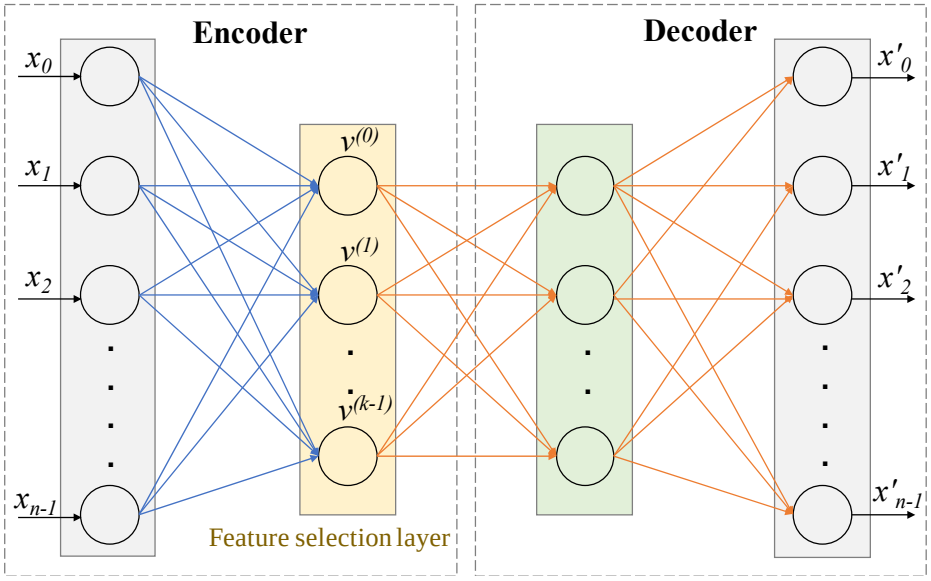

Figure 4.3: Architecture of Concrete Autoencoder. CAE architecture consists of an encoder and a decoder. The layer after input layer of encoder is called concrete feature selection layer shown in yellow. This layer has *k* number of node where each node is for each feature to be selected. During the training stage, the $\boldsymbol{i^{th}}$ node $\boldsymbol{v^{(i)}}$ takes the value $\boldsymbol{X^T}$ *f(i)*, where *f(i)* is the corresponding weight vector of node *i*. During testing stage, these weights are fixed and the element with the highest value is selected by the corresponding $\boldsymbol{i^{th}}$ hidden node. The architecture of the decoder remains the same during training and testing.

Then, the selected feature set x_s can be used to reconstruct the original matrix $\boldsymbol{X}$ and classify the cancer types. In the feature selection layer of CAE (Figure 4.3), the original features are selected based on the temperature of this layer which is tuned using an annealing schedule. More specifically, the concrete selector layer identifies *k* important features as the temperature decreases to zero. For reconstructing the input, a simple decoder similar to the ones associated with a standard AE is used. The temperature τ of the random variable in the selector layer has a significant impact on forming the output of each node. Initially, when τ is high, search space is large since it considers a linear combination of all

features, as shown in Figure 4.4(a). In contrast, the selector layer will not be able to search all possible combinations of features at low τ, and thus, the model converges to a poor local minimum. This means that as temperature goes down, a small number of features are necessary for stable convergence. Annealing or gradual decrease in temperature avoids the model convergence to a poor local minimum. The effect of annealing in feature selection is shown in Figure 4.4(a). For example, at starting temperature, τ_s, the number of input features is 10, and the number of features to be selected, *k*, is 3. At the next epoch, when the temperature is τ_{s+1}, the number of possible features reduces to 6. After some epochs, when the temperature reaches its lower bound τ_{stop}, the number of features further reduces to 3, equal to *k*, the user-specified number of features to be selected. Instead of using a fixed temperature, a simple annealing scheduling scheme is used for every concrete variable. It starts with a user-defined high temperature (τ_s) and steadily lowers the temperature, until it touches the end bound (τ_e), by every epoch as follows:

$$\tau_{(e)} = \tau_s \, (\tau_N / \tau_s)^{e/n} \text{.....(2)}$$

where τ_e is the temperature at epoch *e*, *N* refers to the total number of epochs. Adam optimizer with a learning rate of 0.001 is used for all the experiments for CAE. Figure 4.4(b) shows an example of the effect of temperature in reducing the loss while training the CAE to select a reduced set of 100 features from the original feature space of 12,309 lncRNAs. The starting temperature of CAE was set to 10, and it ends at 0.01. The model was trained for the same number of epochs (*n* = 100) to control the performance.

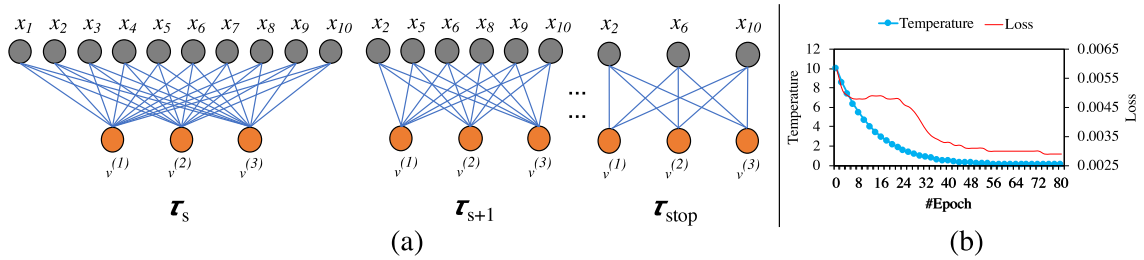

Figure 4.4: Effect of annealing in reducing search space. (a) An example: at starting temperature $\boldsymbol{\tau_s}$, the number of input features is 10, and the number of features to be selected is k = 3; at the next epoch when the temperature is $\boldsymbol{\tau_{s+1}}$, the number of possible features reduces to 6; after some epochs, when the temperature reaches to its lower bound $\boldsymbol{\tau_{stop}}$, the number of features further reduces to 3, equal to k. (b) Effect of temperature change in reducing the loss while training the concrete autoencoder on lncRNA expression data with k = 100 features to be selected from original feature space of 12,309 lncRNAs.

4.3.2 Implementation of LASSO

To select the important features, LASSO 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 the parameter λ, the larger number of coefficients are shrunk to zero, and a smaller number of features are selected. In this experiment, the optimized λ was set in a range of 0.005 to 0.01 to select a different number of features ranging from 10 to 500.

4.3.3 Implementation of RF

Random Forest works based on a tree structure that employs ensemble. RF consists of a number of decision trees. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two branches, so similar response values end up in the same set. The optimal condition is chosen based on impurity. For classification, it is either *Gini* impurity or information gain/entropy. Thus, when the tree is fully developed, it can compute how much each feature decreases the weighted impurity on the tree. For forest,

the impurity decrease from each feature can be measured as a feature rank. The feature importance is calculated as the sum over the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits [45]. RF needs three parameters to be tuned: (i) *n_estimator:* number of estimators, also known as the number of trees in the forest, (ii) *min_sample_split:* minimum number of nodes required to split, and (iii) *criterion:* impurity to measure the quality of a split. In *GridSearch*, the ranges of values assigned to tune *n_estimator* and *min_sample_split* were from 2 to 300 and 1 to 150, respectively. Two options, *Gini* and *entropy*, were used to optimize the impurity parameter *criterion*. The optimum values or options for *n_estimator*, *min_sample_split, and criterion* found by *the GridSearch* method are 100, 120, and *Gini*, respectively.

4.3.4 Implementation of SVM-RFE

Recursive feature elimination is a recursive method in which less important features are eliminated in every iteration. In the RFE technique, SVM was used as the estimator in the present study. A linear kernel with a regularization parameter $C = 0.05$ was used. C controls the tradeoff between the error and norm of the learning weights. The *GridSearch* algorithm was used to estimate the best set of parameters for SVM. In every iteration of RFE, the number of dropped features was set to 100.

LASSO, RF, and SVM-RFE were implemented using the scikit-learn framework [81], whereas CAE was implemented using TensorFlow [88] based deep learning framework, Keras [89]. Experiments are parallelized on NVIDIA Quadro K620 GPU with 384 cores and 2GB memory devices. To avoid overfitting, the dataset was split into the train and test set according to the 75/25 ratio, as shown in Figure 4.2. The training set was used to estimate the learning parameters, and the test set was used for performance evaluation.

4.4 Reconstruction and Classification

The feature selection capability of CAE is compared with standard autoencoder (AE), LASSO, RF, and SVM-RFE in two different ways: (a) reconstruction of all input features using the selected features and (b) classification performance in classifying 33 different cancer types using the selected features. A subset of features by varying *k* from 10 to 500 was extracted using CAE. For the comparison to be fair and along the same grounds with CAE, the same number of lncRNAs were selected using all other models. The SVM was used for classifying 33 cancer types using the selected features. We trained a linear regressor with no regularization to reconstruct all the input features from the selected features.

4.5 Evaluation and Validation

Five evaluation metrics have been used to record the classification and reconstruction performance, such as accuracy, precision, recall, f1 score, and mean squared error (MSE). Accuracy is the number of correct predictions made by the model over all kinds of predictions made. Precision is the number of correct positive results divided by the number of positive results predicted by the model. It indicates the predicted positive portion of the samples. The recall is the number of correct positive results divided by the number of all relevant samples. F1 score is the harmonic mean of precision and recall. Reconstruction performance measure, MSE, was calculated using linear regression on the test set.

All classification performance metrics were measured by comparing the predicted labels with the true labels of independent test samples. The optimal set of features was selected based on two criteria: (a) the number of features should be as few as possible, and (b) classification accuracy using the selected features should be > 90%. The final list of key

lncRNAs is selected from the union of features derived from the binary intersection of four approaches,

$(CAE \cap LASSO) \cup (CAE \cap RF) \cup (CAE \cap SVMRFE) \cup (LASSO \cap RF) \cup (LASSO \cap SVMRFE) \cup (RF \cap SVMRFE)$……..…..(3)

Then each lncRNA discovered in this study was cross-checked with existing literature whether it is already a known biomarker or not. The capability of selected lncRNAs in pan-cancer classification was visually validated using the unsupervised visualization technique t-SNE [26]. To validate the prognostic performance of discovered lncRNAs, survival analysis of cancer patients using the Kaplan-Meier [90] method was performed [91].

4.6 Results

A series of experiments were conducted to compare the performance of CAE with other state-of-the-art feature selection methods such as standard autoencoder, LASSO, RF, and SVM-RFE. Each of these methods was used to select features in the range of 10 to 500 lncRNAs. The expression profiles of these lncRNAs were then used to train a linear classifier SVM to classify 33 cancer types.

4.6.1 Classification Performance Using Selected Sets of Features

Figure 4.5 shows classification performance using different sets of selected features. The initial stages of the experiments were performed with a smaller subset of the selected features as we wanted to understand the performance of the models being compared. The optimal classification performance with CAE (accuracy > 90% with the smallest number of features) was observed with about 100 features. Beyond this point, the increase in performance was not significant.

It is clear from Figure 4.5 that, for all sets of selected features, CAE performed better than LASSO, RF, and SVM-RFE in terms of four evaluation matrices, including accuracy, precision, recall, and f1 score. Of course, it could not beat the standard AE, as expected. It is noticeable that even with a smaller number of features (say 10), the accuracy of CAE was close to 70%, whereas LASSO (55% accuracy), RF (38% accuracy), and SVM-RFE (50% accuracy) showed poor results for the same number of features. The trend remains the same with the increase in the number of features.

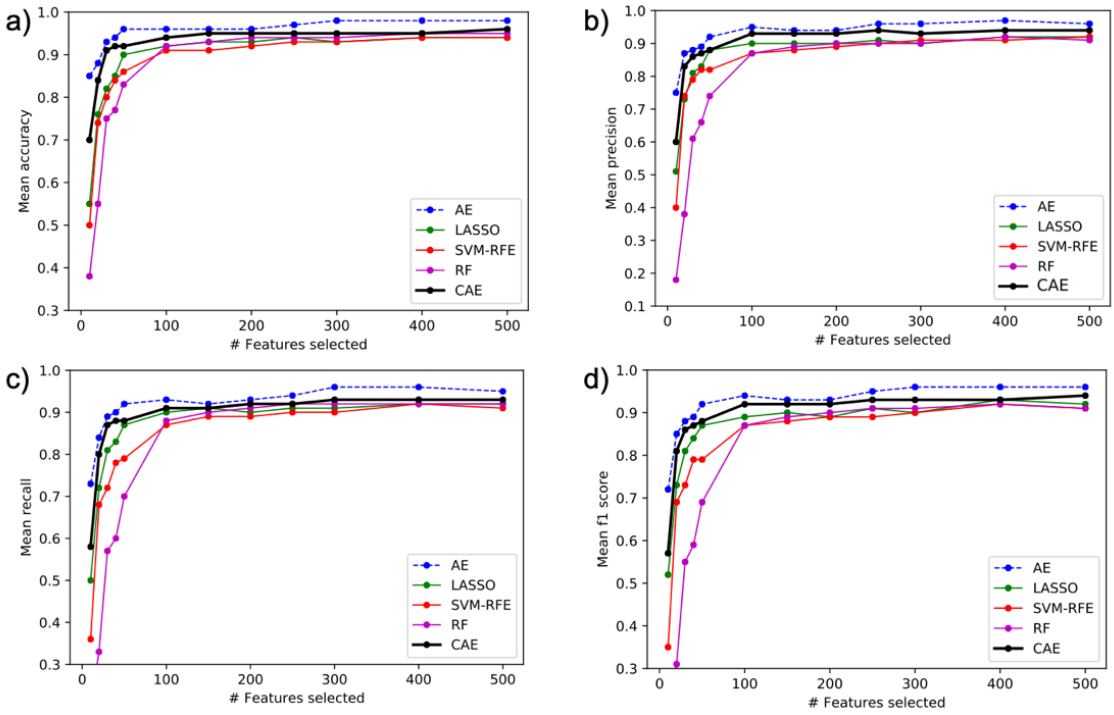

Figure 4.5: Classification performances of the proposed method using selected features. Comparison of CAE with other feature selection methods. Throughout all values of k tested on both (a) Accuracy, (b) Precision, c) Recall, and d) f1 score; CAE has the highest classification performance after AE.

4.6.2 Reconstruction Performance of Feature Selection Algorithms

Figure 4.6 shows the comparison of reconstruction performance among five feature selection algorithms. Note that AE selects latent features, whereas the other four algorithms select actual features. The CAE starts with an MSE of 60 and quickly reduces to a value of

less than 10 within the top 100 features, as shown in Figure 4.6. It is also clear from this figure that CAE has a lower reconstruction error compared to LASSO, RF, and SVM-RFE for any set of selected features. Again, CAE cannot beat AE, as expected, since AE uses latent features.

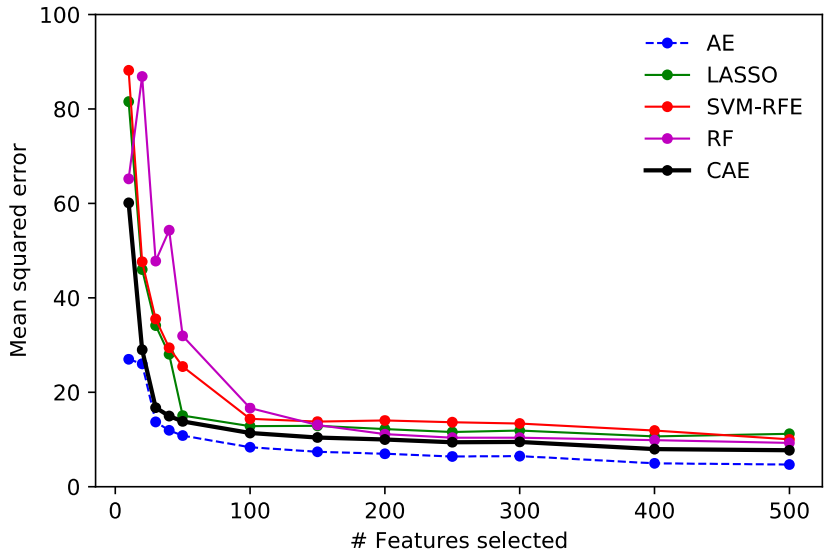

Figure 4.6: Reconstruction mean squared error for different number of features selected by different models.

4.6.3 Combined Set of Features

Based on the performance of CAE, a set of 100 lncRNAs (features) produced an optimal classification. So, to produce a stable set of features for this problem, each of the four feature selection algorithms was run to extract 100 features. The final list of 69 key lncRNAs resulted from the union of features derived from the binary intersection of four approaches as mentioned in eq. 3. Figure 4.7 shows the Venn diagram of the common features extracted by four algorithms. It is clear from the Venn diagram that 67 (100 – 23) out of 69 lncRNAs came from CAE, which dictates the superiority of CAE.

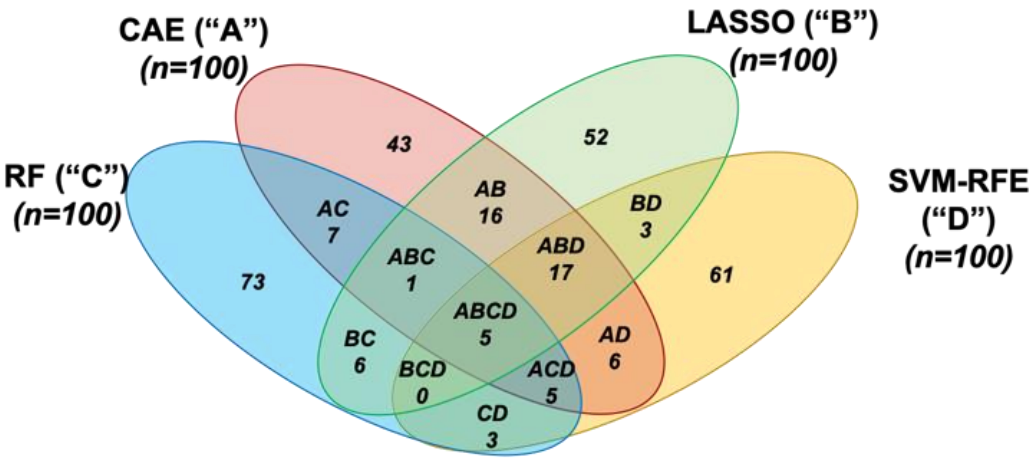

Figure 4.7: Common features selected by different methods.

Table 4.1 shows the comparison of classification and reconstruction performance among the four approaches. Selected features from each method were passed to a linear regressor for reconstructing the input features. It is clear from this table and Figure 4.6 that CAE is more resilient to errors. In comparison, this error is more pronounced in the other competing methods.

Table 4.1: Classification and reconstruction performances using combined lncRNAs and selected lncRNAs using different models.

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**MSE**
Combined	0.93 ±0.02	0.91±0.01	0.91±0.02	0.9±0.03	13.46±0.10
LASSO	0.92±0.01	0.87±0.02	0.88±0.02	0.87±0.01	13.84±0.08
SVM-RFE	0.85±0.03	0.85±0.02	0.82±0.03	0.83±0.02	25.98±0.08
RF	0.89±0.02	0.86±0.03	0.81±0.03	0.81±0.03	22.91±0.12
CAE	0.93±0.01	0.89±0.01	0.9±0.02	0.9±0.02	12.23±0.09

Combined features are also used for classification and showed the highest classification performance as compared to other methods, Table 4.1.

4.6.4 Visual Validation of Selected Features

Figure 4.8 shows the clustering capability of discovered 69 lncRNAs expression profiles using the t-SNE plot [26]. It is clear from the t-SNE plot that the selected lncRNAs can discover the heterogeneity among 33 different cancers. So, the newly identified lncRNAs can be considered essential features for diagnosis, prognosis, and therapeutic target for different cancers. Then each lncRNA was cross-checked with the existing literature whether it is already a known biomarker. Of the 69 lncRNAs, 38 were found in existing literature as known biomarkers for different cancers, as shown in Table 4.2. The remaining 31 lncRNAs were novel discoveries based on the ***lncRNA disease*** **database v2.0.**

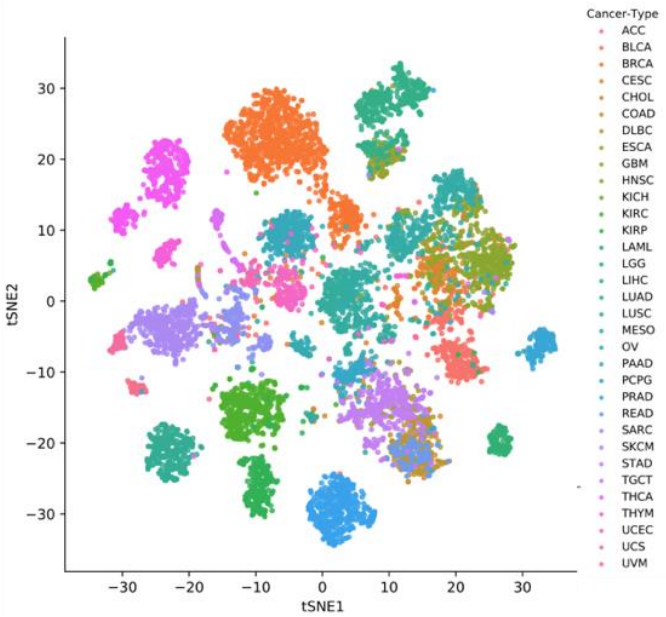

Figure 4.8: t-SNE using top 69 lncRNAs where each dot represents a cancer sample and each color represents a cancer type.

4.7 Discussion

It is clear from the literature that lncRNAs play a key role in cancer development. More research is needed to identify cancer-specific lncRNAs. Existing methods used co-expression networks such as lncRNA-mRNA or lncRNA-miRNA-mRNA. As per our knowledge, no study used only lncRNA expression to classify cancer types except our previous work [92], where feature extraction was not considered.

In this study, we identified 69 key lncRNAs that can identify the origins of 33 different cancers. When compared against the existing literature, 38 (55%) lncRNAs have been reported as important prognostic biomarkers for various cancers. Since the proposed method can identify already known lncRNA biomarkers, it can be concluded that the newly discovered 31 lncRNAs can be considered novel biomarkers for cancers. Survival analysis suggests that some of 31 lncRNAs are novel biomarkers, as shown in Figure 4.9. Many studies have been conducted using mRNA expression for predicting cancer types and developing screening tools. No such tools are available that used expression profiles of lncRNAs. Hence, the identified 69 lncRNAs can be used as a screening tool for cancer diagnosis and as therapeutic targets for different cancers, for which further studies are required.

Table 4.2: 69 key lncRNAs identified in this study.

Known lncRNAs (n=38)
AC005083.1, AC008268.1, AC093850.2, AC133528.2, AFAP1-AS1, CASC9, CRNDE, DNM3OS, EMX2OS, FAM83H-AS1, FENDRR, GATA2-AS1, GATA6-AS1, H19, HAGLR, HAND2-AS1, HCG11, HNF1A-AS1, LHFPL3-AS1, LINC00261, LINC00511, LINC01116, LINC01133, LINC01139, LINC01158, MALAT1, MEG3, MNX1-AS1,

NR2F1-AS1, PIK3CD-AS2, PTCSC2, SATB2-AS1, SFTA1P, TRPM2-AS, UCA1, VPS9D1-AS1, XIST, ZNF667-AS1
Novel lncRNAs (n=31) Based on *lncRNA disease* v2.0 (http://www.rnanut.net/lncrnadisease/) dated: July 2020
AC005082.12, AC079630.4, AP001626.1, CECR7, CTA-384D8.31, CTD-2377D24.4, CTD-3032H12.2, GATA3-AS1, HOXA10-AS, HOXA11-AS, HOXD-AS2, LINC00958, LINC01082, LINC01272, MIR205HG, NKX2-1-AS1, RP1-288H2.2, RP1-60O19.1, RP11-1017G21.5, RP11-1055B8.3, RP11-264B14.2, RP11-3P17.5, RP11-465B22.8, RP11-47A8.5, RP11-807H17.1, RP3-416H24.1, SLCO4A1-AS1, TBX5-AS1, U47924.27, U91324.1, ZFPM2-AS1

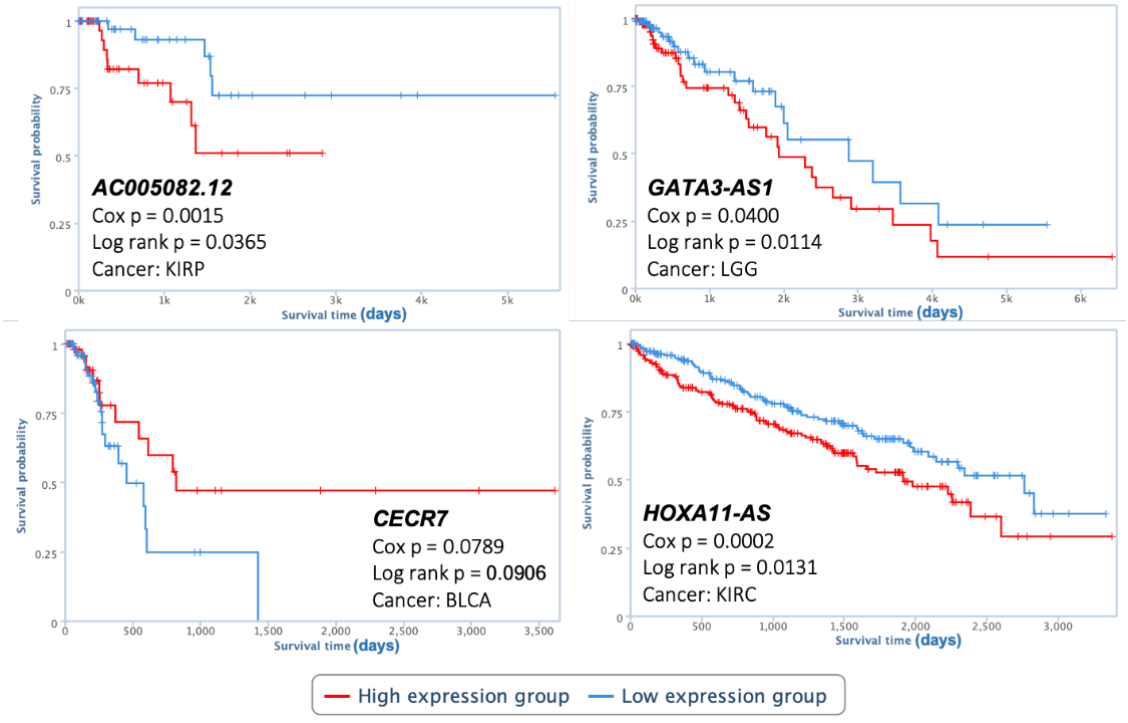

Figure 4.9: Kaplan-Meier survival analysis curve of high-risk and low-risk patients evaluated on novel lncRNA (*AC005082.12, CECR7, GATA3-AS1*, and *HOXA11-AS*).

CHAPTER 5 FEATURE SELECTION AND CANCER CLASSIFICATION (GLYCOME GENES)

This chapter provides a detailed explanation of how glycome genes performed an important role in cancer progression. It also discussed how the proposed feature selection framework could identify important glycome biomarkers for different cancers.

5.1 Introduction

One of the most ubiquitous pathways in nature is cell glycosylation. Post-translational glycosylation of proteins is a common cellular activity, wherein most if not all proteins are glycosylated [93]. While adding structure and stability, protein glycosylations also provide binding motifs for other molecular partners (e.g., Lectins). They often offer physical subtleties that impact protein complexing, membrane/cytosolic dynamics, and functional activity. In cancer, these biological characteristics imparted by cellular glycosylation are fundamentally aberrant due to variances in the 'glycome' gene [94]–[96]. Altered protein glycans and their glycan-modifying enzymes are now considered key features of cancer. Intensive efforts are underway to understand better how aberrant glycosylation can facilitate tumorigenicity, tumor progression, and metastatic behavior [93]. Considering the breadth and mounting evidence for the key role of aberrant glycosylations in cancer progression, we speculate that distinct glycome gene signatures align with a particular cancer glycosylation pattern originating from a particular cell lineage.

Many computational methods fail to identify a small number of relevant features, rather increasing learning costs and deteriorating performance [54]. It may be argued that the larger the feature set, the better the classification. However, not all of these features will be necessary for optimal classification [22], [76]. Only a selected number of significant or

relevant features can lead to optimal classification. Many of the remaining features are not significant and could be either noise, irrelevant to the study, or even redundant [14]. Such insignificant features can lead to unwanted computational complexities and deteriorate the model's performance. This is more pronounced when working with high-dimensional data. Thus, it is essential to identify the set of significant features that can provide us with the optimal classification and clustering. To accomplish this objective, we need a robust method that can eliminate the redundant features and noise that do not carry any information about the data labels, thus providing only relevant features [84].

The problem comes with highly imbalanced distribution of data ranging from 36 for CHOL cancer to 1089 for BRCA. Any supervised feature selection approaches such as LASSO, RF, and RFE will be biased to heavy groups. We need a robust unsupervised feature selection approach to find appropriate features that can differentiate 33 different cancers.

Over the past decade, many unsupervised feature selection algorithms have been developed. The popular algorithms using regularization as the means for selecting discrete features are Multi-Cluster Feature Selection (MCFS) [97], Unsupervised Discriminative Feature Selection (UDFS) [98], and Autoencoder Feature Selector (AEFS) [99]. Recently, Abid et al. [23] developed Concrete Autoencoder (CAE) without resorting to regularization. Rather, they used a continuous relaxation of the discrete random variables, the Concrete distribution [85]. MCFS [97] uses regularization to isolate the features preserving the clustering structure in the data. UDFS [98] incorporates discriminative analysis and $l_{2,1}$-norm minimization on a set of weights applied to the input to select features most useful for local discriminative analysis. AEFS [99] uses $l_{2,1}$ regularization

on the weights of the encoder that maps the input data to a latent space and optimizes these weights for their ability to reconstruct the original input.

The CAE [23] is an end-to-end differentiable method for global feature selection and efficiently identifies a subset of the most informative features. It takes advantage of both (a) autoencoder (AE), which can achieve the highest classification accuracy, and (b) relaxation of the discrete random variables, the Concrete distribution [8], which is capable of selecting actual features instead of latent features. It has also been shown that CAE performs better than MCFS, UDFS, and AEFS in selecting discrete features [23], which motivated us to use CAE for feature selection in this study. The CAE filtered a shorter list of glycome genes related to 33 different cancers from the original larger list.

5.2 Materials and Methods

5.2.1 Data Preparation

The expression profiles and clinical data for 33 different cancers were downloaded from the UCSC Xena database [100]. This dataset contains expression profiles of about 60 thousand RNAs, including coding genes (mRNAs) and non-coding genes (lncRNAs and miRNAs). This study considered the expression profiles of glycome-related genes (n = 498) for analysis and model evaluation. The glycome genes were procured from the study by Sweeney *et al.* [93]. Table 5.1 shows the distribution of glycome genes in 12 different categories at different levels of analysis. The original list consists of 696 genes with some duplicates. After removing duplicates, the unique list consists of 529 glycome genes. Of 529, 498 genes have expression profiles for all the samples for 33 cancers, which were used to select a reduced list of features. It should be noted that this study was based on cancer patients only. So, normal samples available in the same cancer were removed. The

final dataset contains 9,566 cancer patients. The cancer-specific distributions based on the 75/25 (training/testing) split are shown in Figure 5.1. Each mRNA expression was processed using a min-max normalization method to achieve good training performance.

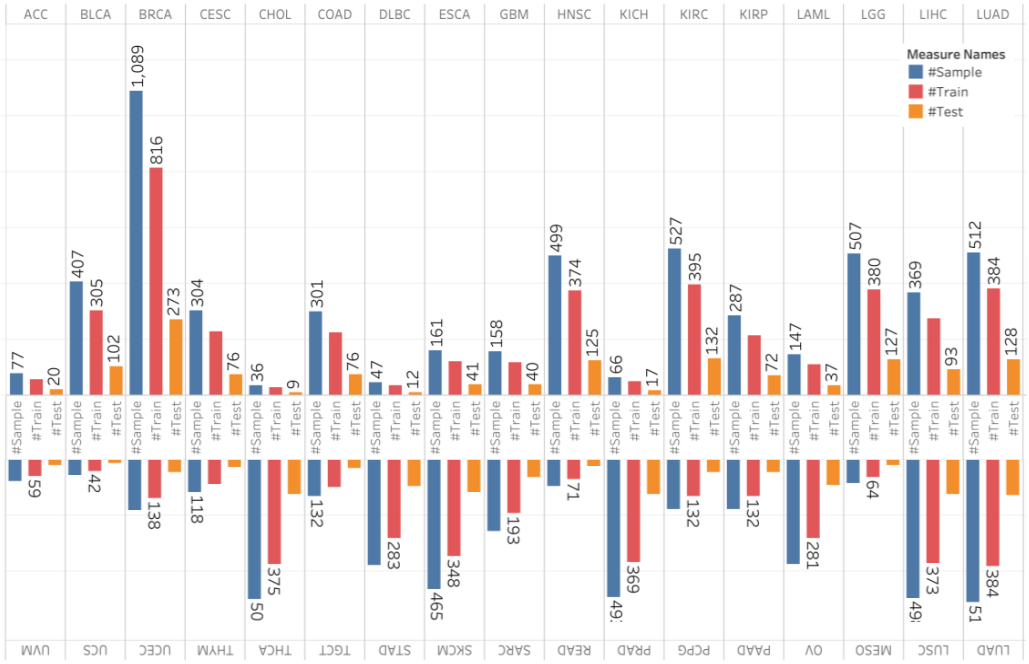

Figure 5.1: Sample distribution for 33 cancers along with 75-25 split for training and testing.

5.2.2 Feature Selection

It is clear from Figure 5.1 that the distribution of cancer samples is highly imbalanced, ranging from 36 for Cholangiocarcinoma (CHOL) to 1089 for Breast Cancer (BRCA). Since the data is highly imbalanced, a choice of supervised feature selection will result in highly biased results toward heavy groups. So, for selecting important features (glycome genes), a state-of-the-art deep learning-based unsupervised algorithm, Concrete Autoencoder (CAE), was used. The CAE takes advantage of both Autoencoder (AE) [47],

capable of producing the highest accuracy, and Concrete Relaxation [85], capable of selecting actual features instead of latent features.

Table 5.1: Distribution of glycome genes among 12 different categories. Original dataset: 696 glycome genes with some duplicates. Unique list: 529 genes. Feature selection experiment: 498 genes used.

Category	**Original**	**Unique**	**Experiment**
Adhesion Molecule	9	7	7
CBP:C-Type Lectin	105	80	74
CBP:I-Type lectin	27	21	20
Galectin	14	13	12
Glycan Degradation	87	61	59
Glycosyltransferases	256	199	187
Glycoproteins	53	38	31
Intracellular protein transport	13	8	8
Miscellaneous	8	6	6
Nucleotide Sugar Transporters	72	57	57
Proteoglycans	41	31	29
Sulfotransferases	11	8	8
Total	**696**	**529**	**498**

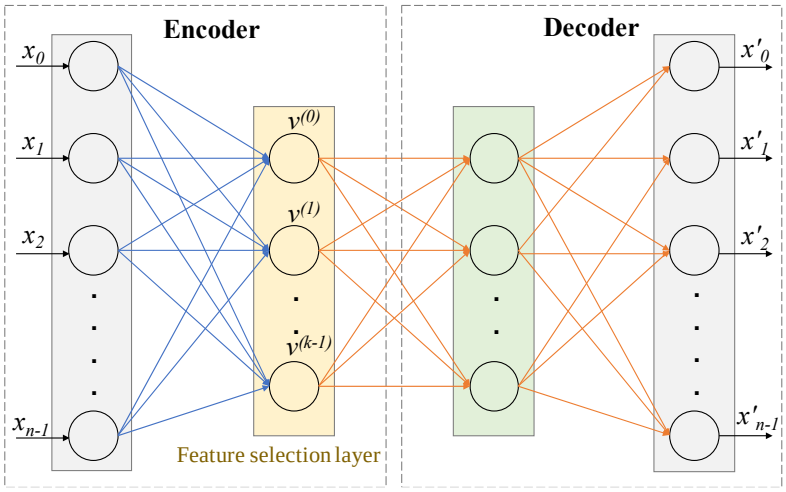

Figure 5.2: Architecture of Concrete Autoencoder. CAE architecture consists of an encoder and a decoder. The layer after the encoder's input layer is called the concrete feature selection layer, as shown in yellow. This layer has k number of nodes where each node is for each feature to be selected. During the training stage, the $\boldsymbol{i^{th}}$ node $\boldsymbol{v^{(i)}}$ takes the value $\boldsymbol{X^T}$ *f(i)*, where *f(i)* is the corresponding weight vector of node *i*. During the testing stage, these weights are fixed, and the element with the highest value is selected by the corresponding $\boldsymbol{i^{th}}$ hidden node. The architecture of the decoder remains the same during training and testing.

The concrete autoencoder (CAE) proposed by [23] is a variation of the original autoencoder (AE) [47], which is used for dimension reduction. An AE is a neural network that consists of two parts: (a) an encoder that selects latent features and (b) a decoder that uses selected latent features to reconstruct an output that matches the input with minimum error. In CAE, instead of using a sequence of fully connected layers in the encoder, a concrete relaxation-based feature selection layer is used where the user can define the number of nodes (features), *k*, as shown in Figure 5.2. This layer selects a probabilistic linear arrangement of input features while training, which converges to a discrete set of *k* features by the end of the training phase, subsequently used in the testing phase.

Let's p(x) is a probability distribution over a d-dimensional vector. The objective is to identify a subset of features, S≡{1…k} of size |*S*|=*k*. Also, learning a reconstruction

function $f_r(.): \mathbb{R}^k \xrightarrow{\Delta} \mathbb{R}^d$, such that the loss between original sample $\boldsymbol{x}$ and reconstructed sample $f_r(x_S)$ is minimized as stated in Eq. 1,

$$argmin_{S,r}\, E_{p(x)}[\|f_r\,(x_S\,) - x\|_2] \ldots\ldots\ldots\ldots\ldots (1)$$

where $x_s \in \mathbb{R}^k$ consists of only selected features x_i s.t. $i \in S$. Note that samples are represented in a *2D* matrix, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, and the aim is to pick *k* columns of $\boldsymbol{X}$ such that sub-matrix $X_s \in \mathbb{R}^{n \times k}$. Later, selected feature set x_s can be used to reconstruct the original matrix $\boldsymbol{X}$ and classify the cancer types.

In the feature selection layer of CAE in Figure 5.2, the original features are selected based on this layer's temperature, which is tuned using an annealing schedule, as shown in Figure 5.3. More specifically, the concrete selector layer identifies *k* important features as the temperature decreases to zero, Figure 5.3(b). For reconstructing the input, a simple decoder similar to the ones associated with a standard AE is used. The temperature τ, of the random variable in the selector layer, has a significant impact on forming each node's output. Initially, when τ is high, search space is large since it considers a linear combination of all features, as shown in Figure 5.3(a). In contrast, the selector layer will not be able to search all possible combinations of features at low τ, and thus, the model converges to a poor local minimum. This means that as temperature goes down, a small number of features are necessary for stable convergence. Annealing or gradual decrease in temperature avoids the model convergence to a poor local minimum. The effect of annealing in feature selection is shown in Figure 5.3(a). For example, at the starting temperature, τ_s, the number of input features is 10, and the number of features to be selected is k = 3. At the next epoch, when the temperature is τ_{s+1}, the number of possible features reduces to 6. After some epochs, when the temperature reaches its lower bound τ_{stop}, the number of features further reduces

to 3, equal to *k*, the user-specified number of features to be selected. Instead of using a fixed temperature, a simple annealing scheduling scheme is used for feature selection. It starts with a user-defined high temperature (τ_s) and steadily lowers the temperature until it touches the end bound (τ_e), by every epoch as follows:

$$\tau_{(e)} = \tau_s \, (\tau_N / \tau_s)^{e/n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where, τ_e is the temperature at epoch *e*, *N* refers to the total number of epochs. Adam optimizer, with a learning rate of 0.001, was used for all the experiments for CAE. The starting temperature of CAE was set to 10, and it ends at 0.01.

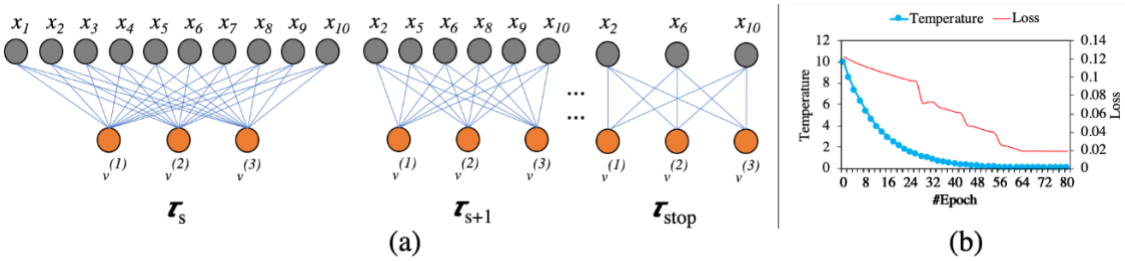

Figure 5.3: Effect of annealing in reducing search space. (a) An example: at starting temperature $\boldsymbol{\tau_s}$, the number of input features is 10 and the number of features to be selected is *k* = 3; at the next epoch when the temperature is $\boldsymbol{\tau_{s+1}}$, the number of possible features reduces to 6; after some epochs, when the temperature reaches to its lower bound $\boldsymbol{\tau_{stop}}$, the number of features further reduces to 3, which is equal to *k*. (b) Effect of temperature change in reducing the loss while training the concrete autoencoder on mRNA expression data to select the desired number of features, k. If the temperature is exponentially decayed (the annealing schedule), the feature selection layer converges to informative features with minimum loss.

5.2.3 Classification

To check the relevance of the selected features (glycome genes) to the origin of 33 different cancers, five classification algorithms, including Gaussian Naïve Bayes (GNB), K-nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) were used. The dataset was split into the train and test set according to a 75/25 ratio to avoid overfitting. The numbers of training and testing samples of 33 cancers are shown in Figure 5.1. The training set was used to estimate the learning parameters, and

the test set was used for performance evaluation. The mean accuracy of 10 different runs was reported in results where the dataset has been shuffled and split (75/25) for every run. Four different evaluation metrics have been used to record the classification performance, such as accuracy, precision, recall, and f1 score. Accuracy is the number of correct predictions made by the model over all kinds of predictions made. Precision is the number of correct positive results divided by the number of positive results predicted by the model. It indicates the predicted positive portion of the samples. The recall is the number of correct positive results divided by the number of all relevant samples. F1 score is the harmonic mean of precision and recall.

All performance metrics are measured on the predicted labels and true labels of independent test samples. The optimal number of features is selected based on two criteria: (a) the number of features should be as few as possible, and (b) the classification accuracy using the selected features should be > 90%.

5.2.4 Comparison

The feature selection capability of concrete autoencoder (CAE) was compared with the standard autoencoder (AE). Both AE and CAE are unsupervised approaches, but the former produces latent features, and the latter produces actual features. It is also known that AE performs better, maybe at the highest level, since it comes up with a reduced number of latent features with maximum variance. The objective of comparing CAE with AE is to check how close CAE's performance is to that of AE.

5.3 Results

5.3.1 Feature Selection and Classification Results

Finding Optimal k-value: The conditions for optimal feature set are (a) the number of features should be as few as possible, and (b) classification accuracy using the optimal feature set should be > 90%. As shown in Figure 5.4(a), a series of experiments were conducted to find the optimal number of features using CAE to classify 33 different cancers.

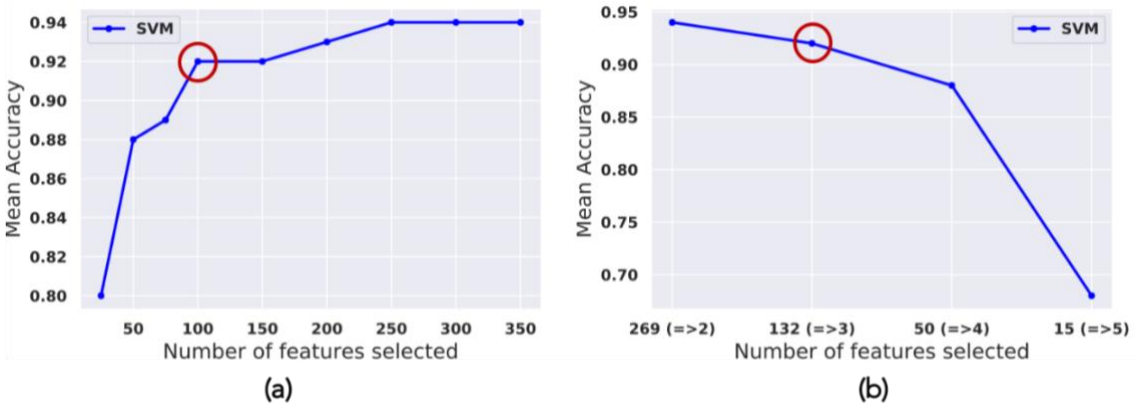

Figure 5.4: Optimal k-value and stable feature set. (a) Optimum k-value: Mean accuracy at a different number of features selected by CAE. The initial increase in the number of selected features from 25 to 100 showed a sharp increase in accuracy from 80% to 92%. Beyond this point, the increase in performance was not significant. From 100 to 200 features, accuracy increased only by 1%, which is not worthwhile. So, 100 features producing 92% accuracy meet the criteria of optimal k-value (number of features as few as possible and accuracy > 90%). (b) Stable feature set: Mean accuracy at a different number of features selected based on the frequency of a feature appearing in 10 runs with optimal k = 100. 132 genes appearing in ≥ 3 runs produced an accuracy of 92%. To increase the accuracy from 92% to 94% (only by 2%), one needs twice as many features (269 genes instead of 132 genes). 132 genes with 92% accuracy meet the optimal criteria ((number of features as few as possible and accuracy > 90%). Thus, the stable feature set consists of 132 genes.

It is clear from this figure that the initial increase in the number of selected features from 25 to 100 showed a sharp increase. Beyond this point, the increase in performance was not significant. For example, to increase the performance from 92% to 93%, one needs to increase the number of features from 100 to 200, which is not worthwhile. The optimal

classification performance for the present problem with CAE (accuracy > 90% with the smallest number of features) was observed with about 100 features. In other words, the optimal k-value for this problem is 100.

Finding a Stable Set of Features: With the same value of k = 100, the CAE produces a different optimal subset of 100 features in different runs. To get a stable set of features, the model was run 10 times with k = 100. Without loss of generality, it can be assumed that a gene that appears in more than one run can be considered an important feature. In 10 runs, it was observed that 269, 132, 50, and 15 genes appeared in ≥ 2, ≥ 3, ≥ 4, and ≥ 5 runs, respectively. The classification performance using these four subsets of features is shown in Figure 5.4((b). The feature sets 269 (≥ 2) and 132 (≥ 3) produced accuracy > 90%. It is noticeable that to increase the accuracy from 92% to 94%, one needs to increase the number of features from 132 to 269. In other words, to increase the accuracy by 2%, we need twice as many features, which is not worthwhile. So, the set of 132 genes that appeared in 3 or more runs were considered the stable feature set (the gene names are shown in *Appendix A.1*).

Comparing CAE with AE: To compare CAE performance with AE, 132 latent features were generated using AE. For completeness, the original feature set of 498 genes was also used for classification. Table 5.2 shows the performance of five classifiers – GNB, KNN, RF, SVM, and LR – in classifying 33 different cancers. Block A, Block B, and Block C of Table 5.2 show the performance of five classifiers using the original feature set (498 genes), reduced and stable feature set (132 genes), and 132 latent features produced by AE. It is clear from this table that SVM performed better with each set of features in terms of four evaluation matrices, including accuracy, precision, recall, and f1 score. It is noticeable

that the accuracy using the original feature set of 498 genes was 95%, which indicates that glycome genes carry the signature of cancers. But to conduct the wet lab experiment to identify the roles of each of these 498 genes is difficult and expensive. A reduced and stable set of features are desired to design a wet lab experiment. The stable set of 132 genes isolated in this study produces an accuracy of 92%, which satisfies the conditions for optimal feature set (number of features should be as few as possible and accuracy should be > 90%). The 132 latent features derived from AE show the upper bound of performance, 94%, for the present problem. The performance of CAE (92% accuracy) is pretty close to AE (94% accuracy), which provides confidence in explaining the role of glycome genes in cancer initiation and progression.

Table 5.2: Classification performance. Block A: Using original features of 498 glycome genes. Block B: Using 132 glycome genes selected by CAE. Block C: Using 132 latent features produced by AE.

#Features	Classifier	Mean Accuracy	Mean Precision	Mean Recall	Mean f1 Score
Block A 498	GNB	0.86 (+/- 0.01)	0.84 (+/- 0.01)	0.84 (+/- 0.01)	0.83 (+/- 0.01)
	KNN	0.91 (+/- 0.01)	0.88 (+/- 0.01)	0.88 (+/- 0.01)	0.87 (+/- 0.01)
	RF	0.91 (+/- 0.01)	0.89 (+/- 0.01)	0.85 (+/- 0.01)	0.85 (+/- 0.01)
	SVM	**0.95 (+/- 0.01)**	**0.93 (+/- 0.01)**	**0.92 (+/- 0.01)**	**0.92 (+/- 0.01)**
	LR	0.94 (+/- 0.01)	0.92 (+/- 0.01)	0.92 (+/- 0.01)	0.92 (+/- 0.01)
Block B CAE 132 (≥ 3)	GNB	0.84 (+/- 0.01)	0.80 (+/- 0.01)	0.83 (+/- 0.01)	0.80 (+/- 0.01)
	KNN	0.89 (+/- 0.01)	0.85 (+/- 0.01)	0.85 (+/- 0.01)	0.85 (+/- 0.01)
	RF	0.90 (+/- 0.01)	0.88 (+/- 0.02)	0.83 (+/- 0.01)	0.83 (+/- 0.01)
	SVM	**0.92 (+/- 0.01)**	**0.88 (+/- 0.01)**	**0.89 (+/- 0.01)**	**0.88 (+/- 0.01)**
	LR	0.92 (+/- 0.01)	0.89 (+/- 0.01)	0.88 (+/- 0.01)	0.88 (+/- 0.01)

Block C AE 132	GNB	0.83 (+/- 0.01)	0.82 (+/- 0.01)	0.85 (+/- 0.01)	0.83 (+/- 0.01)
	KNN	0.91 (+/- 0.01)	0.86 (+/- 0.01)	0.86 (+/- 0.01)	0.86 (+/- 0.01)
	RF	0.92 (+/- 0.01)	0.89 (+/- 0.01)	0.84 (+/- 0.01)	0.85 (+/- 0.01)
	SVM	**0.94 (+/- 0.01)**	**0.91 (+/- 0.01)**	**0.90 (+/- 0.01)**	**0.90 (+/- 0.01)**
	LR	0.91 (+/- 0.01)	0.89 (+/- 0.01)	0.84 (+/- 0.01)	0.85 (+/- 0.01)

5.3.2 Capability of Selected Features

Figure 5.5 shows the capability of selected 132 glycome genes in identifying the origin of 33 cancers with the t-SNE plot and confusion matrix. It is clear from the t-SNE plot that 132 glycome genes can distinguish 33 different types of cancer by forming distinct clusters. It is also clear from the confusion matrix that most cancers were identified with high accuracy except CHOL, ESCA, and READ. The number of CHOL samples was very low (36 only) compared to other cancers, which might play some role in poor performance. Though the number of samples (161 patients) for ESCA is not low, poor performance could be due to its complexity. The rectal adenocarcinoma (READ) was confused with colon adenocarcinoma (COAD). Similarly, some of the COAD samples were also confused with the READ samples. The reason is that both COAD and READ share many common features since the colon and rectum are two parts of one large organ.

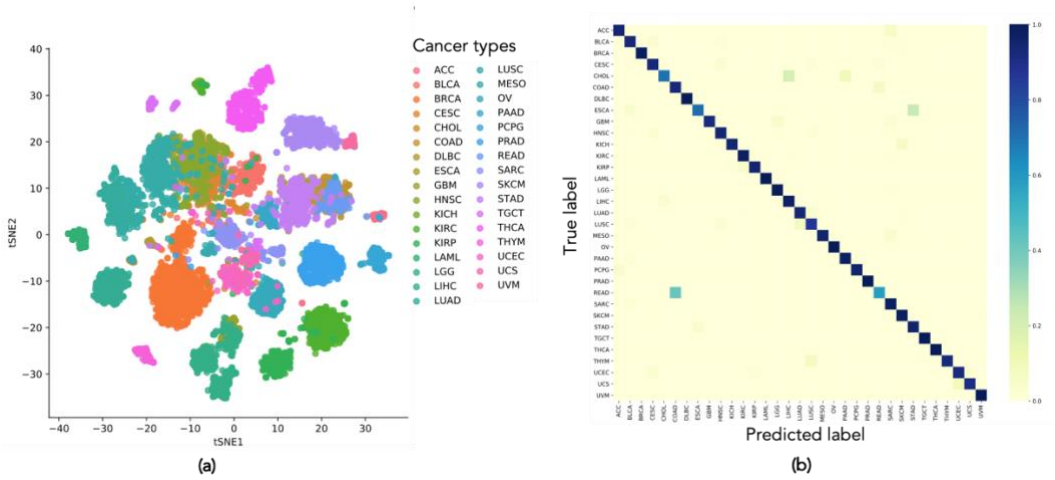

Figure 5.5: Capability of selected 132 glycome genes in identifying the origin of 33 cancers. (a) Confusion matrix generated using 132 glycome genes from SVM. (b) t-SNE using 132 glycome genes where each dot represents a cancer sample, and each color represents a cancer type.

5.3.3 Importance of Selected Features

Table 5.3 shows the distribution of glycome genes before and after feature selection by CAE. There were 498 and 132 genes before and after the selection process, respectively. The classification accuracies using 498 and 132 genes were 95% and 92%, respectively (last row of the table). The objective of this study was to find as few features (glycome genes) as possible with an accuracy > 90%, which helps design a wet lab experiment to investigate further the role of glycome genes in the process of cancer initiation and progression. It is clear from Table 5.3 that the number of genes in each category has been significantly reduced after the feature selection process. This means that glycosylation can be explained with fewer genes in each category. For example, to explain glycosylation in terms of Adhesion Molecule, one can use only two genes instead of seven genes. Similarly, to explain Glycan degradation, one can use 17 genes instead of 59 genes.

Table 5.3: Distribution of glycome genes before and after selection using CAE. Total genes: 498 (before) and 132 (after). Accuracy: 95% (before) and 92% (after). *Remarks:* Provide a smaller list of 132 glycome genes capable of identifying the origin of 33 cancers with an accuracy > 90%. This list of 132 genes could be used to design a wet lab experiment to investigate their role in tumorigenesis further.

Category	**Before**	**After**
Adhesion Molecule	7	2
CBP:C-Type Lectin	74	20
CBP:I-Type lectin	20	7
Galectin	12	3
Glycan Degradation	59	17
Glycosyltransferases	187	54
Glycoproteins	31	4
Intracellular protein transport	8	1
Miscellaneous	6	0
Nucleotide Sugar Transporters	57	15
Proteoglycans	29	7
Sulfotransferases	8	2
Total	**498**	**132**
Classification Accuracy	**95%**	**92%**

CHAPTER 6 CLASS-SPECIFIC FEATURE SELECTION AND CANCER SUBTYPE CLASSIFICATION

This chapter contains a class-specific feature selection framework to identify biomarkers associated with each molecular subtype of breast cancer. It also provides the necessary information on predicting breast cancer subtypes using different machine learning methods. In addition, it shows the prognostic evaluation of identified biomarkers. A detailed discussion on how novel biomarkers can be used for cancer diagnosis and prognosis is also provided.

6.1 Introduction

Breast cancer (BRCA) is heterogeneous with multiple subtypes, and treatment varies based on the subtype, even their prognostic consequences might be partially identical [101]. Using mRNA expression pattern with a hierarchical clustering method, Sorlie *et al.* [102] and Perou *et al.* [103] identified five subtypes as basal-like, HER2, Luminal A, Luminal B, and Normal-like. These subtypes are based on the presence or absence of estrogen, progesterone, and HER2 receptor (ER/PR/HER2). For example, if a patient's receptor statuses are ER-, PR-, and HER2+, the patient is diagnosed with HER2 subtype, while status with ER-, PR-, and HER2- represent basal or triple negative breast cancer (TNBC) (Table 6.1). These five subtypes have become the gold standard for breast cancer treatment. Identifying clinically relevant molecules of each subtype is essential for disease management and therapeutic decision-making. Existing methods for breast cancer subtyping are highly restricted to protein-coding genes and ignore the non-coding genes, which occupies close to 98% of the whole genome. The non-coding RNAs (ncRNAs) also play vital regulatory roles in breast cancer development [104], [105]. High-throughput

transcriptome data of breast cancer patients support the fact that in addition to mRNAs, ncRNAs also show a differential expression profile when tumor samples are compared with normal samples [106]. Non-coding RNAs can be categorized into two sets: (a) small ncRNAs (<200 nucleotides) and (b) long non-coding RNAs (lncRNAs) (>200 nucleotides). LncRNAs are classified according to the nearest protein-coding genes as intergenic, intronic, sense, or antisense [107]. LncRNA can be expressed either ubiquitous or tissue-specific. Later, they might be released in an unchanging form into the blood circulation during the disease progression [108]. The diverse regulatory roles for these lncRNAs as diagnostic and prognostic biomarkers for breast cancer have been implicated [24], [25], [92], [108]–[112]. For example, overexpression of *EPIC1* is associated with poor prognosis in Luminal B breast cancer patients [113]. This lncRNA is epigenetically activated in up to 90% of tumor samples across ten cancer types, including breast cancer. It has been shown that if some lncRNAs are recurrently targeted by DNA methylation alterations in tumors, they may play an important role in tumor initiation and progression [113].

Table 6.1: Molecular subtypes based on the presence or absence of estrogen, progesterone, and HER2 receptor (ER/PR/HER2) expression.

Subtypes	**Estrogen Receptor**		**Progesterone Receptor**		**HER2 Receptor**	**% Cancers**
Luminal A	+	and/or	+		-	30-40
Luminal B	+	and/or	+/-	or	+/-	20-30
HER2	-		-		+	12-20
Basal/Triple Negative	-		-		-	15-20
Normal-like	+	and/or	+		-	N/A

The “intrinsic” subtypes Basal, HER2, Luminal A, Luminal B, and Normal-like, have been extensively studied by hierarchical clustering of microarray data using different sets of “intrinsic” genes [21], [102], [103], [114]–[116]. Intrinsic genes are genes with significantly greater variation in expression between different tumors than between paired samples from the same tumor [103]. Starting with an expanded set of 1,906 “intrinsic” genes comprised of genes found in studies [102], [114]–[116], Parker *et al.* [20] developed a 50-gene supervised subtype predictor using Prediction Analysis of Microarray (PAM) [117], thus calling it PAM50. The limitation of PAM50 is that it has been developed ignoring normal-like subtype of breast tumor, which has distinct characteristics for ER, PR, and HER2 receptors. Recently, Zhang *et al.* identified a shortlist of genes and lncRNAs as the signatures for four types of breast tumors and did not consider normal-like subtype, a major limitation of this study [101]. The second limitation is that the hazard ratio values of the shortlist of genes in terms of survival analysis are close to 1, which raises a concern about the prognostic capability of those genes. To understand the biology concerning lncRNA expression, one needs to consider all five subtypes of breast tumors. In the present study, we considered five subtypes to have a comprehensive understanding of the role of lncRNAs in the clinical outcome of each subtype. We proposed a Recursive Feature Elimination (RFE) approach with L1-norm Multiclass Support Vector Machine (L1MSVM) as the estimator by taking advantage of both RFE [37] and L1MSVM [118], thus calling it Recursive L1-norm Multiclass Support Vector Machine (RL1MSVM).

6.2 Materials and Methods

The genome-wide lncRNA expression profiles of breast cancer patients are publicly available in The Cancer Genome Atlas (TCGA) and were used to discover the subtype-

specific key lncRNAs. To discover the subtype-specific key lncRNAs, we used the newly proposed RL1MSVM method along with two frequently used wrapper-based methods, L1MSVM and RF. First, an optimal number of features were isolated using the newly proposed RL1MSVM model. Then the same number of features were identified using the other two models, L1MSVM and RF. Finally, a feature isolated by at least two techniques is added to the final list of key lncRNAs. Figure 6.1 represents the flowchart of data preparation, feature selection using three algorithms (L1MSVM, RF, and RL1MSVM), classification of five subtypes using the selected sets of features, and corresponding performance evaluation. The details of the materials and methods are described in the following subsections.

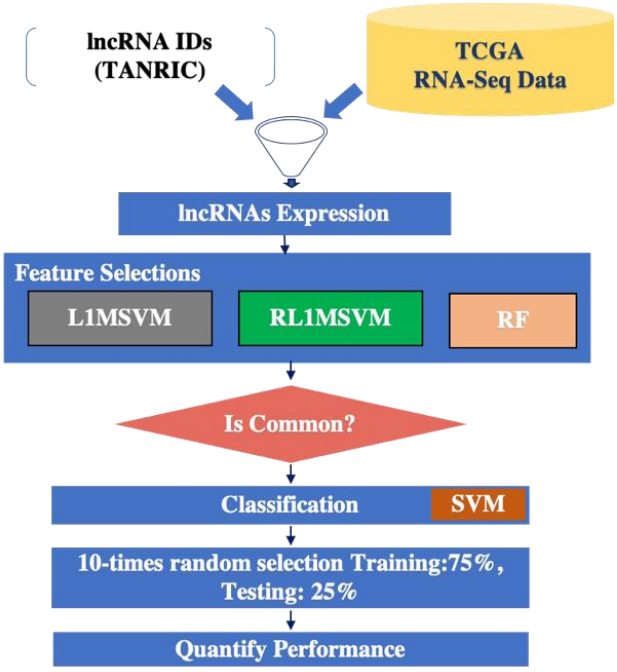

Figure 6.1: Process flow diagram: data preparation, feature selection, classification, and performance evaluation.

6.2.1 Acquisition of Breast Cancer Data

The breast cancer RNAseq FPKM normalized expression profiles and clinical data were downloaded (July 2019) from the UCSC Xena database [77] for analysis and validating

the idea. Out of 1,218 available patients, 1,207 were labeled with molecular subtypes, and the remaining 11 without subtypes information were excluded from the analysis. Each sample consists of 60,483 RNA (coding and non-coding combined) expression values. The row and column headings of the dataset represent the RNA Ensemble IDs and sample IDs, respectively. The value of each cell represents the normalized read counts of an RNA for a specific sample. Each RNA expression was further processed using a min-max normalization method to achieve proper training performance.

Since this study focuses on identifying key lncRNAs for breast cancer, expression values of lncRNAs are isolated from the combined dataset using lncRNA IDs available in The Atlas of non-coding RNA in Cancer (TANRIC) [78]. This mapping resulted in 12,309 common lncRNAs with expression values of 1105 breast cancer patients with five different subtypes, which were used for analysis in the present study. Table 6.2 summarizes the processed data: the number of samples ranges from 40 for Normal-like to 577 for Luminal A, and an average survival ranges from 1090 days for HER2 to 1410 days for Normal-like patients.

Table 6.2: Number of samples and average survival of breast cancer patients in each subtype.

Subtype	No. of Samples	Average Survival, Days
Luminal A	577	1,293
Luminal B	216	1,112
HER2	81	1,090
Basal	191	1,311
Normal-like	40	1,410

6.2.2 The Proposed Recursive l1-norm Multiclass SVM

Not all feature selection techniques can select class-specific features, such as (a) Analysis of Variance (ANOVA), a filter-based method, (b) Least Absolute Shrinkage and Selection Operator (LASSO) [36], a wrapper-based method, and (c) Recursive Feature Elimination (RFE) [37], an embedded method. However, the embedded feature selection approach, L1MSVM, can discover class-specific essential features [118]. Although RFE cannot identify class-specific features but has a role in reducing the number of irrelevant and noisy features in each iteration and selecting the top-ranking features to improve the classification performance [118]. The L1MSVM generates a sparse weight matrix with many zeros, which provides the first level of reduction from a large number of features. Applying RFE to it further reduces the number of features with smaller weights. Thus, the proposed RL1MSVM approach, a combination of L1MSVM and RFE, is appropriate for the present problem of discovering subtype-specific lncRNAs for breast cancer due to the high dimensionality (12,309 lncRNAs) of the data compared to the number of samples (1,105 samples). The formulation of RL1MSVM starts from a standard linear SVM classifier with l2-norm. The general setup of a supervised classification problem consists of sample: $i \in \{1 \cdots n\}$, feature: $j \in \{1 \cdots d\}$, and class: $k \in \{1 \cdots c\}$. The training set is represented by $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{1 \le i \le n}$, where $\boldsymbol{x}_i = (x_{i1} \cdots x_{id})^T$ represents the *i*th sample over a d-dimensional feature vector and $\boldsymbol{y}_i = (y_{i1} \cdots y_{ic})^T$ represents its label vector. $y_{ik} = 1$, if the sample belongs to *the k*th class, otherwise $y_{ik} = -1$. For any class k, a linear classifier uses a *d*-dimensional weight vector, $\boldsymbol{w}_k = (w_{k1} \cdots w_{kd})^T$. To find appropriate $\boldsymbol{w}_k$, SVM minimizes the following objective function.

$$\frac{1}{2}\sum_{j=1}^{d} {w_{kj}}^2 + \frac{C}{2}\sum_{i=1}^{n} J\left(y_{ik}(\boldsymbol{w}_k.\boldsymbol{x}_i)\right) \qquad (1)$$

Where $l2\text{-norm} = \sqrt{\sum_{j=1}^{d} {w_{kj}}^2}$, and J refers to the loss function. The regularization parameter C controls the trade-off between achieving a low error on the training data and minimizing the norm of the weights. It has been shown that replacing *l2*-norm with *l1*-norm ($\sum_j |w_{kj}|$) performs better for multiclass feature selection problems [119]. The final form of L1MSVM that solves the simultaneous feature selection for the multiclass problem is as follows.

$$\sum_{k=1}^{c}\sum_{j=1}^{d} |w_{kj}| + \frac{C}{2}\sum_{\mathrm{i}=1}^{\mathrm{n}} J\left(y_{ik}(w_k.x_i)\right) \qquad (2)$$

However, the above objective function is not differentiable, and we cannot employ linear programming techniques for optimization [120]. One can do something similar using the RFE technique, which tells that the significance of selecting a feature *j* for the class *k* should be related to the magnitude of its weight w_{kj}. In other words, feature *j* should be eliminated if its weight is the smallest; thus, it selects top-ranking features for class *k*. A feature with the lowest weight is determined using the equation given below.

$$j = \ argmin \sum_{k} {w_{kj}}^2 \qquad (3)$$

Combining the idea of *l1*-norm SVM for multiclass and backward elimination technique RFE provides the complete solution of class-specific feature selection for a multiclass problem.

6.2.3 Implementation of Feature Selection Approaches

In the present study, we used two state-of-the-art feature selection approaches, L1MSVM and RF, to compare the performance with the newly proposed simultaneous feature selection and classification approach RL1MSVM. To select the optimized features, tuning of hyperparameters is a must. The *GridSearch* cross-validation technique was used to find the optimum values or options for hyperparameters for all feature selection approaches.

Implementation of LIMSVM

Three kernels used in L1MSVM are linear, polynomial, and radial basis function (RBF). The hyperparameters are different for each kernel: (i) *C* for the linear kernel, (ii) *C* and *degree* for the polynomial kernel, and (iii) *C*, *degree*, and *gamma* for RBF kernel. The regularization parameter *C* controls the trade-off between the number of selected features and the model's accuracy. For example, if *C* is small, the model will select a low number of features for classification, resulting in poor prediction accuracy. The *degree* of the polynomial is used to find the hyperplane to split the data. Using degree equal to 1 is the same as using a 'linear' kernel. Also, increasing the degree leads to higher training times. The *gamma* is a parameter for non-linear hyperplanes. The higher value of *gamma* leads to overfitting as the classifier tries to perfectly fit the training data. The *GridSearch* found a linear kernel with *C* equal to 0.07 as the global optimum hyperparameters. The details of identifying the optimal value of *C,* are provided in section 6.3.1 Selecting Optimum Features.

Implementation of RF

RF needs three hyperparameters to be tuned: (i) *n_estimator:* number of estimators, or number of trees in the forest, (ii) *min_sample_split:* minimum number of nodes required

to split, and (iii) *criterion:* impurity to measure the quality of a split. In *GridSearch*, the ranges of values assigned to tune *n_estimator* and *min_sample_split* were from 2 to 300 and 1 to 150, respectively. Two options, *Gini* and *entropy*, were used to optimize the impurity parameter *criterion*. The optimum values or options for *n_estimator*, *min_sample_split, and criterion* found by *the GridSearch* method are 100, 120, and *Gini*, respectively.

Implementation of RL1MSVM

The newly proposed feature selection approach, RL1MSVM, is an RFE feature selection approach, where L1MSVM is the estimator. Three steps to implement RL1MSVM are: *Step-1:* Train L1MSVM on the active features. *Step-2:* Remove the feature with the smallest weight (RFE). *Step-3:* Go back to step 1 or stop if the algorithm finds the desired number of features. In every iteration of RFE, the number of dropped features was set to 100.

A python machine learning package, Scikit-learn [81], was used for model deployment. All models are executed on a CPU Intel Core i7 with 16GB RAM.

6.2.4 Classification and Performance Evaluation

To compare the performance in classifying the five subtypes of breast cancer using the sets of features selected by L1MSVM, RF, and RL1MSVM, four different performance metrics, including accuracy, precision, recall, and f1 score, were evaluated. For a fair comparison, the same number of features were selected using all three approaches. In our previous study, it was shown that Support Vector Machine (SVM) outperformed in classifying multiple cancer types using lncRNA expression [121]. Therefore, the same classifier, a linear SVM without regularization was employed to classify the breast cancer

samples into five subtypes. The dataset was split into training and test set according to a 75/25 ratio to avoid overfitting (Figure 6.1). The training set was used to estimate the learning parameters, and the independent test set was used for performance evaluation.

6.2.5 Final Set of Key lncRNAs

The final set of key lncRNAs was derived from the union of the intersections of two sets of features selected by two approaches. More specifically, lncRNAs found in at least two methods were considered key lncRNA, as stated in Eq. 4.

$$Key\ lncRNAs = (L1MSVM \cap RF) \cup (RF \cap RL1MSVM) \cup (RL1MSVM \cap L1MSVM) \quad (4)$$

6.3 Results

Here, we report the performance of the proposed simultaneous feature selection and classification approach, RL1MSVM, compared to the similar state-of-the-art methods L1MSVM and RF. For a fair comparison, we need to select the same number of lncRNAs using three approaches. Since the L1MSVM is the estimator in RFE approach RL1MSVM, the optimum number of features selected by L1MSVM would provide the basis for comparison and subsequent analysis.

6.3.1 Selecting Optimum Features

Before selecting optimum features by L1MSVM model, we need to find the optimal parameters for it. Figure 6.2 shows the results of *GridSearch* cross-validation technique in finding the optimal parameters for L1MSVM. In *GridSearch*, the ranges of values assigned to tune the parameters *C*, *degree*, and *gamma* were from 0.0001 to 1000, 1 to 5, and 0.001 to 100, respectively. The *GridSearch* found a linear kernel with *C* equal to 0.1 as the local

optimum hyperparameter, as shown in Figure 6.2(a). Then, we ran the same experiment with C ranging between 0.01 and 0.1 and found C = 0.07 as optimal (Figure 6.2b). We selected the optimal value of C based on the mean test score of 5-fold cross-validation.

After training with the optimized parameters, the L1MSVM model identified 239 important lncRNAs (*Appendix A Table S1*) as the feature set for classification and prediction. Of 239 lncRNAs, the number of subtype-specific lncRNAs for Basal, HER2, Luminal A, Luminal B, and Normal-like were 27, 43, 82, 72, and 15, respectively. The combined number of unique lncRNAs was 196, as some lncRNAs contributed to multiple subtypes. Thus, the optimum number of features selected by L1MSVM was 196, as shown in Figure 6.2(b). The same number of features (196 lncRNAs) were selected using RL1MSVM and RF for comparison and subsequent analysis.

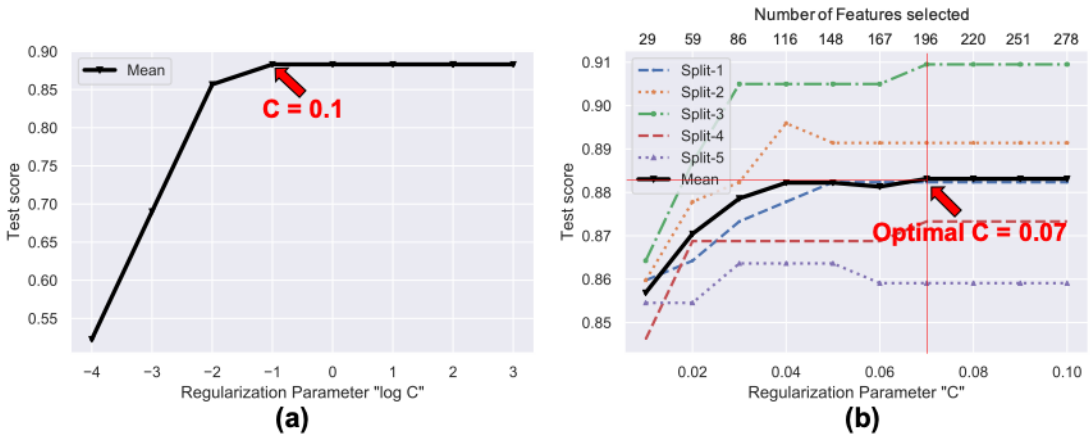

Figure 6.2: 5-fold cross-validation test score of L1MSVM at different values of regularization parameter C. (a) *C* ranges from 0.0001 to 1000 and found *C* = 0.1 as local optimum. (b) *C* ranges from 0.01 to 0.1 and found *C* = 0.07 as global optimum where the number of features is 196. The optimal C is based on the mean test score of 5-fold cross-validation.

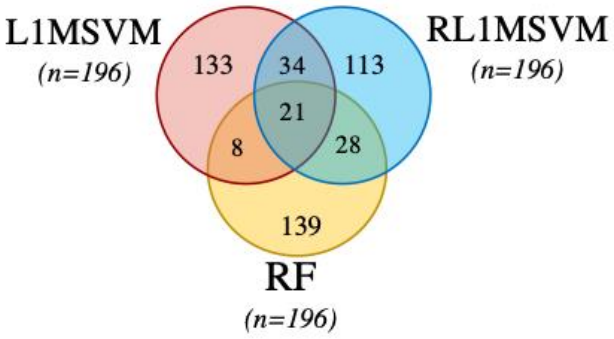

Figure 6.3: Venn diagram: Number of common features among three methods.

The Venn diagram, Figure 6.3, shows the number of common features among different methods. It is clear from the Venn diagram that there is a considerable variation among the features selected by three different approaches, even the numbers of selected features are the same. Of 196 lncRNAs selected by each model, only 21 lncRNAs were common between three models, 55 (34 + 21) between L1MSVM and RL1MSVM, 29 (21 + 8) between L1MSVM and RF, and 49 (21 + 28) between RL1MSVM and RF.

6.3.2 Discovering Final Set of LncRNAs

Due to heterogeneous nature of breast cancers, classifying them into five molecular subtypes using the genome-wide expression profiles of lncRNAs is complex, and thus, different feature selection algorithms select different sets 196 lncRNAs. We argued that a feature discovered by two models would be regarded as key feature. Using this criterion, 91 key lncRNAs were found and distributed over the subtypes as 20, 14, 27, 20, and 13 for Basal, HER2, Luminal A, Luminal B, and Normal-like, respectively (*Appendix A Table S2)*. A few lncRNAs appeared in more than one subtype (highlighted in blue color), such as *HOTAIR* and *STK4-AS1* belong to Luminal A and Luminal B, and *MEG3* belong to Luminal A and Basal subtypes.

6.3.3 Performance of Selected Features in Classifying Breast Cancer Subtypes

Table 6.3 shows the performance of three sets of 196 lncRNAs selected by L1MSVM, RF, and RL1MSVM, in classifying five subtypes of breast cancer. Support vector machine (SVM) was used for subtype classification. It is clear that the features selected by RL1MSVM performed better than that of L1MSVM and RF, in four evaluation metrics, including accuracy, precision, recall, and f1 score, which is also supported by the confusion matrix derived from classification (*Appendix A Figure S1*). For example, accuracies in classifying breast cancer subtypes are 0.90, 0.84, and 0.92 using three sets of 196 lncRNAs selected by L1MSVM, RF, and RL1MSVM, respectively. We also checked the classification performance using the 91 key lncRNAs. It is clear that the 91 key lncRNAs (accuracy 0.83) are as good as 196 lncRNAs (accuracy 0.84) discovered by RF in classifying subtypes, which is significant.

Table 6.3: Comparison of feature selection performance of L1MSVM, RF, and RL1MSVM. Three sets of 196 lncRNAs were selected by three approaches. SVM was used to classify the breast cancer samples into five subtypes using the selected features. Values of four performance metrics, including Accuracy, Precision, Recall, and f1 Score, are evaluated. The last row shows the classification performance using the 91 key lncRNAs.

Model Name	# of features	Accuracy	Precision	Recall	f1 Score
L1MSVM	196	0.90 (+/- 0.02)	0.83 (+/- 0.04)	0.82 (+/- 0.04)	0.82 (+/- 0.03)
RF	196	0.84 (+/- 0.02)	0.78 (+/- 0.05)	0.76 (+/- 0.03)	0.76 (+/- 0.03)
RL1MSVM	**196**	**0.92 (+/- 0.02)**	**0.87 (+/- 0.03)**	**0.85 (+/- 0.03)**	**0.85 (+/- 0.02)**
Key lncRNA	***91***	***0.83 (+/- 0.03)***	***0.78 (+/- 0.03)***	***0.75 (+/- 0.03)***	***0.75 (+/- 0.04)***

6.3.4 Clusters of Subtype-Specific Patients

To check the clustering capability using expression profiles of 196 lncRNAs discovered by three feature selection techniques along with the 91 key lncRNAs, we applied hierarchical

clustering and unsupervised visualization technique t-SNE [26]. It is clear from the heatmap of hierarchical clustering, Figure 6.4, that the 91 key lncRNAs performed better than the three sets of 196 lncRNAs, specially in clustering normal-like samples. The t-SNE plots, Figure 6.5, shows that the 91 key lncRNAs perform at the same level as 196 lncRNAs. Thus, the newly identified 91 key lncRNAs can be considered as possible features for diagnosis, prognosis, and therapeutic target for breast cancer.

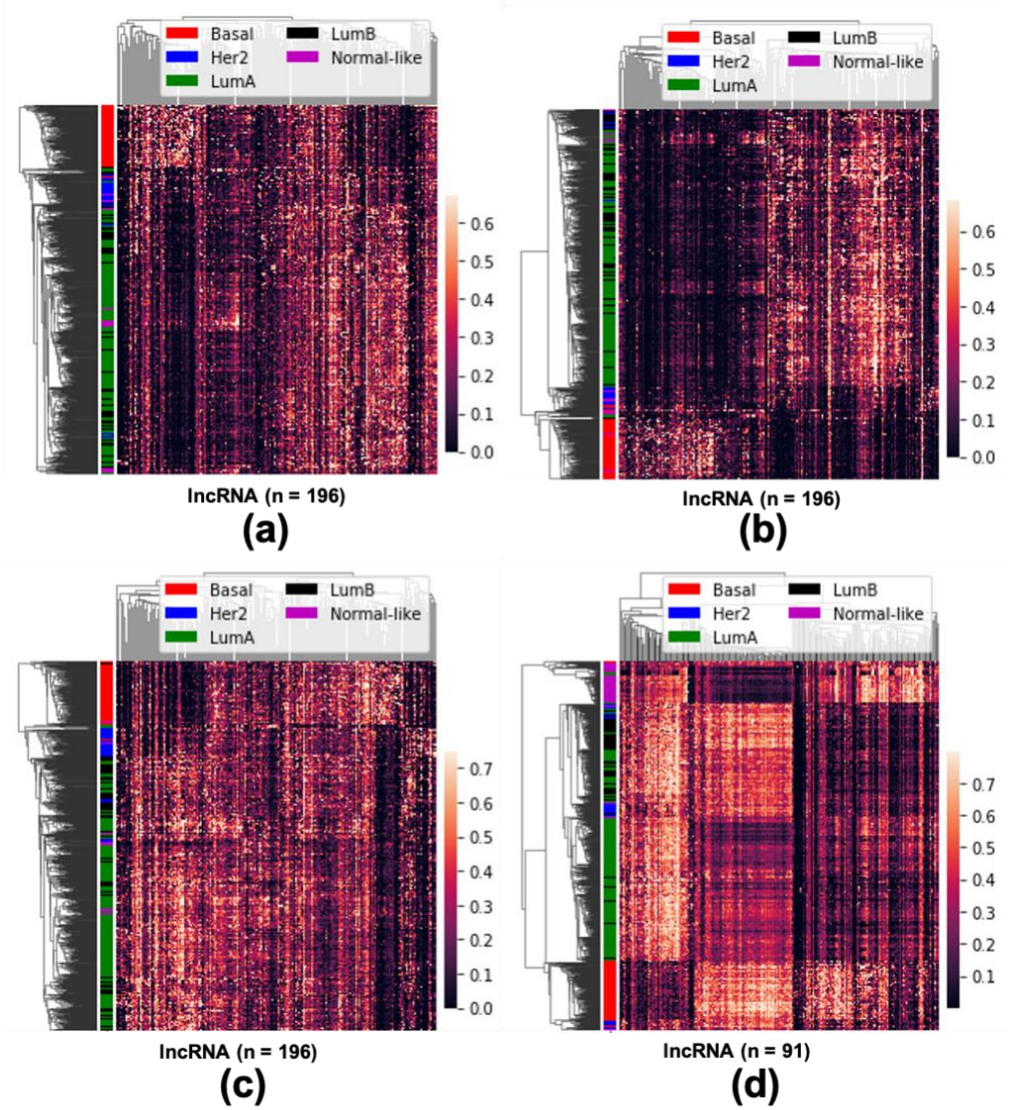

Figure 6.4: Heatmap of breast cancer subtypes clustering using the expression profiles of 196 lncRNAs discovered by (a) L1MSVM, (b) RF, and (c) RL1MSVM, and d) 91 key lncRNAs.

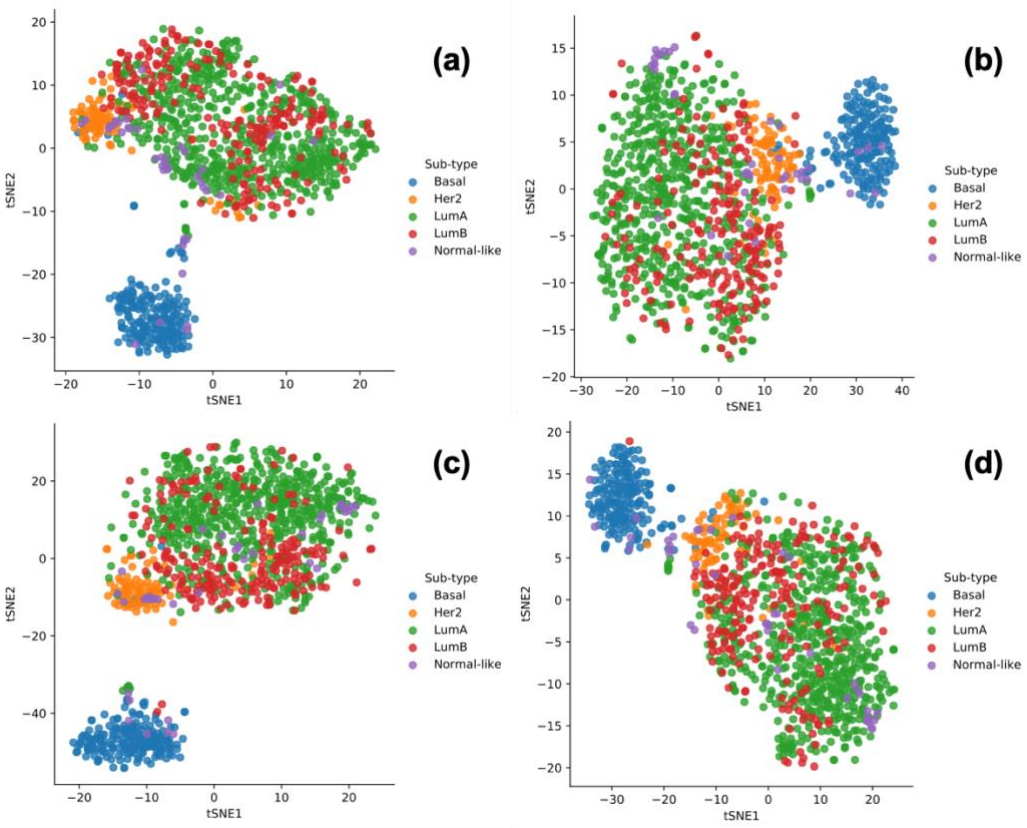

Figure 6.5: t-SNE plots to cluster breast cancer subtypes using the expression profiles of 196 lncRNAs discovered by (a) L1MSVM (b) RF (c) RL1MSVM, and d) 91 key lncRNAs.

6.3.5 Literature Validation of Discovered lncRNAs

Table 6.4 summarizes the literature validation of the final set of 91 lncRNAs. Fifty-three of these lncRNAs are known for any kind of disease (*Appendix A Table S3*) and 38 are novel discoveries. Of 53 lncRNAs, 25 are known for breast cancer and other diseases (*Appendix A Table S4*) and 6 are known for breast cancer only (*Appendix A Table S5*). It should be noted that 6 of the novel lncRNAs (*Appendix A Table S6*) are discovered by all three approaches, which could be further studied for potential therapeutic targets.

Table 6.4: Summary of literature validation of discovered lncRNAs.

lncRNA Type	# lncRNAs	Tables
Discovered key lncRNAs	91	Table S2
Known for any kind of diseases	53	Table S3
Known for breast cancer and other diseases	25	Table S4
Known for breast cancer only	6	Table S5
Novel	38	Table 6.5
Common lncRNAs between novel (*n*=38) and overlapping in all three methods (*n*=21) (Figure 6.3. Venn diagram)	6	Table S6

6.3.6 Prognostic Evaluation of Novel lncRNAs

To evaluate the prognostic capabilities of novel lncRNAs, survival analyses were performed on the whole cohort as well as on subtype-specific cohort. The patients with values less than or equal to the median were labeled group A (low-expression group) and greater than the median group B (high-expression group). After dividing into two groups, a log-rank test was conducted. Then the hazard ratio was calculated as the hazard rate of group A divided by the hazard rate of group B to check the prognostic capability of an individual lncRNA. The criteria for a lncRNA to be prognostic are log-rank test P-value $\leq$ 0.05 and Hazard Ratio (HR) $\neq$ 1.0. Table 6.5 shows the list of 38 novel lncRNAs with associated subtype, genomic coordinates, and prognostic significance. Of 38 novel lncRNAs, 23 are found prognostically significant.

Table 6.5: List of 38 novel lncRNAs and corresponding breast cancer subtype along with their genomic coordinate. LncRNAs highlighted in blue color belong to more than one subtype or pleiotropic. Of 38 Novel lncRNAs, 23 lncRNAs were found prognostically significant.

LncRNA	Sub-type	Chrom	Start	End	Prognostically Significant?
AC005152.3	Basal	chr17	72021851	72034092	Yes
AC087491.2	Basal	chr17	39619613	39622513	Yes
CTD-3032H12.1	Basal	chr16	54937786	54938671	
LINC00152	Basal	chr2	87455368	87606805	Yes
RP11-279F6.1	Basal	chr15	69463026	69571440	
RP11-281O15.4	Basal	chr5	178969390	178990116	Yes
TTC39A-AS1	Basal	chr1	51329654	51335324	Yes
CTB-33O18.1	HER2	chr5	173562478	173573199	Yes
CTD-2284J15.1	HER2	chr8	86333274	86343314	Yes
ELOVL2-AS1	HER2	chr6	11043524	11078226	Yes
KIRREL3-AS1	HER2	chr11	126543947	126610948	
LINC00839	HER2	chr10	42475543	42495336	
RP1-232P20.1	HER2	chr6	5451683	5458075	Yes
RP11-20F24.2	HER2	chr10	37240887	37242049	
RP11-28F1.2	HER2	chr18	63313802	63314376	Yes
STK4-AS1	LumA, B	chr20	44963794	44966402	Yes
CTD-2015G9.2	LumA	chr16	86722091	86741059	Yes
CTD-2081C10.7	LumA	chr5	53880293	53881051	
LINC00324	LumA	chr17	8220642	8224043	
LINC00922	LumA	chr16	65284499	65576300	
LINC01272	LumA	chr20	50267486	50279795	Yes
PARD3-AS1	LumA	chr10	34815767	34816386	
PRKAG2-AS1	LumA	chr7	151877042	151879223	Yes
RERG-IT1	LumA	chr12	15112363	15114698	
RP11-21L23.2	LumA	chr11	76800364	76804555	Yes
SEMA3B-AS1	LumA	chr3	50266641	50267371	Yes
AC016735.2	LumB	chr2	43027853	43039547	Yes
AP000439.3	LumB	chr11	69477133	69479940	
DOCK9-AS2	LumB	chr13	99087819	99088625	Yes
LINC00992	LumB	chr5	117415509	117546298	Yes
RARA-AS1	LumB	chr17	40340867	40343136	
SYN2	LumB	chr3	12004402	12191400	Yes
TPTEP1	LumB	chr22	16601887	16698742	
CTB-51J22.1	Normal-like	chr7	74059576	74062284	Yes

DYNLL1-AS1	Normal-like	chr12	120490328	120495940	
LINC00087	Normal-like	chrX	135095028	135098634	
LINC00504	Normal-like	chr4	14470465	14888169	Yes
MIR205HG	Normal-like	chr1	209428820	209432838	Yes

6.4 Discussion

We hypothesize that there should be a shortlist of salient features or important lncRNAs with prognostic capability that could differentiate cancer subtypes. Our investigation showed that the lncRNAs discovered in this study carry significant information on having the prognostic capability of differentiating high- and low-risk groups of patients of a particular breast cancer subtype, as explained in the section 6.3.6. We also discussed the biological relevance of the selected lncRNAs comparing with the existing literature in section 6.3.5.

6.4.1 Novel lncRNAs Associated with Breast Cancer Subtypes

This research identified 91 key lncRNA associated with breast cancer (*Appendix A Table S2*). When compared against the existing literature, 53 (58%) have been reported as known important prognostic biomarkers for various diseases, including breast cancer. The remaining 38 lncRNAs are novel discovery. Of 53 known lncRNAs, 25 are related to breast cancer and other diseases, and six have been previously associated only with breast cancer, including, *AC008268.1, FGF14-AS2, LINC00993, LINC01016, PTPRG-AS1, ST8SIA6-AS1* (*Appendix A Table S5*). Classification and clustering capabilities of 91 key lncRNAs are shown in Table 6.3 and Figure 6.5.

Since the proposed method can identify already known lncRNA biomarkers, we can conclude that the newly discovered 38 lncRNAs (Table 6.5) have the potential to be considered as novel biomarkers associated with Basal (7), HER2 (8), Luminal A + Luminal

B (18), and Normal-like (5) breast cancer subtypes. However, some well-known critical lncRNAs are still missing in our results, such as *ANRIL, ATAB, NEAT1*, and *TP53TG1* [106]. *Reason 1:* the list of lncRNAs used for this analysis was obtained from the TANRIC repository annotated by GENCODE v2.0, which is not a complete list; for example, *ANRIL* and *ATAB* are missing in the list. *Reason 2:* the model produced the optimum accuracy with 196 lncRNAs, and it remained at the optimum level with the increase of the number of features up to 400 (Figure 6.2). Thus, selecting the minimum number of optimal features (196 lncRNAs) for analysis might miss some key lncRNAs.

6.4.2 LncRNAs as Screening Tool and Therapeutic Target

Many studies have been conducted using mRNA expression for predicting breast cancer molecular subtypes as well as developing screening tools such as PAM50 [20] and 70-genes [21]. However, none of these tools used lncRNAs for predicting breast cancer molecular subtypes. As per our knowledge, this study is the first attempt to discover subtype-specific lncRNAs while predicting breast cancer molecular subtypes. Intra-tumor heterogeneity is likely to have implications for cancer therapeutics [122]. Hence, the identified 91 lncRNAs, which are subtype-specific, can be used not only as a screening tool for breast cancer diagnosis but also as therapeutic targets.

CHAPTER 7 MULTI-RUN CONCRETE AUTOENCODER FOR FEATURE SELECTION

This chapter provides the limitations of concrete autoencoder in feature selection from high dimensional transcriptomic data. It also discussed how the multi-run approach could overcome its limitations. This chapter contains the necessary information on data preprocessing, model development, model training, hyper-parameter tuning, and performance evaluation. It also contains the biological validations of identified biomarkers.

7.1 Introduction

Recent studies showed that long non-coding RNAs (lncRNAs), which are longer than 200 nucleotides, play key roles in tumorigenesis [49]–[51]. The lncRNAs also have key functions in transcriptional, post-transcriptional, and epigenetic gene regulation [9]. Schmitt and Chang discussed the impact of lncRNA in cancer pathways [10]. Hanahan and Weinberg described the involvement of lncRNAs in six hallmarks of cancer such as proliferation, growth suppression, motility, immortality, angiogenesis, and viability [11]. Hoadley *et al.* showed that cell of origin patterns dominate the molecular classification of tumors available in The Cancer Genome Atlas (TCGA) [52]. Their analysis used copy number, mutation, DNA methylation, RPPA protein, mRNA, and miRNA expression. However, they did not consider another important molecular signature of cancer, lncRNA expression. This work motivated us to investigate the importance of lncRNAs in identifying different types of cancer. We hypothesize that there should be a shortlist of salient features or important lncRNAs with prognostic capability that could dictate the origin of multiple cancers.

In general, feature selection is worthwhile when the whole set of features is difficult to collect or expensive to generate [23]. For example, in TCGA, the lncRNA expression profile dataset contains more than 12,000 features (lncRNAs) for each of 33 different cancers, and it is expensive to generate this data. Consequently, it is important to answer the question: *Is there a set of salient features (lncRNAs) capable of identifying the origin of 33 cancers or a subset of 33 cancers?*

Standard dimension reduction methods, such as principal component analysis (PCA) [58] and autoencoders [47], can generate a greatly reduced set of ***latent features***. However, these latent features are not ***the original features*** but are functional combinations of the original features. Identifying original features increases the "explainability" of the results and allows us to perform biological interpretation in diagnosing various deadly diseases, such as cancers. Recently, few deep learning-based feature selection methods showed improvement in selecting original features in both supervised and unsupervised settings [23], [59]–[61].

In our previous study [24], we showed that a deep learning-based unsupervised feature selection algorithm CAE [23], performs better in feature selection, especially, in selecting a small number of features, compared to the state-of-the-art supervised feature selection methods such as LASSO, RF, and SVM-RFE. However, the study was based on the expression profiles of cancer patients only. The questions that remained unanswered are: (a) Were the identified lncRNAs cancer-specific or organ-specific? (b) CAE produces different sets of features in different runs, which raises questions about which set to use as the final feature set. (c) Do the identified lncRNAs have prognostic capability? (d) How to validate the identified lncRNAs?

In this research, to address question (a), we analyzed data from 12 cancers having a healthy to cancer sample ratio of at least 1:10. To address question (b), we proposed running CAE multiple times with a fixed number of features selected in each run and taking the most frequently appearing features in multiple runs as the final set of features. To address question (c), survival analysis was performed to show that the identified features have the prognostic capability. To address question (d), we checked the existence of identified lncRNAs in experimental works of literature, drug-lncRNA network, and cancer hallmarks.

The distribution of the number of samples for 12 cancers in TCGA is highly imbalanced, ranging from 36 for CHOL cancer to 1089 for BRCA cancer. Any supervised feature selection approach will be biased to heavy groups. So, the unsupervised nature of CAE can handle this issue in identifying appropriate features related to 12 different cancers. The proposed multi-run CAE approach filtered the key lncRNAs from 12,309 lncRNAs that are related to 12 different cancers with higher classification accuracy and better diagnosis of cancer origin compared to the state-of-the-art embedded feature selection approaches, including LASSO [123], RF [45], and SVM-RFE [37], and unsupervised feature selection approaches such as MCFS and UDFS.

Contributions of this study are as follows: 1) development of an optimal and stable feature selection framework, mrCAE. 2) Discovery of an optimal and stable set of 128 lncRNAs capable of identifying the origin of organs for 12 different cancers with an accuracy of 95%. 3) It has been shown that the lncRNAs identified using mrCAE from the expression profiles of cancer patients are truly cancer-specific, not organ-specific. 4) Survival or

prognostic analysis of discovered lncRNAs. 5) Identified features, lncRNAs, are validated with existing literature, drug-lncRNA networks, and hallmark lncRNAs.

7.2 Materials and Methods

7.2.1 Data Preparation

To characterize the cancer-associated lncRNA, expression profiles and clinical data for 33 different cancers were downloaded from the UCSC Xena database [77]. Each lncRNA expression was processed using a min-max normalization method to achieve good training performance. For this study, we considered the cancer types for which the number of normal samples is at least 10% of cancer samples, and 12 cancer types met this criterion. The distributions of cancer and normal samples for 12 cancers are shown in Table 7.1.

Table 7.1: Sample distributions of 12 cancers were considered in this experiment.

	BRCA	CHOL	COAD	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA
Normal	113	9	41	23	72	32	50	57	49	52	9	58
Cancer	1088	36	301	65	527	286	369	510	498	493	94	501

This dataset contains about 60 thousand RNAs expression profiles, including coding genes (mRNAs) and non-coding genes (lncRNAs and miRNAs). In this study, only the expression profiles of lncRNA (n=12,309) were considered for analysis and model evaluation. The final dataset contains 4,768 cancer patients and 565 normal patients.

7.2.2 Features Selection Using Multi-Run Concrete Autoencoder

For selecting important features (lncRNAs), a state-of-the-art deep learning-based unsupervised algorithm, Concrete Autoencoder (CAE) [23], was run multiple times iteratively. We name this approach multi-run CAE (mrCAE). The reason for using mrCAE is that CAE selects the most informative features in a *stochastic manner*, meaning different

sets of informative features are selected in different runs. We made the assumption while running CAE multiple times that if a feature appears in more than one run, that can be considered a stable feature.

Architecture and Working Principle of CAE

The architecture of a CAE (Figure 7.1) consists of a single encoding layer, also known as the "Feature selection layer" shown in yellow, and arbitrary decoding layers (e.g., a deep feedforward neural network), shown in the box on the right. The detailed algorithm is available in [23]. The function of the encoder is to select a given number of k actual features (not latent features in the case of a traditional Autoencoder) in a stochastic manner from the original large input feature space, **X** of size n. The function of the decoder is to reconstruct the original features (**X′** is the reconstructed feature vector) using the k features selected by the encoder.

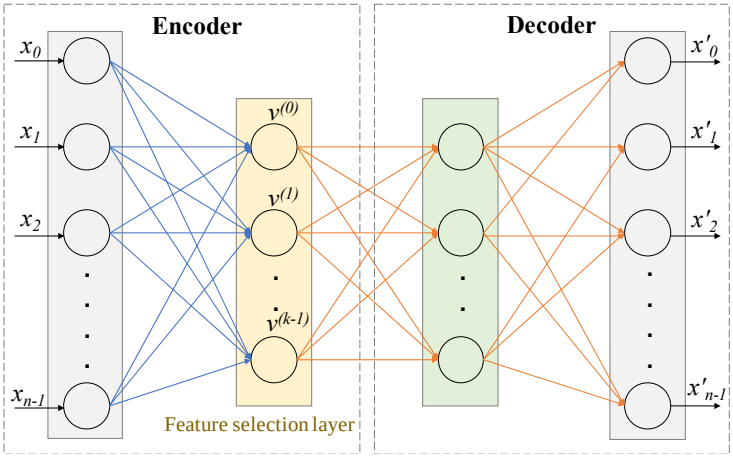

Figure 7.1: Architecture of Concrete Autoencoder. CAE architecture consists of an encoder and a decoder. The layer after input layer of encoder is called concrete feature selection layer shown in yellow. This layer has k number of node where each node is for each feature to be selected. Decoder is to check how well the input features can be reconstructed using the selected *k* features. Output layer has the same number of nodes as input layer. $X = [x_0, x_1, \ldots, x_{n-1}]$ = Input features. $X' = [x'_0, x'_1, \ldots, x'_{n-1}]$ = Reconstructed features.

How input features are selected depends on the *temperature* of the selection layer, which is modulated from a high value to a small value using a simple annealing schedule [23]. As the temperature of the selection layer approaches zero, the layer selects *k* individual input features. The decoder of a concrete autoencoder serves as the reconstruction function. It is the same as that of a standard autoencoder. Thus, the concrete autoencoder is a method for selecting a discrete set of *k* features optimized for an arbitrarily complex reconstruction function.

Training and Testing/Validation of CAE: The samples in a cohort are divided into 80/20 split for training and testing. In the training phase, 80% of samples are used to select the *k* informative features. In the testing/validation phase, 20% of samples are used to reconstruct their original features using the selected *k* features.

Hyperparameter Tuning for CAE

The hyperparameters of CAE were tuned for lncRNA expression data of 12 TCGA cancer types. We kept two of the parameters the same as used in the original CAE, developed by Abid et al. [23]. These two parameters are leaky ReLU with a threshold value of 0.1 and a 10% dropout rate. To tune the number of nodes in two hidden layers of the decoder, the model was tested by varying the number of nodes from 240 to 340, with a step size of 10. It was found that a decoder with 300 nodes in both layers yields the highest accuracy. So, the number of nodes in two hidden layers of the decoder was selected as 300.

To tune the number of epochs and learning rate, the random search [124] approach was used. The values used for the number of epochs were 200, 300, 500, 1000, 1500, 2000, 2500, and 3000. Similarly, for learning rate, the values were 0.001, 0.002, 0.005, 0.0005, 0.01, and 0.05. In every run of CAE, the values of the two hyperparameters were randomly

selected. With 300 epochs and a 0.002 learning rate, the 100 features selected by the CAE produced the highest accuracy in classifying 12 cancer types using SVM. So, these parameter values were chosen for further analysis. Details of hyperparameter tuning are available in *Appendix B.1*.

For every iteration of a single run in the hyperparameter tuning phase, temperature, mean-max probability (mean of maximum probabilities of the selected features), training loss, and validation loss were observed and plotted. The plot paints a clear picture of the learning process in CAE at every epoch, thereby naming it as the characteristic plot of CAE, as shown in Figure 7.2.

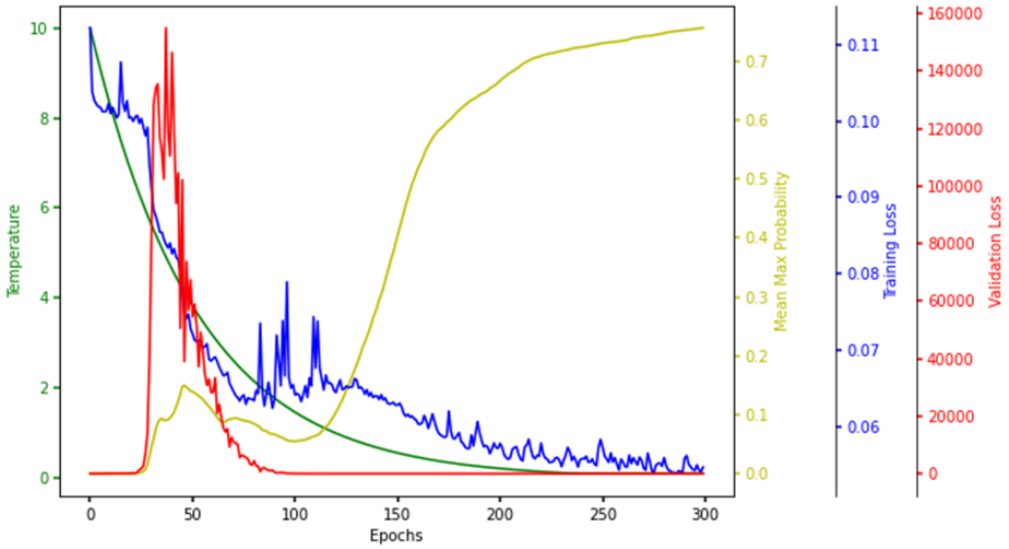

Figure 7.2: Characteristic Plot of Concrete Autoencoder. Temperature (green), Mean-max probability (yellow), training loss (blue), and validation loss (red) are plotted at different scales.

One of the main objectives of this plot is to see if the model is converging in terms of loss, which is evident in Figure 7.2, as the training loss and validation loss converge to a lower value. Each node in the concrete selection layer learns a probability value for every feature, and the node selects the one with the highest probability. The higher the mean-max

probability is, the more each node in the concrete selector is confident of one of the features. So, it is one of the goals to have the mean-max probability as high as possible.

7.2.3 Comparing mrCAE with Other Feature Selection Approaches

The feature selection capability of mrCAE was compared with the standard autoencoder (AE), three frequently used embedded feature selection models, including LASSO [123], Random Forest (RF) [45], and Support Vector Machine with Recursive Feature Elimination (SVM-RFE) [37], and two unsupervised feature selection models, MCFS [97] and UDFS [98]. The same number of features were selected using each approach, and those features were used to evaluate the classification performance in classifying 12 different cancer types. A stratified 5-fold cross-validation using SVM with linear kernel was conducted to evaluate the classification performance. Four different evaluation metrics - accuracy, precision, recall, f1 score - have been used to record the classification performance.

7.2.4 Implementation of Feature Selection Algorithms

All feature selection algorithms except mrCAE were implemented using the scikit-learn framework (https://scikit-learn.org/), whereas mrCAE was implemented using a deep learning framework named Keras (https://keras.io/). Experiments are parallelized on NVIDIA Quadro K620 GPU with 384 cores and 2GB memory devices. The dataset was split into the train and test set according to the 80/20 ratio to avoid overfitting. The training set was used to estimate the learning parameters, and the test set was used for performance evaluation.

7.3 Results

We analyzed the lncRNA expression profiles of 12 cancers with the goal of identifying the key lncRNAs using mrCAE. ***First***, we showed that the features selected by CAE are truly cancer-specific, not organ-specific. ***Second***, we showed the stochastic nature of CAE in selecting equally significant different sets of features in different runs. ***Third***, we showed that mrCAE performed better than the single-run CAE and other state-of-the-art feature selection methods, including LASSO, RF, SVM-RFE, MCFS, and UDFS. ***Fourth***, we determined a stable set of lncRNAs that not only can stratify 12 different cancer types but also have the highest number of lncRNAs with prognostic behavior. ***Fifth***, we identified the optimal number of runs for mrCAE.

7.3.1 Features Selected from Tumor Tissues are Cancer-Specific, not Organ-Specific

To check that the features selected by CAE from the lncRNA expression profiles of cancer samples are truly cancer-specific, not organ-specific, we ran CAE separately on tumor and normal samples to identify two sets of 80 features (lncRNAs). Figure 7.3(a) shows only five commons between 80 tumor and 80 normal features, which evidenced that 75 out of 80 features are unique to both tumor and normal tissues. It is clear from the t-SNE plots of Figure 7.3b and 3c that tumor and normal features can distinctively cluster 12 tumor tissues and corresponding normal tissues, respectively. However, when we do the cross, meaning the t-SNE plot of tumor tissues using normal features (Figure 7.3d) and t-SNE plot of normal tissues using tumor features (Figure 7.3e), there are no distinct clusters for 12 tumor and corresponding normal tissues. *Appendix B.2* shows similar results for 40-feature (*Figure S8)* and 60-feature scenarios (*Figure S9)*. These experiments proved that the features derived from tumor samples are truly cancer-specific, not organ-specific.

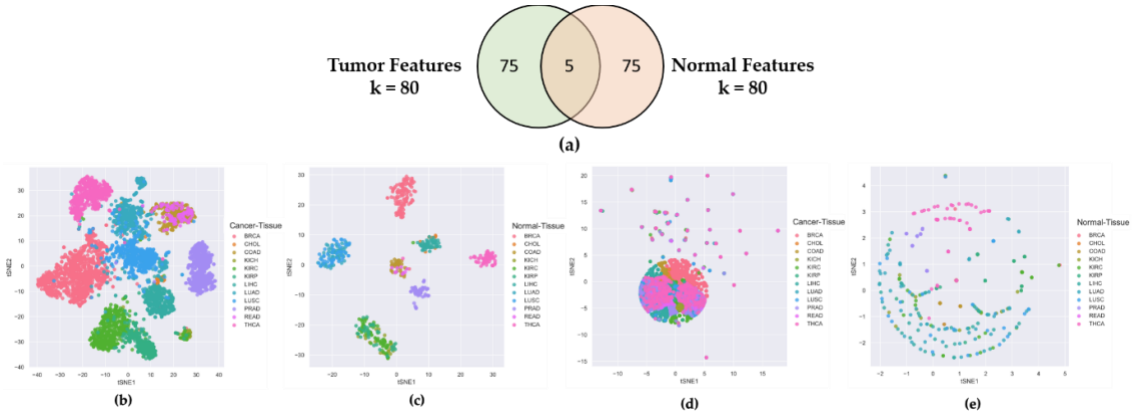

Figure 7.3: Comparing Tumor Features with Normal Features. a) Venn diagram of 80 tumor features and 80 normal features derived from CAE; b) t-SNE plot of tumor samples using tumor features; c) t-SNE plot of normal samples using normal features; d) t-SNE plot of tumor samples using normal features; e) t-SNE plot of normal samples using tumor features.

7.3.2 CAE Produces Different Sets of Significant Features in Different Runs

Though CAE selects a subset of the most significant features from a given dataset, it produces different sets of significant features in different runs due to its stochastic nature [23]. To show the stochastic nature of CAE, three sets of 60 features were selected for the experiment. Figure 7.4 shows (a) the Venn diagram, (b) classification accuracy of 12 cancer types, (c) mean squared error (MSE) of reconstructing original features, and (d) t-SNE plots of visualizing 12 different types of cancer samples, respectively, using three sets of features.

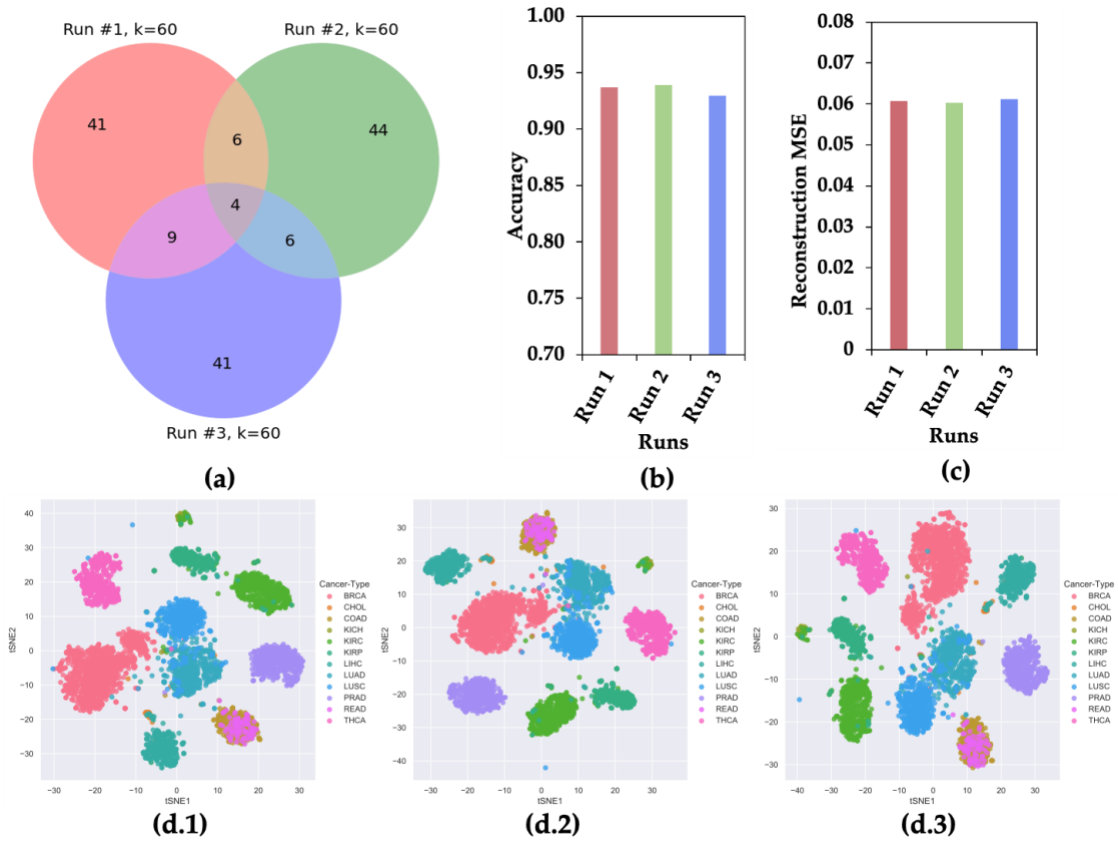

Figure 7.4: CAE Property of Selecting Different Sets of Features in Different Runs. a) Venn Diagram, b) accuracy of classifying 12 cancer types, (c) reconstruction mean squared error (MSE), and (d) t-SNE plots for 12 cancer samples using three sets of 60 features selected in three runs.

It is clear from the Venn diagram that few lncRNAs are common between any two runs (10, 10, and 13 of 60 lncRNAs). Though the three sets of 60 lncRNAs are different, they are equally good in classifying 12 different cancer types (Figure 7.4b) and reconstructing original features (Figure 7.4c). The t-SNE plots, Figure 7.4d, also support that the three sets of 60 lncRNAs are equally good in differentiating 12 cancer types. Thus, it is clear from Figure 7.4 that CAE selects different sets of most informative features in different runs. This observation motivated us to **hypothesize** that a feature appearing in multiple runs of CAE (mrCAE) carries the most meaningful information for a given dataset. In the next section, we showed that mrCAE performed better than the single-run CAE and other

state-of-the-art feature selection methods, including LASSO, RF, SVM-RFE, MCFS, and UDFS.

7.3.3 Comparison of mrCAE with Existing Feature Selection Approaches

Before comparing mrCAE with the existing feature selection approaches, we evaluated the performance of single-run CAE with a different number of selected features, which will guide us on how many features we should select for comparison. In Figure 7.5(a), it is noticeable that even with a smaller number of features, only with ten features, the average accuracy of CAE was close to 85%. There is a sharp increase in average accuracy (91%) with 20 features, followed by a slight increase (92% accuracy) up to 60 features. Then the curve reaches a plateau. From this figure, it seems like 40 features (before starting plateau) selected using different algorithms would be a good choice for comparison.

Selection of 40 Features from mrCAE: CAE was run 100 times to select 100 features in each run. In 100 runs, it selected a total of 534 unique features. The frequency of appearing these features in 100 runs ranges between 1 and 98. The 40 most frequent features, the top 40 features from the sorted list in descending order based on frequency, were used to measure the performance of mrCAE.

Figure 7.5(b) shows the classification performance using the sets of 40 lncRNAs selected from different feature selection algorithms, including LASSO, RF, SVM-RFE, MCFS, UDFS, AE, CAE, and mrCAE. It is clear that mrCAE performed better than any other feature selection approaches regarding the accuracy, recall, precision, and F1 score.

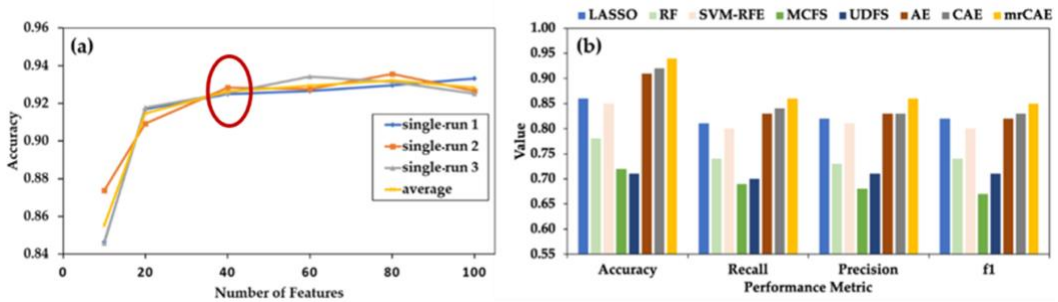

Figure 7.5: Comparing mrCAE with other feature selection approaches. (a) Behavior of single-run CAE to decide the number of features to be selected for comparison. CAE was run three times to select six sets of 10, 20, 40, 60, 80, and 100 features. "single avg" represents the average accuracy of three runs. (b) Classification performance using 40 features selected by LASSO, RF, SVM-RFE, MCFS, UDFS, AE, CAE, and mrCAE. Note that, each approach selects 40 actual features except AE, which selects 40 latent features.

7.3.4 mrCAE to Select a Stable Set of Features

mrCAE Systems: To identify a unique and stable set of lncRNAs that not only can distinguish between 12 different cancer types but also have the highest number of features with prognostic behavior, we designed mrCAE systems with 10, 20, 40, 60, 80, 100, and 120 runs. In each of the single runs of a mrCAE system, 100 lncRNAs were selected. Table 7.2 shows the summary statistics of mrCAE systems, including the total number of unique lncRNAs selected and the maximum frequency of a lncRNA appearing in each mrCAE system. The minimum frequency was 1 for all the different mrCAE systems. As shown in Table 7.2, a total of 223 unique lncRNAs (combined list of 10 sets of 100 lncRNAs) were selected by the 10-run mrCAE system, and the frequency of a lncRNA appearing in multiple runs ranged between 1 and 10. Similarly, a total of 575 unique lncRNAs were selected by the 120-run mrCAE system, and the frequency of a lncRNA appearing in multiple runs ranged between 1 and 117.

Table 7.2: Summary statistics of mrCAE systems in selecting lncRNAs. 100 lncRNAs were selected in each run of mrCAE.

mrCAE	Total LncRNAs	Min Frequency	Max Frequency
10-run mrCAE	223	1	10

20-run mrCAE	313	1	20
40-run mrCAE	400	1	40
60-run mrCAE	464	1	60
80-run mrCAE	499	1	80
100-run mrCAE	534	1	98
120-run mrCAE	575	1	117

Frequent and Stable Features: Features appearing more than once in mrCAE system were considered frequent features. Features with higher frequencies were considered stable features.

Table 7.3: Ranges of frequency for the top features in six categories.

	Ranges of Frequency					
mrCAE	**Top-10**	**Top-20**	**Top-40**	**Top-60**	**Top-80**	**Top-100**
10-run mrCAE	(10–10)	(9–10)	(6–10)	(4–10)	(3–10)	(2–10)
20-run mrCAE	(19–20)	(15–20)	(11–20)	(8–20)	(5–20)	(4–20)
40-run mrCAE	(36–40)	(29–40)	(22–40)	(15–40)	(11–40)	(8—40)
60-run mrCAE	(53–60)	(44–60)	(31–60)	(21–60)	(16–60)	(13–60)
80-run mrCAE	(69–80)	(60–80)	(42–80)	(28–80)	(22—80)	(17–80)
100-run mrCAE	(84–98)	(74–98)	(53–98)	(35–98)	(27–98)	(21–98)
120-run mrCAE	(99–117)	(85–117)	(62–117)	(44–117)	(34–117)	(25–117)

Top Frequent Features: The top frequent features, for example, Top-10 features in any mrCAE system, were the first ten features from the combined list sorted in descending order based on frequency. To identify a stable set of lncRNAs, we selected the top features from each of the seven mrCAE systems in six different categories: Top-10, Top-20, Top-40, Top-60, Top-80, and Top-100. Table 7.3 shows the ranges of frequency for the top features in six different categories. It is noticeable from both Tables 7.2 and 7.3 that the ***most frequent feature*** appeared in 10, 20, 40, 60, and 80 runs in the cases of 10-, 20-, 40-, 60- and 80-run mrCAE systems, respectively, but the trend was not maintained for 100-run (appeared in 98 runs) and 120-run (appeared in 117 runs) systems. In other words, the

most frequent feature appeared in each run of each mrCAE system except for the 100-run and 120-run systems, for which it (most frequent feature) appeared in 98 and 117 runs, respectively. It can be concluded that for the given lncRNA expression profile dataset of 12 cancers, the mrCAE system with 100 or more runs could not produce the most frequent features in each run. Thus, a 100-run mrCAE can be considered the optimal configuration for this dataset, and the results from 120-run mrCAE were not considered for subsequent analyses.

Finally, this experiment resulted in six unique sets of features corresponding to Top-10, Top-20, Top-40, Top-60, Top-80, and Top-100 features, as shown in the Venn diagram of Figure 7.6. For example, combining six sets of top-10 features from 10-, 20-, 40-, 60-, 80-, and 100-run mrCAE systems produced a unique list of 14 lncRNAs.

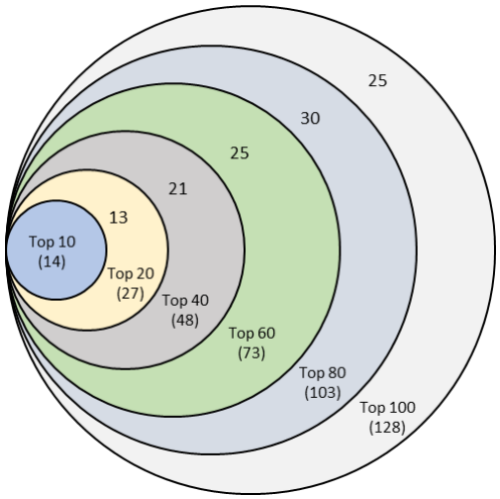

Figure 7.6: Venn diagram of six sets of unique features identified from six mrCAE systems. The mrCAE consisted of 10, 20, 40, 60, 80, and 100 runs. Each of these runs was conducted to select 100 features. The smallest set (light blue), containing 14 features, represents the unique features coming from six sets of 10 most frequent features from 10-, 20-, 40-, 60-, 80-, and 100-run mrCAE systems. Similarly, the 2nd smallest set contains 27 (14 + 13) unique features from six sets of Top-20 features selected.

The Venn diagram shows that each set of unique features was a subset of the following more extensive unique feature set. Finally, we can conclude that the 128 unique features (*Appendix B.3 Table S9*)—produced from the union of six sets of Top-100 features coming from 10-, 20-, 40-, 60-, 80-, and 100-run mrCAE systems—represented the stable and optimal feature set. We used this set of lncRNAs to conduct the downstream study, including survival and prognostic analyses and validation.

7.3.5 Prognostic Capability of Significant lncRNAs

To evaluate the prognostic capabilities of the selected 128 stable lncRNAs, survival analyses of patients with different cancer types were performed. Any lncRNAs with zero expression values for most cancer samples were excluded from the survival analysis of that cancer. The patients with values less than or equal to the median were labeled group A. Those with values greater than the median were labeled group B. After dividing into two groups, a log-rank test was conducted, and the hazard ratio was calculated as the hazard rate of group A vs. hazard rate of group B to check the prognostic capability of a lncRNA. The criteria for a lncRNA to be prognostic are log-rank test p-value ≤ 0.05 and Hazard Ratio (HR) $\neq 1.0$. Kaplan–Meier curves were plotted to show the prognostic behavior of lncRNAs.

Figure 7.7(a) shows the Kaplan–Meier plot for GATA3-AS1, one of the 11 prognostic lncRNAs for breast cancer, and Figure 7.7(b) shows the forest plot of survival analyses for 11 prognostic lncRNAs. It can be observed from Figure 7.7a that group B (red) had a higher rate of survival than group A (blue), meaning that lncRNA GATA3-AS1 could successfully distinguish the high-risk group (Group A) of BRCA patients from the low-risk group (Group B). In other words, the cohort with a low expression (blue) of GATA3-AS1 had a

1.53-times higher rate of death than the high-expression cohort (red). Thus, the cohorts with low-expression values for seven lncRNAs (HR > 1.0) showed higher chances of death compared to the high-expression cohorts (Figure 7.7b). On the other hand, the cohorts with low-expression values for four lncRNAs (HR < 1.0) showed lower chances of death compared to the high-expression cohorts (Figure 7.7b). *Appendix B.4* shows the forest plots for other cancer types.

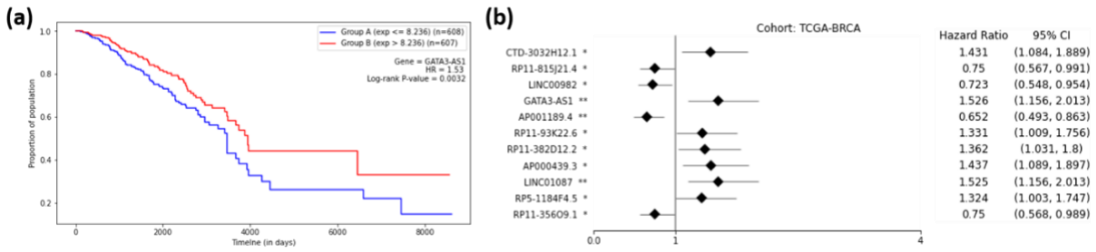

Figure 7.7: Survival Analysis of TCGA-BRCA. (a) Kaplan–Meier Curve for the GATA3-AS1 lncRNA on the TCGA-BRCA cohort. Group A (blue) is the group with an expression less than or equal to the median, and Group B (red) is the group with an expression greater than the median. (b) Forest plot of survival analysis for 11 prognostic lncRNAs on the BRCA cohort. The asterisks represent the log-rank *p*-values (*—$p \leq 0.05$, **—$p \leq 0.01$, ***—$p \leq 0.001$, ****—$p \leq 0.0001$).

The number of prognostically significant lncRNAs for each type of cancer is given in Table 7.4. The highest number of prognostic lncRNAs were discovered for KIRC (31 lncRNAs), followed by LUAD (22 lncRNAs) and LUSC (18 lncRNAs). The proposed approach failed to discover any prognostic lncRNA for CHOL, potentially because the cohort consisted of only 36 patients (Table 7.1). Some of the lncRNAs were found to be prognostic for more than one cancer. Of the stable set of 128 lncRNAs, 76 were prognostic.

Table 7.4: Summary of survival analysis regarding the number of prognostic lncRNAs for each of the 12 TCGA cancer types.

BRCA	CHOL	COAD	KICH	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA	Total
11	0	3	3	31	15	1	22	18	4	4	10	**76**

7.3.6 Validations

The stable set of 128 lncRNAs derived from mrCAE was validated with the existing literature [125]. Of 128 lncRNAs, 103 were found to be known lncRNAs associated with different cancer types, as shown in Figure 7.8(a). For example, 98 lncRNAs are associated with BRCA, 52 lncRNAs are related to LUAD, and 37 lncRNAs are related to KIRP. Some lncRNAs were also found in four different cancer hallmarks, Figure 7.8(b); for example, six lncRNAs were found to be related to cancer prognosis. We also validated the top 128 lncRNAs with existing drug–lncRNA networks. A drug–lncRNA network was formed based on the Spearman correlation coefficient between lncRNA expression levels and the IC50 values of the drug [126]. We found that 113 out of 128 lncRNAs are associated with 24 different drugs primarily used in cancer-related treatments, as shown in Figure 7.8(c&d). For example, the drug nilotinib is mainly used to treat a specific type of blood cancer associated with 18 different lncRNAs, Figure 7.8(e).

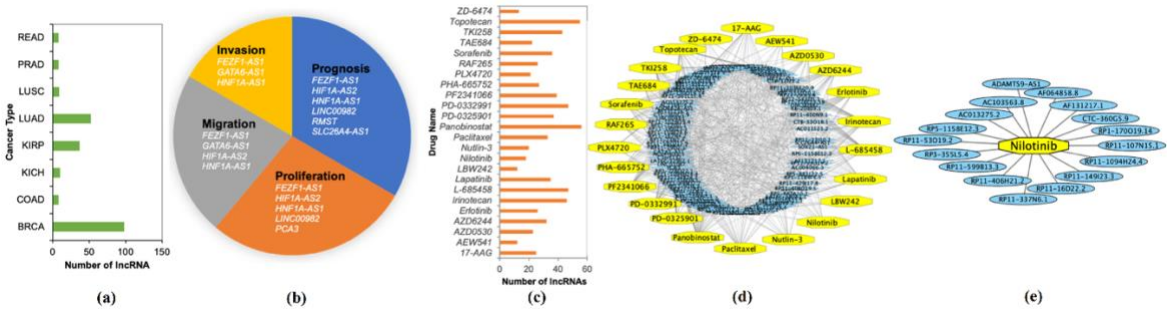

Figure 7.8: Validation of Identified lncRNAs. (a) Number of known lncRNAs derived by mrCAE related to different cancer types found in [127]–[130]; (b) mrCAE derived lncRNAs related to different cancer hallmarks [131]. (c) number of lncRNAs related to different cancer drugs [126]; (d) drug–lncRNA networks for all 24 drugs; (e) an example lncRNA–drug network for nilotinib, which is used to treat certain blood cancers associated with 18 different lncRNAs.

7.4 Discussion

The objective of the present study was to identify significant lncRNAs that carry meaningful information on (a) identifying the origins of multiple cancers, (b) evaluating the prognostic capability of differentiating high-risk and low-risk groups of patients of particular cancers, and (c) having potential for targeted therapy. The original CAE algorithm is capable of identifying subsets of important features. However, due to the stochastic nature of the algorithm, it produces different subsets in different runs [23], prohibiting its application in precision medicine. Thus, we hypothesized that the most frequently appearing lncRNAs in multiple runs of CAE (mrCAE) would produce a biologically meaningful and stable set of features.

Our investigation showed that the lncRNAs selected by the proposed mrCAE outperformed both the single-run CAE and the standard autoencoder, along with other feature selection approaches, in identifying the origins of multiple cancers, as shown in Figure 7.5. Thus, the current results confirmed that the proposed mrCAE could be utilized as a tool for identifying a stable set of meaningful features. It should be noted that the proposed mrCAE approach is very similar to a common bioinformatic approach of bootstrapping analysis used to evaluate the stability of results. The bootstrapping effect could be the reason that mrCAE performs better than the CAE and standard AE.

Our results showed that the lncRNAs selected by the proposed mrCAE carry meaningful information on the prognostic capability of differentiating high- and low-risk groups of patients of particular cancer, as explained in Section 7.3.5. We also showed the biological relevance of the selected lncRNAs by comparing them with existing literature, hallmark

lncRNAs, and drug–lncRNA networks, Figure 7.8. Thus, the lncRNAs selected by mrCAE can be used as possible targets for therapy.

CHAPTER 8 CONCLUSIONS AND FUTURE DIRECTIONS

8.1 Feature Selection and Cancer Type Classification (8 Cancers)

We developed a computational framework to identify key lncRNAs for multiple cancer types, employing two feature selection and five classification methods only using lncRNA expression of cancer samples. This study identified 37 key lncRNAs that can classify eight cancer types with an accuracy ranging from 95% to 98%. The t-SNE plot showing eight distinct clusters for eight cancer types supports that the discovered 37 lncRNAs can differentiate eight different cancer types. The survival analyses of individual lncRNA support that the discovered lncRNAs also have the prognostic capability of differentiating between high-risk and low-risk patients. Thus, the discovered lncRNAs can be used as diagnostic and prognostic features for eight different cancer types considered in this study.

8.2 Feature Selection and Cancer Type Classification (33 Cancers)

This research extended our feature selection framework by integrating a deep learning-based unsupervised feature selection algorithm, concrete autoencoder, to identify the key features. The proposed method was evaluated in identifying the origin of 33 different cancer types using the expression profiles of selected features (69 lncRNAs) from the original feature space of 12 thousand lncRNAs. Validation with the existing literature and survival analyses supports that the selected lncRNAs could be potential biomarkers for the diagnosis and prognosis of 33 different cancers. This research accounts for feature selection and identifying the origin of different cancers.

8.3 Feature Selection and Cancer Type Classification (Glycome genes)

We also developed an in-silico framework to identify significant glycome genes related to the origins of 33 different cancers. The same deep learning-based unsupervised feature

selection algorithm, concrete autoencoder, was used to develop the framework. The developed framework identified an optimal set of glycome genes related to 33 cancer types. This optimal set of glycome genes can differentiate 33 cancers using expression profiles with an accuracy of 92%. This part of the research accounts for feature selection and identifying the origin of different cancer types using a subset of glycome genes. These findings highlight the importance of cell-type-specific glycosylation in cancer development and offer subsets of glycome genes in several molecular categories that can be investigated for their respective role in cancer-specific malignancy.

8.4 Class-Specific Feature Selection

An embedded feature selection method, recursive l1-norm multiclass SVM, was proposed to identify class-specific features for a multiclass problem while classifying. Using lncRNA expression profiles, the proposed method experimented with five molecular subtypes of breast cancer patients. The proposed method effectively identified small subsets of subtype-specific important lncRNAs, while classifying the breast cancer patients into five subtypes. Experimental results and validation support that the selected lncRNAs could be potential biomarkers for breast cancer diagnosis and prognosis.

8.5 Multi-Run Concrete Autoencoder for Feature Selection

Finally, we proposed a multi-run concrete autoencoder (mrCAE) to identify prognostic lncRNAs for multiple cancers. We tested the extended model in analyzing the lncRNA expression profiles of 12 cancers. The model selected a stable set of lncRNAs that could differentiate 12 cancers with high accuracy and provide subsets of prognostic lncRNAs for 12 cancers. The lncRNAs selected by the proposed mrCAE outperformed the lncRNAs selected by the single-run CAE and other feature selection approaches. Additionally, the

proposed mrCAE outperformed the standard autoencoder, which selected the latent features and was thought to be the upper limit in dimension reduction. Since the proposed mrCAE outperformed AE and can select actual features in contrast to latent features by AE, it can provide meaningful information that can be used for precision medicine, such as identifying prognostic lncRNAs for different cancers.

8.6 Future Directions

This work can be extended to the simultaneous discovery of three types of relevant RNAs, considering the expression profiles of 60K RNAs, including mRNA, lncRNA, and miRNA, which will provide a comprehensive list of RNA biomarkers related to each type or subtype of cancer. It can also be used to integrate multi-omics data such as DNA methylation and histone modification.

This study considers only cancer patients to identify cancer-related glycome genes. In the future, we will use the same framework for normal samples corresponding to different cancers to find the glycome genes related to normal tissues. Comparing these two sets will help pinpoint the glycome gene signatures for cancers.

Though the proposed mrCAE model was applied to multiple cancer types, we can also use it on a single cancer type to identify the cancer-specific informative features, as used in identifying the informative features for single-digit MNIST data by the developer of CAE.

BIBLIOGRAPHY

[1] S. Chakraborty and T. Rahman, "The difficulties in cancer treatment," *Ecancermedicalscience*, vol. 6, p. ed16, 2012, doi: 10.3332/ECANCER.2012.ED16.

[2] D. Ramazzotti, A. Lal, B. Wang, … S. B.-N., and undefined 2018, "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival," *nature.com*, Accessed: Oct. 03, 2020. [Online]. Available: https://www.nature.com/articles/s41467-018-06921-8?cid=tw%26p.

[3] A. Kinnaird, S. Zhao, K. W.-N. R. Cancer, and undefined 2016, "Metabolic control of epigenetics in cancer," *nature.com*, Accessed: Oct. 03, 2020. [Online]. Available: https://www.nature.com/articles/nrc.2016.82?cacheBust=1508292600345.

[4] M. M. Suzuki and A. Bird, "DNA methylation landscapes: Provocative insights from epigenomics," *Nature Reviews Genetics*, vol. 9, no. 6. Nature Publishing Group, pp. 465–476, Jun. 2008, doi: 10.1038/nrg2341.

[5] J. E. Audia and R. M. Campbell, "Histone Modifications and Cancer," *cshperspectives.cshlp.org*, doi: 10.1101/cshperspect.a019521.

[6] C. M. Croce, "Causes and consequences of microRNA dysregulation in cancer," *Nature Reviews Genetics*, vol. 10, no. 10. Nature Publishing Group, pp. 704–714, Oct. 2009, doi: 10.1038/nrg2634.

[7] H. S. Chiu *et al.*, "Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context," *Cell Rep.*, vol. 23, no. 1, pp. 297-312.e12, Apr. 2018, doi: 10.1016/j.celrep.2018.03.064.

[8] J. Prensner, A. C.-C. discovery, and undefined 2011, "The emergence of lncRNAs in cancer biology," *AACR*, Accessed: Feb. 05, 2022. [Online]. Available: https://cancerdiscovery.aacrjournals.org/content/1/5/391.short.

[9] H. Tao, J.-J. Yang, X. Zhou, Z.-Y. Deng, K.-H. Shi, and J. Li, "Emerging role of long noncoding RNAs in lung cancer: Current status and future prospects," *Respir. Med.*, vol. 110, pp. 12–19, 2016.

[10] A. M. Schmitt and H. Y. Chang, "Long noncoding RNAs in cancer pathways," *Cancer Cell*, vol. 29, no. 4, pp. 452–463, 2016.

[11] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011, doi: https://doi.org/10.1016/j.cell.2011.02.013.

[12] J. Harrow *et al.*, "GENCODE: The reference human genome annotation for The ENCODE Project," *genome.cshlp.org*, doi: 10.1101/gr.135350.111.

[13] Y. Kondo, K. Shinjo, and K. Katsushima, “Long non-coding RNAs as an epigenetic regulator in human cancers,” *Cancer Science*, vol. 108, no. 10. Blackwell Publishing Ltd, pp. 1927–1933, Oct. 01, 2017, doi: 10.1111/cas.13342.

[14] J. Pirgazi, M. Alimoradi, T. E. Abharian, and M. H. Olyaee, “An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, 2019.

[15] S. Liang, A. Ma, S. Yang, Y. Wang, and Q. Ma, “A review of matched-pairs feature selection methods for gene expression data analysis,” *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 88–97, 2018.

[16] Y. Li *et al.*, “A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data,” *BMC Genomics*, vol. 18, no. 1, pp. 1–13, Jul. 2017, doi: 10.1186/S12864-017-3906-0/FIGURES/4.

[17] S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, and E. Medico, “Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer,” *Sci. Reports 2020 101*, vol. 10, no. 1, pp. 1–13, Aug. 2020, doi: 10.1038/s41598-020-70832-2.

[18] Z. Yu, Z. Wang, X. Yu, and Z. Zhang, “RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches,” *Comput. Intell. Neurosci.*, vol. 2020, 2020, doi: 10.1155/2020/4737969.

[19] P. Domingos, “review articles Tapping into the ‘folk knowledge’ needed to advance machine learning applications,” vol. 55, no. 10, 2012, doi: 10.1145/2347736.2347755.

[20] J. S. Parker *et al.*, “Supervised risk predictor of breast cancer based on intrinsic subtypes,” *J. Clin. Oncol.*, vol. 27, no. 8, p. 1160, 2009.

[21] C. Fan *et al.*, “Concordance among gene-expression--based predictors for breast cancer,” *N. Engl. J. Med.*, vol. 355, no. 6, pp. 560–569, 2006.

[22] A. Al Mamun and A. M. Mondal, “Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM)*, 2019, pp. 2825–2831.

[23] A. Abid, M. F. Balin, and J. Zou, “Concrete autoencoders: Differentiable feature selection and reconstruction,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 694–711, 2019.

[24] A. Al Mamun, W. Duan, and A. M. Mondal, “Pan-cancer Feature Selection and Classification Reveals Important Long Non-coding RNAs,” in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, Dec. 2020, pp. 2417–2424, doi: 10.1109/BIBM49941.2020.9313332.

[25] A. Al Mamun *et al.*, "Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers," *Int. J. Mol. Sci. 2021, Vol. 22, Page 11919*, vol. 22, no. 21, p. 11919, Nov. 2021, doi: 10.3390/IJMS222111919.

[26] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[27] T. Speed, *Statistical analysis of gene expression microarray data*. Chapman and Hall/CRC, 2003.

[28] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 02, pp. 185–205, 2005.

[29] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, 2015.

[30] Y. Sun, C. Lu, and X. Li, "The cross-entropy based multi-filter ensemble method for gene selection," *Genes (Basel).*, vol. 9, no. 5, p. 258, 2018.

[31] A. Rau, M. Flister, H. Rui, and P. L. Auer, "Exploring drivers of gene expression in the Cancer Genome Atlas," *Bioinformatics*, vol. 35, no. 1, pp. 62–68, 2019.

[32] I.-S. Jeong, H.-K. Kim, T.-H. Kim, D. H. Lee, K. J. Kim, and S.-H. Kang, "A feature selection approach based on simulated annealing for detecting various denial of service attacks," *Softw. Netw.*, vol. 2018, no. 1, pp. 173–190, 2018.

[33] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, 2012.

[34] Y.-L. Wu, C.-Y. Tang, M.-K. Hor, and P.-F. Wu, "Feature selection using genetic algorithm and cluster validation," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2727–2732, 2011.

[35] M. M. Kabir, M. Shahjahan, and K. Murase, "A new hybrid ant colony optimization algorithm for feature selection," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3747–3763, 2012.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[37] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.

[38] J. Fang, "Tightly integrated genomic and epigenomic data mining using tensor decomposition," *Bioinformatics*, vol. 35, no. 1, pp. 112–118, 2019.

[39] M. B. Pouyan and D. Kostka, “Random forest based similarity learning for single cell RNA sequencing data,” *Bioinformatics*, vol. 34, no. 13, pp. i79--i88, 2018.

[40] M. Ram, A. Najafi, and M. T. Shakeri, “Classification and biomarker genes selection for cancer gene expression data using random forest,” *Iran. J. Pathol.*, vol. 12, no. 4, p. 339, 2017.

[41] R. Wang, “AdaBoost for feature selection, classification and its relation with SVM, a review,” *Phys. Procedia*, vol. 25, pp. 800–807, 2012.

[42] T. T. Le, R. J. Urbanowicz, J. H. Moore, and B. A. McKinney, “Statistical inference Relief (STIR) feature selection,” *Bioinformatics*, vol. 35, no. 8, pp. 1358–1365, 2019.

[43] X. Lu, H. Gu, Y. Wang, J. Wang, and P. Qin, “Autoencoder based feature selection method for classification of anticancer drug response,” *Front. Genet.*, vol. 10, p. 233, 2019.

[44] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[45] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, “Variable selection using random forests,” *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2225–2236, 2010.

[46] A. Abid, M. F. Balin, and J. Zou, “Concrete autoencoders for differentiable feature selection and reconstruction,” *arXiv Prepr. arXiv1901.09346*, 2019.

[47] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, 2006.

[48] C. Cortes, V. Vapnik, and L. Saitta, “Support-Vector Networks Editor,” *Mach. Leaming*, vol. 20, pp. 273–297, 1995.

[49] S. W. Cheetham, F. Gruhl, J. S. Mattick, and M. E. Dinger, “Long noncoding RNAs and the genetics of cancer,” *Br. J. Cancer*, vol. 108, no. 12, p. 2419, 2013.

[50] Y. Fang and M. J. Fullwood, “Roles, functions, and mechanisms of long non-coding RNAs in cancer,” *Genomics. Proteomics Bioinformatics*, vol. 14, no. 1, pp. 42–54, 2016.

[51] X. Zhang *et al.*, “Mechanisms and functions of long non-coding RNAs at multiple regulatory levels,” *International Journal of Molecular Sciences*, vol. 20, no. 22. MDPI AG, Nov. 02, 2019, doi: 10.3390/ijms20225573.

[52] K. A. Hoadley *et al.*, “Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer,” *Cell*, vol. 173, no. 2, pp. 291–304, 2018.

[53] B. Lyu and A. 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*, 2018, pp. 89–96.

[54] L. Sun, X. Kong, J. Xu, R. Zhai, S. Zhang, and others, “A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification,” *Sci. Rep.*, vol. 9, no. 1, p. 8978, 2019.

[55] F. Zhang, H. L. Kaufman, Y. Deng, and R. Drabier, “Recursive SVM biomarker selection for early detection of breast cancer in peripheral blood,” *BMC Med. Genomics*, vol. 6, no. 1, p. S4, 2013.

[56] Y. Zhang, Q. Deng, W. Liang, and X. Zou, “An efficient feature selection strategy based on multiple support vector machine technology with gene expression data,” *Biomed Res. Int.*, vol. 2018, 2018.

[57] L. Han *et al.*, “The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes,” *Nat. Commun.*, vol. 5, p. 3963, 2014.

[58] H. Hotelling, “Analysis of a complex of statistical variables into principal components.,” *J. Educ. Psychol.*, vol. 24, no. 6, p. 417, 1933.

[59] A. Mirzaei, V. Pourahmadi, M. Soltani, and H. Sheikhzadeh, “Deep feature selection using a teacher-student network,” *Neurocomputing*, vol. 383, pp. 396–408, 2020.

[60] Y. Lu, Y. Fan, J. Lv, and W. S. Noble, “DeepPINK: reproducible feature selection in deep neural networks,” in *Advances in Neural Information Processing Systems*, 2018, pp. 8676–8686.

[61] V. Borisov, J. Haug, and G. Kasneci, “CancelOut: A Layer for Feature Selection in Deep Neural Networks,” in *International Conference on Artificial Neural Networks*, 2019, pp. 72–83.

[62] Y. Kotake *et al.*, “Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15p15 INK4B tumor suppressor gene,” *Oncogene*, vol. 30, no. 16, p. 1956, 2011.

[63] F. Yang *et al.*, “Up-regulated long non-coding RNA H19 contributes to proliferation of gastric cancer cells,” *FEBS J.*, vol. 279, no. 17, pp. 3159–3165, 2012.

[64] S. Redon, P. Reichenbach, and J. Lingner, “The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase,” *Nucleic Acids Res.*, vol. 38, no. 17, pp. 5797–5806, 2010.

[65] L. H. Schmidt *et al.*, “The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth,” *J. Thorac. Oncol.*, vol. 6, no. 12, pp. 1984–1992, 2011.

[66] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, and P. P. Pandolfi, “A coding-independent function of gene and pseudogene mRNAs regulates tumour biology,” *Nature*, vol. 465, no. 7301, p. 1033, 2010.

[67] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: the next generation,” *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[68] C.-H. Wu, C.-L. Hsu, P.-C. Lu, W.-C. Lin, H.-F. Juan, and H.-C. Huang, “Identification of lncRNA functions in lung cancer based on associated protein-protein interaction modules,” *Sci. Rep.*, vol. 6, p. 35939, 2016.

[69] G. Zhang *et al.*, “Identification of cancer-related miRNA-lncRNA biomarkers using a basic miRNA-lncRNA network,” *PLoS One*, vol. 13, no. 5, p. e0196681, 2018.

[70] W. Xing *et al.*, “Genome-wide identification of lncRNAs and mRNAs differentially expressed in non-functioning pituitary adenoma and construction of an lncRNA-mRNA co-expression network,” *Biol. Open*, vol. 8, no. 1, p. bio037127, 2019.

[71] L.-X. Wang, C. Wan, Z.-B. Dong, B.-H. Wang, H.-Y. Liu, and Y. 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,” *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.*, vol. 25, p. 2896, 2019.

[72] J. Sui *et al.*, “Integrated analysis of long non-coding RNA-associated ceRNA network reveals potential lncRNA biomarkers in human lung adenocarcinoma,” *Int. J. Oncol.*, vol. 49, no. 5, pp. 2023–2036, 2016.

[73] Y. Chen, Y. Pan, Y. Ji, L. Sheng, and X. Du, “Network analysis of differentially expressed smoking-associated mRNAs, lncRNAs and miRNAs reveals key regulators in smoking-associated lung cancer,” *Exp. Ther. Med.*, vol. 16, no. 6, pp. 4991–5002, 2018.

[74] A. Lanzós *et al.*, “Discovery of cancer driver long noncoding RNAs across 1112 tumour genomes: new candidates and distinguishing features,” *Sci. Rep.*, vol. 7, p. 41544, 2017.

[75] X. Zhang, J. Wang, J. Li, W. Chen, and C. Liu, “CRlncRC: a machine learning-based method for cancer-related long noncoding RNA identification using integrated features,” *BMC Med. Genomics*, vol. 11, no. 6, p. 120, 2018.

[76] A. Al Mamun and A. M. 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*, 2019, p. 541.

[77] M. Goldman *et al.*, “The UCSC Xena Platform for cancer genomics data visualization and interpretation,” *BioRxiv*, p. 326470, 2019.

[78] J. Li *et al.*, “TANRIC: an interactive open platform to explore the function of lncRNAs in cancer,” *Cancer Res.*, vol. 75, no. 18, pp. 3728–3737, 2015.

[79] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *J. Stat. Softw.*, vol. 33, no. 1, p. 1, 2010.

[80] R.-J. Palma-Mendoza, D. Rodriguez, and L. De-Marcos, “Distributed ReliefF-based feature selection in Spark,” *Knowl. Inf. Syst.*, vol. 57, no. 1, pp. 1–20, 2018.

[81] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python ,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[82] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in {P}ython,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[83] E. D. D. Team, “Deeplearning4j: Open-source distributed deep learning for the JVM,” *Apache Softw. Found. Licens. 2.0. http//deeplearning4j.org*.

[84] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*, vol. 454. Springer Science & Business Media, 2012.

[85] C. J. Maddison, A. Mnih, and Y. W. Teh, “The concrete distribution: A continuous relaxation of discrete random variables,” *arXiv Prepr. arXiv1611.00712*, 2016.

[86] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv Prepr. arXiv1312.6114*, 2013.

[87] M. Goldman, B. Craft, A. Brooks, J. Zhu, and D. Haussler, “The UCSC Xena Platform for cancer genomics data visualization and interpretation,” *BioRxiv*, p. 326470, 2018.

[88] Mart\’\in~Abadi *et al.*, “{TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems.” 2015, [Online]. Available: https://www.tensorflow.org/.

[89] F. Chollet and others, “Keras.” GitHub, 2015.

[90] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, 1958.

[91] E. A. Mauger, R. A. Wolfe, and F. K. Port, “Transient effects in the Cox proportional hazards regression model,” *Stat. Med.*, vol. 14, no. 14, pp. 1553–1565, 1995.

[92] A. Al Mamun and A. M. Mondal, “Long Non-coding RNA Based Cancer Classification using Deep Neural Networks,” in *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM BCB’19)*, 2019, pp. 541–541.

[93] J. G. Sweeney *et al.*, “Loss of GCNT2/I-branched glycans enhances melanoma growth and survival,” *Nat. Commun.*, vol. 9, no. 1, pp. 1–18, Dec. 2018, doi: 10.1038/s41467-018-05795-0.

[94] C. J. Dimitroff, A. Sharma, and R. J. Bernacki, “Cancer metastasis: A search for therapeutic inhibition,” *Cancer Investigation*, vol. 16, no. 4. Informa Healthcare, pp. 279–290, 1998.

[95] S. R. Barthel, J. D. Gavino, L. Descheny, and C. J. Dimitroff, “Targeting selectins and selectin ligands in inflammation and cancer,” *Expert Opinion on Therapeutic Targets*, vol. 11, no. 11. pp. 1473–1491, 2007.

[96] C. J. Dimitroff, “Galectin-binding O-glycosylations as regulators of malignancy,” *Cancer Research*, vol. 75, no. 16. American Association for Cancer Research Inc., pp. 3195–3202, Aug. 2015.

[97] D. Cai, C. Zhang, and X. He, “Unsupervised feature selection for Multi-Cluster data,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 333–342.

[98] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, “L2,1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning,” in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[99] K. Han, Y. Wang, C. Zhang, C. Li, and C. Xu, “Autoencoder Inspired Unsupervised Feature Selection,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April, pp. 2941–2945, doi: 10.1109/ICASSP.2018.8462261.

[100] M. Goldman *et al.*, “The UCSC Xena platform for public and private cancer genomics data visualization and interpretation,” *bioRxiv*, p. 326470, Sep. 2018, doi: 10.1101/326470.

[101] S. Zhang *et al.*, “lncRNA gene signatures for prediction of breast cancer intrinsic subtypes and prognosis,” *Genes (Basel).*, vol. 9, no. 2, p. 65, 2018.

[102] T. Sørlie *et al.*, “Repeated observation of breast tumor subtypes in independent gene expression data sets,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 14, pp. 8418–8423, 2003.

[103] C. M. Perou *et al.*, “Molecular portraits of human breast tumours.,” *Nature*, vol. 406, no. 6797, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.

[104] R. A. Youness and M. Z. Gad, “Long non-coding RNAs: functional regulatory players in breast cancer,” *Non-coding RNA Res.*, 2019.

[105] M. Pecero, J. Salvador-Bofill, and S. Molina-Pinelo, “Long non-coding RNAs as monitoring tools and therapeutic targets in breast cancer,” *Cell. Oncol.*, vol. 42, no. 1, pp. 1–12, 2019.

[106] J. J. Chan and Y. Tay, “Noncoding RNA: RNA regulatory networks in cancer,” *Int. J. Mol. Sci.*, vol. 19, no. 5, p. 1310, 2018.

[107] L. Ma, V. B. Bajic, and Z. Zhang, “On the classification of long non-coding RNAs,” *RNA Biol.*, vol. 10, no. 6, pp. 924–933, 2013.

[108] C. Jiang *et al.*, “Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs,” *Oncotarget*, vol. 7, no. 6, p. 7120, 2016.

[109] F. Yang, S. Lyu, S. Dong, Y. Liu, X. Zhang, and O. Wang, “Expression profile analysis of long noncoding RNA in HER-2-enriched subtype breast cancer by next-generation sequencing and bioinformatics,” *Onco. Targets. Ther.*, vol. 9, p. 761, 2016.

[110] X. Shen *et al.*, “Identification of novel long non-coding RNAs in triple-negative breast cancer,” *Oncotarget*, vol. 6, no. 25, p. 21730, 2015.

[111] T. Aqila, A. Al Mamun, and A. M. Mondal, “Pseudotime Based Discovery of Breast Cancer Heterogeneity,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2019)*, 2019.

[112] A. Al Mamun and A. M. Mondal, “Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers,” in *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, 2019, pp. 2825–2831, doi: 10.1109/BIBM47256.2019.8983413.

[113] Z. Wang *et al.*, “lncRNA Epigenetic Landscape Analysis Identifies EPIC1 as an Oncogenic lncRNA that Interacts with MYC and Promotes Cell-Cycle Progression in Cancer,” *Cancer Cell*, vol. 33, no. 4, pp. 706-720.e9, Apr. 2018.

[114] T. Sørlie *et al.*, “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 19, pp. 10869–10874, Sep. 2001.

[115] Z. Hu *et al.*, “The molecular portraits of breast tumors are conserved across microarray platforms,” *BMC Genomics*, vol. 7, no. 1, p. 96, Apr. 2006.

[116] L. Perreard *et al.*, “Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay,” *Breast Cancer Res.*, vol. 8, no. 2, p. R23, Apr. 2006, doi: 10.1186/bcr1399.

[117] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, “Diagnosis of multiple cancer types by shrunken centroids of gene expression,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 10, pp. 6567–6572, May 2002, doi: 10.1073/pnas.082099299.

[118] L. Wang and X. Shen, “On L 1-norm multiclass support vector machines: methodology and theory,” *J. Am. Stat. Assoc.*, vol. 102, no. 478, pp. 583–594, 2007.

[119] O. Chapelle and S. S. Keerthi, “Multi-class feature selection with support vector machines,” in *Proceedings of the American statistical association*, 2008, vol. 58.

[120] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, “Feature selection for SVMs,” in *Advances in neural information processing systems*, 2001, pp. 668–674.

[121] A. Al Mamun and A. M. Mondal, “Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers,” *2019 IEEE Int. Conf. Bioinforma. Biomed. (IEEE BIBM)*, pp. 2825–2831, 2019.

[122] R. B. Tanvir, M. Sobhan, A. Al Mamun, and A. M. Mondal, “Quantifying Intratumor Heterogeneity by Key Genes Selected Using Concrete Autoencoder,” *bioRxiv*, p. 2021.09.06.459161, Sep. 2021, doi: 10.1101/2021.09.06.459161.

[123] M. Cargnello and P. P. Roux, “Activation and Function of the MAPKs and Their Substrates, the MAPK-Activated Protein Kinases,” *Microbiol. Mol. Biol. Rev.*, vol. 75, no. 1, pp. 50–83, Mar. 2011, doi: 10.1128/MMBR.00031-10.

[124] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, 2012.

[125] J. Chen *et al.*, “LncSEA: A platform for long non-coding RNA related sets and enrichment analysis,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D969–D980, Jan. 2021, doi: 10.1093/nar/gkaa806.

[126] Y. Li *et al.*, “LncMAP: Pan-cancer Atlas of long noncoding RNA-mediated transcriptional network perturbations,” *Nucleic Acids Res.*, vol. 46, no. 3, pp. 1113–1123, Feb. 2018, doi: 10.1093/nar/gkx1311.

[127] T. Cui *et al.*, “MNDR v2.0: An updated resource of ncRNA-disease associations in mammals,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D371–D374, Jan. 2018, doi: 10.1093/nar/gkx1025.

[128] G. Chen *et al.*, “LncRNADisease: A database for long-non-coding RNA-associated diseases,” *Nucleic Acids Res.*, vol. 41, no. D1, pp. D983–D986, Jan. 2013, doi: 10.1093/nar/gks1099.

[129] S. Ning *et al.*, “Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D980–D985, Jan. 2016, doi: 10.1093/nar/gkv1094.

[130] B. Zhou *et al.*, “EVLncRNAs: A manually curated database for long non-coding RNAs validated by low-throughput experiments,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D100–D105, Jan. 2018, doi: 10.1093/nar/gkx677.

[131] J. Wang, X. Zhang, W. Chen, J. Li, and C. Liu, “CRlncRNA: A manually curated database of cancer-related long non-coding RNAs with experimental proof of functions on clinicopathological and molecular features,” *BMC Med. Genomics*, vol. 11, no. 6, pp. 29–37, Dec. 2018, doi: 10.1186/s12920-018-0430-2.

APPENDIX

8.7 Appendix A

8.7.1 Appendix A.1: 132 Glycome genes in 11 molecular categories

1. **Adhesion Molecule:** *EMCN, PODXL2*
2. **CBP:C-Type Lectin:** *ASGR2, CD207, CD209, CLEC10A, CLEC11A, CLEC12A, CLEC14A, CLEC1A, CLEC1B, CLEC2L, CLEC3A, CLEC4C, CLEC4G, CLEC4M, CLEC5A, MBL2, MRC2, PKD1L2, SFTPA1, THBD*
3. **CBP:I-Type Lectin:** *CD22, ICAM1, MAG, PECAM1, SIGLEC1, SIGLEC6, VCAM1*
4. **Galectin:** *LGALS13, LGALS3, LGALS3BP*
5. **Glycan Degradation**: *ARSD, ARSE, ARSF, ASAH2, GALC, GALNS, GLA, GNS, HEXA, HYAL3, MAN1C1, MAN2A1, MAN2B1, NAGA, NEU1, NEU2, SULF1*
6. **Glycosyltransferases:** *ABO, ALG10B, ALG5, ALG6, ALG9, B3GALT1, B3GALT4, B3GNT2, B3GNT3, B3GNT4, B3GNT8, B4GALT1, B4GALT3, B4GALT5, CHST12, CHST14, CHST3, CHSY3, CSGALNACT2, DPAGT1, DPM3, DSEL, EXT1, EXTL1, EXTL2, FUT11, FUT2, FUT5, FUT8, GALNT12, GALNT14, GALNT2, GALNT3, GALNT7, GALNT8, GALNTL5, GCNT4, GLCE, HS3ST3B1, HS3ST6, HS6ST2, LARGE, LFNG, MGAT2, NDST3, PIGH, PIGQ, ST3GAL1, ST6GALNAC5, ST8SIA1, ST8SIA3, ST8SIA6, WBSCR17, XYLT1*
7. **Glycoproteins:** *CD164, EMR1, MUC6, UMOD*
8. **Intracellular protein transport:** *COG1*
9. **Nucleotide Sugar Transporters:** *CMAS, GALT, HK1, MPI, PAPSS1, PGM1, PMM1, SLC35B1, SLC35B3, SLC35B4, SLC35D2, SLC35D3, SLC35E4, SLC35F3, UGP2*
10. **Proteoglycans:** *CD44, GPC3, PTPRZ1, SDC4, SMC3, SPOCK3, SRGN*
11. **Sulfotransferases:** *SULT1A2, SULT1A3*

8.7.2 Appendix A.2

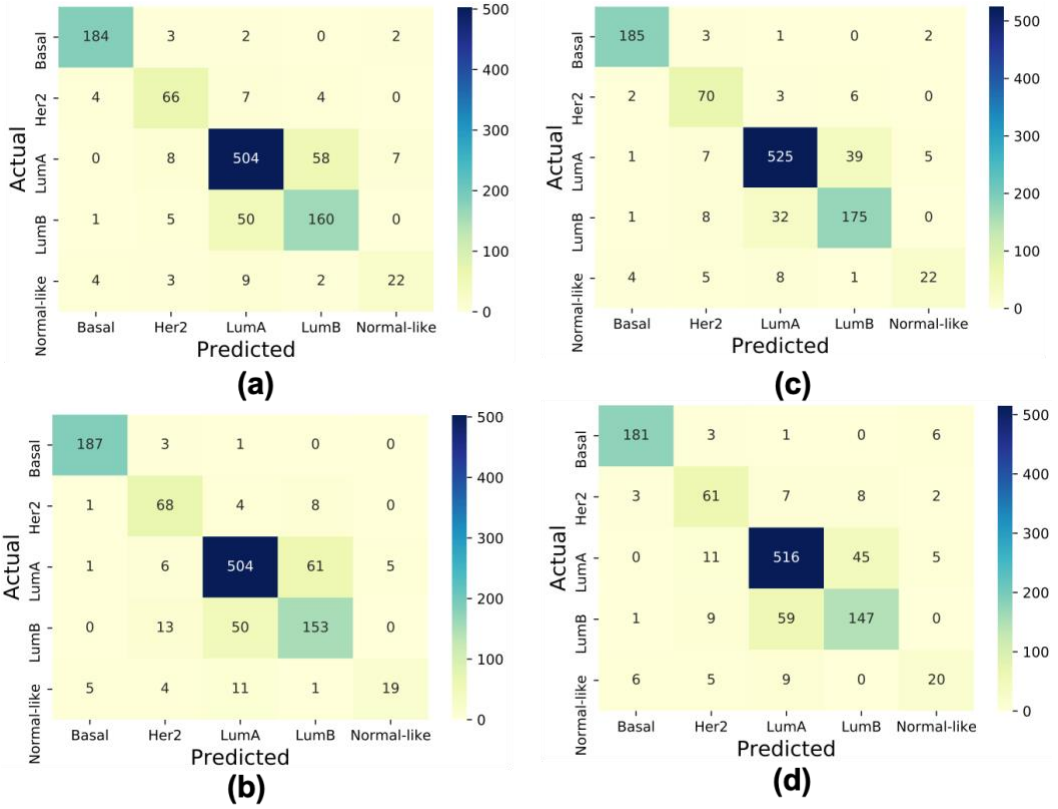

Figure S1: Confusion matrix derived from classification using the expression profiles of 196 lncRNAs discovered by (a) L1MSVM, (b) RF (c), and RL1MSVM, and (d) 91 key lncRNAs.

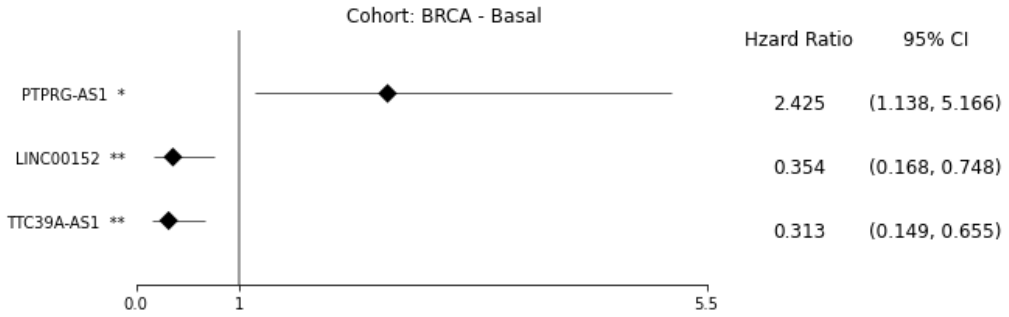

Figure S2: Forest plot of survival analysis for 3 prognostic lncRNAs on BRCA - Basal cohort. The asterisks represent the Log-rank P-values (*p ≤ 0.05, **p ≤ 0.01).

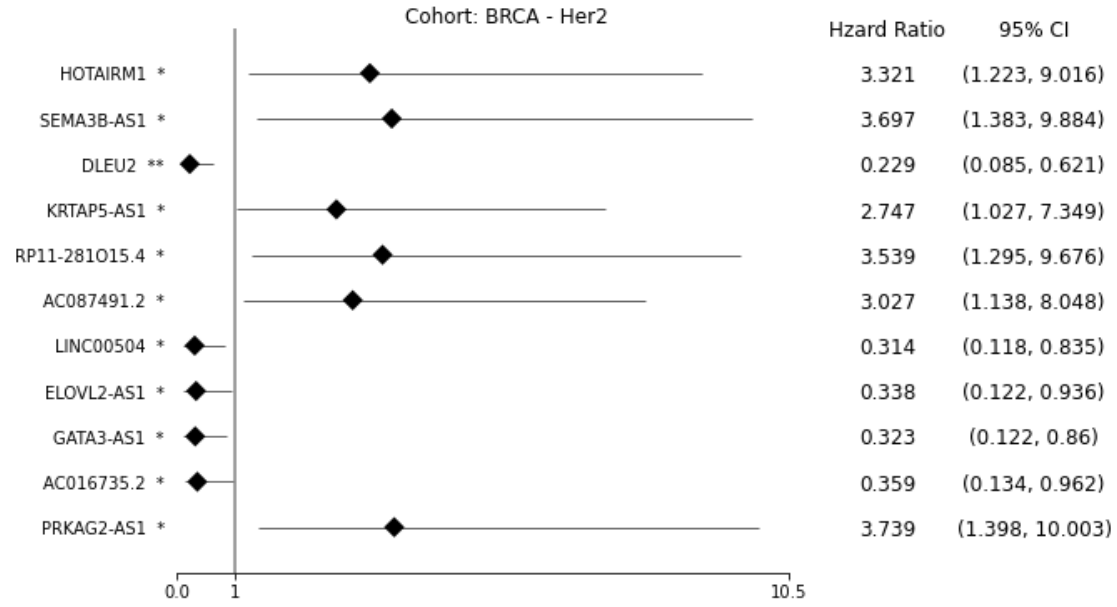

Figure S3: Forest plot of survival analysis for 11 prognostic lncRNAs on BRCA – HER2 cohort. The asterisks represent the Log-rank P-values (*p ≤ 0.05, **p ≤ 0.01).

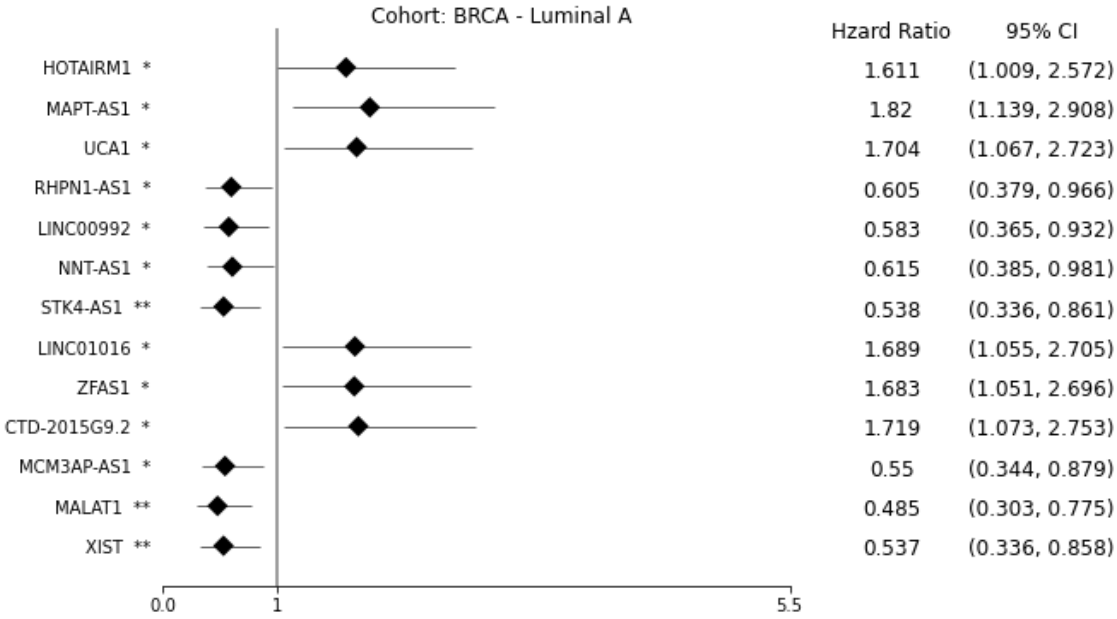

Figure S4: Forest plot of survival analysis for 13 prognostic lncRNAs on BRCA – Luminal A cohort. The asterisks represent the Log-rank P-values (*p ≤ 0.05, **p ≤ 0.01).

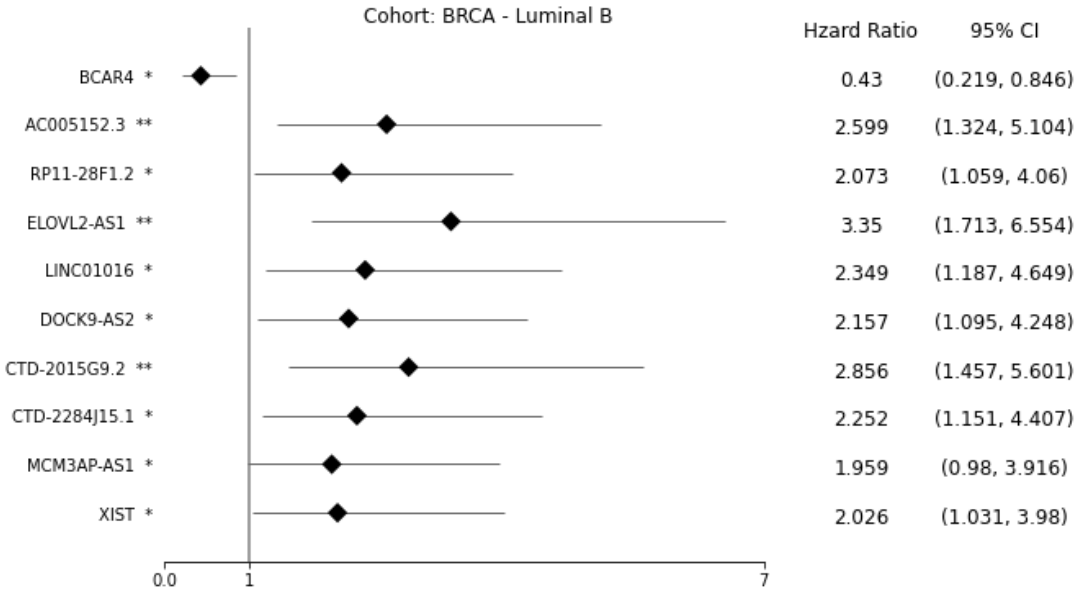

Figure S5: Forest plot of survival analysis for 10 prognostic lncRNAs on BRCA – Luminal B cohort. The asterisks represent the Log-rank P-values (*p ≤ 0.05, **p ≤ 0.01).

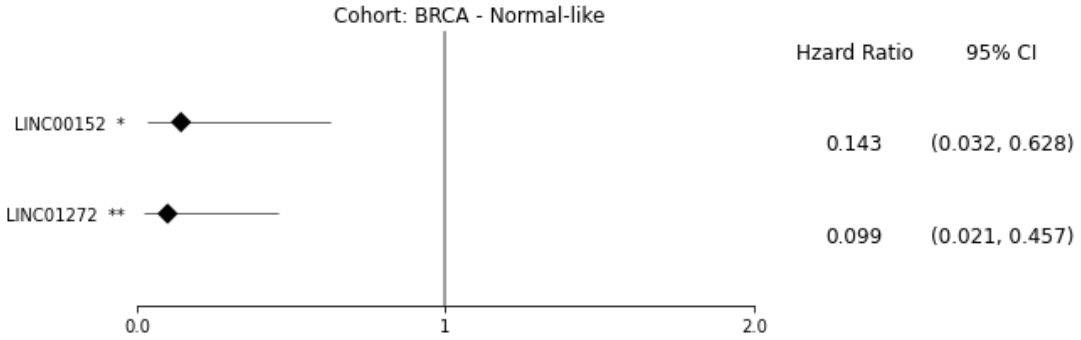

Figure S6: Forest plot of survival analysis for 2 prognostic lncRNAs on BRCA – Normal-like cohort. The asterisks represent the Log-rank P-values (*p ≤ 0.05, **p ≤ 0.01).

Table S1: List of subtype-specific lncRNAs (*n* = 239).

lncRNA	Chrom	Start	End	Strand	Subtype
AC008268.1	chr2	95666084	95668715	+	Basal
AC016735.2	chr2	43027853	43039547	-	Basal
AC144450.1	chr2	1546665	1620113	-	Basal
AFAP1-AS1	chr4	7754090	7778928	+	Basal
CTD-2015G9.2	chr16	86722091	86741059	+	Basal
KB-1991G8.1	chr8	100337595	100350707	+	Basal
KIRREL3-AS1	chr11	126543947	126610948	+	Basal
LINC00504	chr4	14470465	14888169	-	Basal
LINC00511	chr17	72323123	72640472	-	Basal
LINC00839	chr10	42475543	42495336	+	Basal
LINC00993	chr10	37309185	37347031	+	Basal
NFE4	chr7	102973522	102988856	+	Basal
RP11-10A14.5	chr8	9189011	9202854	+	Basal
RP11-19E11.1	chr2	118833700	118835110	-	Basal
RP11-206M11.7	chr3	149284782	149333653	+	Basal
RP11-226L15.5	chr1	160024953	160026794	-	Basal
RP11-281O15.4	chr5	178969390	178990116	+	Basal
RP11-321G12.1	chr15	63390136	63438320	+	Basal
RP11-363E7.4	chr9	19453209	19455173	+	Basal
RP11-378A13.1	chr2	218255319	218257366	+	Basal
RP11-395G23.3	chr8	106270144	106272899	+	Basal
RP11-597D13.9	chr4	158170752	158202877	+	Basal
RP11-616M22.7	chr16	1294551	1299166	-	Basal
RP11-672A2.4	chr11	76654169	76656712	-	Basal
RP11-834C11.4	chr12	54126098	54132843	+	Basal
TTC39A-AS1	chr1	51329654	51335324	+	Basal
WDFY3-AS2	chr4	84965682	85011277	+	Basal
AC005152.3	chr17	72021851	72034092	-	HER2
AC009948.5	chr2	178413939	178440243	+	HER2
AC010729.1	chr2	5696220	5708095	+	HER2
AC025016.1	chr11	5938751	5944984	+	HER2
AC087491.2	chr17	39619613	39622513	+	HER2
AC114730.3	chr2	241808312	241812016	-	HER2
AF127936.9	chr21	14818843	15014430	-	HER2
ASB16-AS1	chr17	44175973	44186717	-	HER2
BDNF-AS	chr11	27506838	27698174	+	HER2
CRNDE	chr16	54918863	54929189	-	HER2

CTB-33O18.1	chr5	173562478	173573199	+	HER2
CTC-537E7.2	chr5	68531690	68533530	-	HER2
CTD-2081C10.7	chr5	53880293	53881051	-	HER2
CTD-3157E16.1	chr17	15787787	15788205	-	HER2
GATA3-AS1	chr10	8050450	8053484	-	HER2
KRTAP5-AS1	chr11	1571353	1599184	+	HER2
LINC00883	chr3	107240692	107326964	+	HER2
LINC01105	chr2	5932687	6001275	+	HER2
PVT1	chr8	127794533	128101253	+	HER2
RP11-10L12.4	chr4	102828055	102844075	+	HER2
RP11-157J24.2	chr6	1528364	1528911	-	HER2
RP11-17M16.2	chr18	76491652	76493918	+	HER2
RP11-206M11.7	chr3	149284782	149333653	+	HER2
RP11-20F24.2	chr10	37240887	37242049	+	HER2
RP11-218M22.1	chr12	630891	663706	+	HER2
RP11-321G12.1	chr15	63390136	63438320	+	HER2
RP11-356O9.2	chr14	37556158	37567095	-	HER2
RP11-369C8.1	chr14	45706250	45715952	-	HER2
RP11-431J24.2	chrX	16152941	16170869	-	HER2
RP11-44N21.1	chr14	105093609	105099004	+	HER2
RP11-490M8.1	chr2	36354749	36355114	-	HER2
RP11-510J16.5	chr16	82044336	82139631	-	HER2
RP11-61L19.3	chr18	9519449	9520199	+	HER2
RP11-635N19.1	chr18	63367328	63381629	+	HER2
RP11-999E24.3	chr14	57993545	57994525	-	HER2
RP3-443C4.2	chr6	151813276	151814179	-	HER2
RP5-1121A15.3	chr7	156944721	156945645	+	HER2
SMG7-AS1	chr1	183460874	183472265	-	HER2
ST8SIA6-AS1	chr10	17386936	17413503	+	HER2
TRIM52-AS1	chr5	181261212	181272307	+	HER2
VPS9D1-AS1	chr16	89711856	89718165	+	HER2
XXyac-YX65C7_A.2	chr6	169213254	169239565	+	HER2
ZNF667-AS1	chr19	56477250	56500666	+	HER2
AC005624.2	chr19	2458935	2462185	-	Luminal A
AC006273.5	chr19	782755	785080	+	Luminal A
AC008268.1	chr2	95666084	95668715	+	Luminal A
AC010729.1	chr2	5696220	5708095	+	Luminal A
AC114730.3	chr2	241808312	241812016	-	Luminal A
AP000251.3	chr21	31559245	31560487	+	Luminal A

CTA-126B4.7	chr22	42438023	42446195	+	Luminal A
CTA-984G1.5	chr22	29260889	29262037	+	Luminal A
CTB-51J22.1	chr7	74059576	74062284	-	Luminal A
CTC-548K16.1	chr19	14333743	14343916	+	Luminal A
CTC-558O2.1	chr5	168706567	168720884	+	Luminal A
CTD-2081C10.7	chr5	53880293	53881051	-	Luminal A
CTD-2263F21.1	chr5	38460925	38468339	-	Luminal A
CTD-2291D10.4	chr19	23075201	23100361	+	Luminal A
CTD-2319I12.4	chr17	60126535	60135644	-	Luminal A
CTD-2514K5.4	chr17	74256896	74262020	-	Luminal A
CTD-2523D13.2	chr11	119729583	119739623	+	Luminal A
CTD-3032H12.1	chr16	54937786	54938671	-	Luminal A
ELOVL2-AS1	chr6	11043524	11078226	+	Luminal A
HOTAIR	chr12	53962308	53974956	-	Luminal A
KRTAP5-AS1	chr11	1571353	1599184	+	Luminal A
LINC00337	chr1	6236240	6239444	+	Luminal A
LINC00471	chr2	231508426	231514339	-	Luminal A
LINC00885	chr3	196142636	196160890	+	Luminal A
LINC00922	chr16	65284499	65576300	-	Luminal A
LINC00993	chr10	37309185	37347031	+	Luminal A
LINC01272	chr20	50267486	50279795	+	Luminal A
MAPT-AS1	chr17	45843651	45895600	-	Luminal A
NCK1-AS1	chr3	136841726	136862054	-	Luminal A
PARD3-AS1	chr10	34815767	34816386	+	Luminal A
PRKAG2-AS1	chr7	151877042	151879223	+	Luminal A
RAMP2-AS1	chr17	42753914	42761257	-	Luminal A
RERG-IT1	chr12	15112363	15114698	-	Luminal A
RP1-102D24.5	chr22	45435864	45448743	-	Luminal A
RP11-100E13.1	chr1	224615296	224616220	-	Luminal A
RP11-105N14.1	chr2	213152970	213153659	+	Luminal A
RP11-108L7.15	chr10	101060029	101061005	-	Luminal A
RP11-108M9.4	chr1	16888538	16889649	-	Luminal A
RP11-120K24.3	chr13	112964835	112966131	-	Luminal A
RP11-180M15.3	chr12	12668982	12685075	+	Luminal A
RP11-206M11.7	chr3	149284782	149333653	+	Luminal A
RP11-21L23.2	chr11	76800364	76804555	+	Luminal A
RP11-22C11.2	chr8	94637285	94639467	-	Luminal A
RP11-244M2.1	chr18	39206924	39800318	-	Luminal A
RP11-251M1.1	chr9	136648610	136660421	-	Luminal A

RP11-28F1.2	chr18	63313802	63314376	-	Luminal A
RP11-299G20.2	chr15	101295419	101305737	+	Luminal A
RP11-303E16.5	chr16	81055301	81056426	+	Luminal A
RP11-314C16.1	chr6	8784178	8785445	+	Luminal A
RP11-321G12.1	chr15	63390136	63438320	+	Luminal A
RP11-349I1.2	chr14	94430633	94464730	+	Luminal A
RP11-356O9.1	chr14	37564047	37579125	+	Luminal A
RP11-431J24.2	chrX	16152941	16170869	-	Luminal A
RP11-473M20.11	chr16	3106764	3109576	+	Luminal A
RP11-506M12.1	chr7	100115214	100127139	-	Luminal A
RP11-527H14.4	chr18	14903580	14915628	+	Luminal A
RP11-53O19.1	chr5	44744900	44808777	-	Luminal A
RP11-585P4.5	chr12	75483454	75489820	-	Luminal A
RP11-672A2.4	chr11	76654169	76656712	-	Luminal A
RP11-680B3.2	chr3	148850933	148960112	-	Luminal A
RP11-733C7.1	chr4	138277115	138281784	-	Luminal A
RP11-758P17.3	chr7	100436204	100438504	+	Luminal A
RP11-773H22.4	chr18	12984694	12991173	-	Luminal A
RP11-774O3.3	chr4	8355090	8358338	-	Luminal A
RP11-800A3.7	chr11	73307235	73309361	-	Luminal A
RP11-80H18.4	chr3	58329965	58330118	+	Luminal A
RP11-932O9.10	chr15	30648797	30649529	+	Luminal A
RP13-638C3.3	chr17	82587313	82588411	-	Luminal A
RP13-977J11.2	chr12	132186735	132189695	-	Luminal A
RP3-337H4.8	chr6	43588230	43591362	-	Luminal A
RP3-395M20.8	chr1	2549920	2557031	-	Luminal A
RP4-533D7.5	chr1	46134531	46139081	+	Luminal A
RP4-564M11.2	chr1	77067920	77078482	+	Luminal A
RP5-1061H20.4	chr1	229258281	229271028	-	Luminal A
RP5-821D11.7	chr22	41831215	41834665	-	Luminal A
RP5-965F6.2	chr1	112177234	112360528	-	Luminal A
SEMA3B-AS1	chr3	50266641	50267371	-	Luminal A
SNHG3	chr1	28505980	28510892	+	Luminal A
SYNPR-AS1	chr3	63423596	63550051	-	Luminal A
TMPO-AS1	chr12	98512973	98516422	-	Luminal A
TUG1	chr22	30970677	30979395	+	Luminal A
XXbac-BPG13B8.10	chr6	29497509	29510556	+	Luminal A
AC005152.3	chr17	72021851	72034092	-	Luminal B
AC006126.4	chr19	45238632	45245370	-	Luminal B

AC006273.5	chr19	782755	785080	+	Luminal B
AC016735.2	chr2	43027853	43039547	-	Luminal B
AC061961.2	chr2	154696462	154697817	-	Luminal B
AC087491.2	chr17	39619613	39622513	+	Luminal B
AC093620.5	chr7	5419827	5420767	+	Luminal B
AC093673.5	chr7	143379692	143380495	-	Luminal B
AF015262.2	chr21	35136638	35139222	+	Luminal B
AF131215.9	chr8	11104691	11106704	-	Luminal B
AP000473.5	chr21	16630827	16640683	+	Luminal B
CTA-221G9.12	chr22	25102433	25112692	-	Luminal B
CTD-2081C10.7	chr5	53880293	53881051	-	Luminal B
CTD-2134A5.3	chr14	103875055	103877478	+	Luminal B
CTD-2201E18.3	chr5	43014414	43067419	-	Luminal B
CTD-2231H16.1	chr5	92151	139863	+	Luminal B
CTD-2291D10.4	chr19	23075201	23100361	+	Luminal B
CTD-2336O2.1	chr8	1761990	1764502	-	Luminal B
CTD-2639E6.9	chr19	48963975	48965158	+	Luminal B
DLEU2	chr13	49982552	50125720	-	Luminal B
DOCK9-AS2	chr13	99087819	99088625	+	Luminal B
DSCAM-AS1	chr21	40383083	40385358	+	Luminal B
HOTAIR	chr12	53962308	53974956	-	Luminal B
KB-1440D3.13	chr22	21661934	21662363	+	Luminal B
KCNK15-AS1	chr20	44694892	44746021	-	Luminal B
LINC00467	chr1	211382803	211435333	+	Luminal B
LINC00992	chr5	117415509	117546298	+	Luminal B
MIR205HG	chr1	209428820	209432838	+	Luminal B
MIR99AHG	chr21	16070522	16627397	+	Luminal B
PIK3CD-AS2	chr1	9672426	9687555	-	Luminal B
PVT1	chr8	127794533	128101253	+	Luminal B
RARA-AS1	chr17	40340867	40343136	-	Luminal B
RP1-102D24.5	chr22	45435864	45448743	-	Luminal B
RP1-232P20.1	chr6	5451683	5458075	-	Luminal B
RP11-108M9.4	chr1	16888538	16889649	-	Luminal B
RP11-120K24.3	chr13	112964835	112966131	-	Luminal B
RP11-127O4.3	chr10	104474939	104480274	-	Luminal B
RP11-181G12.2	chr1	2181794	2184389	-	Luminal B
RP11-21L23.2	chr11	76800364	76804555	+	Luminal B
RP11-22C11.2	chr8	94637285	94639467	-	Luminal B
RP11-244M2.1	chr18	39206924	39800318	-	Luminal B

RP11-251M1.1	chr9	136648610	136660421	-	Luminal B
RP11-279F6.1	chr15	69463026	69571440	+	Luminal B
RP11-283G6.4	chr12	26211164	26335856	-	Luminal B
RP11-314C16.1	chr6	8784178	8785445	+	Luminal B
RP11-356O9.1	chr14	37564047	37579125	+	Luminal B
RP11-356O9.2	chr14	37556158	37567095	-	Luminal B
RP11-395N3.2	chr2	226800146	226811029	+	Luminal B
RP11-401P9.4	chr16	50645809	50649249	+	Luminal B
RP11-429J17.7	chr8	143758153	143771822	-	Luminal B
RP11-459E5.1	chr8	22690150	22798616	+	Luminal B
RP11-47A8.5	chr10	102642792	102644140	-	Luminal B
RP11-485G7.5	chr16	11341809	11345211	-	Luminal B
RP11-506M12.1	chr7	100115214	100127139	-	Luminal B
RP11-507K2.3	chr14	88551597	88552493	+	Luminal B
RP11-53O19.1	chr5	44744900	44808777	-	Luminal B
RP11-613M10.6	chr9	37509150	37510299	+	Luminal B
RP11-629O1.2	chr8	133573183	133573861	+	Luminal B
RP11-635N19.1	chr18	63367328	63381629	+	Luminal B
RP11-680B3.2	chr3	148850933	148960112	-	Luminal B
RP11-773H22.4	chr18	12984694	12991173	-	Luminal B
RP11-774O3.3	chr4	8355090	8358338	-	Luminal B
RP11-92K15.3	chr8	80032724	80033300	-	Luminal B
RP3-416H24.1	chr12	52245048	52247448	-	Luminal B
RP3-443C4.2	chr6	151813276	151814179	-	Luminal B
RP5-1061H20.4	chr1	229258281	229271028	-	Luminal B
SYN2	chr3	12004402	12191400	+	Luminal B
TINCR	chr19	5558167	5568034	-	Luminal B
TMPO-AS1	chr12	98512973	98516422	-	Luminal B
TPTEP1	chr22	16601887	16698742	+	Luminal B
TUSC8	chr13	44400250	44405984	-	Luminal B
ZNF667-AS1	chr19	56477250	56500666	+	Luminal B
AC093850.2	chr2	215718043	215719424	+	Normal-like
DYNLL1-AS1	chr12	120490328	120495940	-	Normal-like
FAM83H-AS1	chr8	143734140	143746337	+	Normal-like
FOXD3-AS1	chr1	63320884	63324441	-	Normal-like
LHX4-AS1	chr1	180269653	180502954	-	Normal-like
LINC00086	chrX	135421943	135428074	+	Normal-like
LINC00467	chr1	211382803	211435333	+	Normal-like
LINC01087	chr2	131637025	131649615	+	Normal-like

MIR205HG	chr1	209428820	209432838	+	Normal-like
NNT-AS1	chr5	43571594	43603230	-	Normal-like
RP11-218M22.1	chr12	630891	663706	+	Normal-like
RP11-394O4.5	chr5	149430286	149432834	+	Normal-like
RP11-428L9.2	chr10	8970125	8973468	+	Normal-like
RP11-65J21.3	chr16	14302288	14326353	+	Normal-like
RP11-738B7.1	chr7	129783370	129785185	-	Normal-like

Table S2: List of subtype-specific key lncRNAs (*n* = 91)

LncRNA	**Sub-type**	**Chrom**	**Start**	**End**
AC005152.3	Basal	chr17	72021851	72034092
AC087491.2	Basal	chr17	39619613	39622513
AC144450.1	Basal	chr2	1546665	1620113
CCAT1	Basal	chr8	127207866	127219088
CTD-3032H12.1	Basal	chr16	54937786	54938671
DANCR	Basal	chr4	52712404	52720351
HIF1A-AS2	Basal	chr14	61715558	61751097
HOTAIRM1	Basal	chr7	27095647	27100265
KCNK15-AS1	Basal	chr20	44694892	44746021
LINC00152	Basal	chr2	87455368	87606805
LINC00993	Basal	chr10	37309185	37347031
MALAT1	Basal	chr11	65497762	65506516
RMST	Basal	chr12	97431653	97565035
RP11-279F6.1	Basal	chr15	69463026	69571440
RP11-281O15.4	Basal	chr5	178969390	178990116
SNHG14	Basal	chr15	24978583	25419462
TTC39A-AS1	Basal	chr1	51329654	51335324
XIST	Basal	chrX	73820651	73852753
ZNF667-AS1	Basal	chr19	56477250	56500666
CTB-33O18.1	Her2	chr5	173562478	173573199
CTD-2284J15.1	Her2	chr8	86333274	86343314
ELOVL2-AS1	Her2	chr6	11043524	11078226
FGF14-AS2	Her2	chr13	102394630	102395703
GATA3-AS1	Her2	chr10	8050450	8053484
KIRREL3-AS1	Her2	chr11	126543947	126610948
LINC00839	Her2	chr10	42475543	42495336
LINC01016	Her2	chr6	33867506	33896914
MIR31HG	Her2	chr9	21455642	21559669
RP1-232P20.1	Her2	chr6	5451683	5458075
RP11-20F24.2	Her2	chr10	37240887	37242049
RP11-28F1.2	Her2	chr18	63313802	63314376
TINCR	Her2	chr19	5558167	5568034
VPS9D1-AS1	Her2	chr16	89711856	89718165

BCAR4	Luminal A	chr16	11819829	11828845
BCYRN1	Luminal A	chr2	47331060	47344517
CTD-2015G9.2	Luminal A	chr16	86722091	86741059
CTD-2081C10.7	Luminal A	chr5	53880293	53881051
H19	Luminal A	chr11	1995163	2001470
HOXA-AS2	Luminal A	chr7	27107777	27134302
LINC00324	Luminal A	chr17	8220642	8224043
LINC00472	Luminal A	chr6	71344344	71420769
LINC00511	Luminal A	chr17	72323123	72640472
LINC00922	Luminal A	chr16	65284499	65576300
LINC01272	Luminal A	chr20	50267486	50279795
MAPT-AS1	Luminal A	chr17	45843651	45895600
PTPRG-AS1	Luminal A	chr3	62260865	62369330
RP11-21L23.2	Luminal A	chr11	76800364	76804555
ST8SIA6-AS1	Luminal A	chr10	17386936	17413503
TMPO-AS1	Luminal A	chr12	98512973	98516422
KRTAP5-AS1	Luminal A	chr11	1571353	1599184
NCK1-AS1	Luminal A	chr3	136841726	136862054
PARD3-AS1	Luminal A	chr10	34815767	34816386
PRKAG2-AS1	Luminal A	chr7	151877042	151879223
RAMP2-AS1	Luminal A	chr17	42753914	42761257
RERG-IT1	Luminal A	chr12	15112363	15114698
SEMA3B-AS1	Luminal A	chr3	50266641	50267371
SNHG3	Luminal A	chr1	28505980	28510892
HOTAIR	Luminal A, B	chr12	53962308	53974956
STK4-AS1	Luminal A, B	chr20	44963794	44966402
MEG3	Luminal A, Basal	chr14	100779410	100861031
AC008268.1	Luminal B	chr2	95666084	95668715
AC016735.2	Luminal B	chr2	43027853	43039547
AP000439.3	Luminal B	chr11	69477133	69479940
DLEU7-AS1	Luminal B	chr13	50807856	50849905
DSCAM-AS1	Luminal B	chr21	40383083	40385358
GAS5	Luminal B	chr1	173863900	173868882
HAGLR	Luminal B	chr2	176173195	176188958
LINC00992	Luminal B	chr5	117415509	117546298
MIR99AHG	Luminal B	chr21	16070522	16627397
PVT1	Luminal B	chr8	127794533	128101253
SNHG17	Luminal B	chr20	38420588	38435353
SYN2	Luminal B	chr3	12004402	12191400
ZFAS1	Luminal B	chr20	49278178	49295738
DOCK9-AS2	Luminal B	chr13	99087819	99088625
PIK3CD-AS2	Luminal B	chr1	9672426	9687555
RARA-AS1	Luminal B	chr17	40340867	40343136
TPTEP1	Luminal B	chr22	16601887	16698742
TUSC8	Luminal B	chr13	44400250	44405984

AFAP1-AS1	Normal-like	chr4	7754090	7778928
CTB-51J22.1	Normal-like	chr7	74059576	74062284
DLEU2	Normal-like	chr13	49982552	50125720
LINC00087	Normal-like	chrX	135095028	135098634
LINC00504	Normal-like	chr4	14470465	14888169
MCM3AP-AS1	Normal-like	chr21	46229217	46259390
MIR205HG	Normal-like	chr1	209428820	209432838
PCAT6	Normal-like	chr1	202810954	202812156
RHPN1-AS1	Normal-like	chr8	143366631	143368548
UCA1	Normal-like	chr19	15828961	15836320
LHX4-AS1	Normal-like	chr1	180269653	180502954
DYNLL1-AS1	Normal-like	chr12	120490328	120495940
NNT-AS1	Normal-like	chr5	43571594	43603230

Table S3: List of known lncRNAs associated with any kind of diseases (*n* = 53)

lncRNA	**Disease Name**	**PubMed ID**
AC008268.1	breast cancer	26910840
AC144450.1	astrocytoma	26252651
AC144450.1	Nasopharyngeal carcinoma	24379026
AFAP1-AS1	cancer	27471399
AFAP1-AS1	cholangiocarcinoma	28535506
AFAP1-AS1	colorectal cancer	27261589
AFAP1-AS1	esophagus adenocarcinoma	23333711
AFAP1-AS1	esophagus squamous cell carcinoma	26756568
AFAP1-AS1	gallbladder cancer	27810781
AFAP1-AS1	hepatocellular carcinoma	26803513
AFAP1-AS1	lung adenocarcinoma	27797003
AFAP1-AS1	lung cancer	26245991
AFAP1-AS1	Nasopharyngeal carcinoma	26246469
AFAP1-AS1	non-small cell lung carcinoma	26463625
AFAP1-AS1	ovarian cancer	28051261
AFAP1-AS1	pancreatic cancer	25910082
AFAP1-AS1	pancreatic ductal adenocarcinoma	25925763
AFAP1-AS1	stomach cancer	28451917
AFAP1-AS1	thyroid cancer	29331858
AFAP1-AS1	tongue squamous cell carcinoma	29310682
BCAR4	breast cancer	16778085
BCAR4	cervical cancer	28112728
BCAR4	chondrosarcoma	28399646
BCAR4	colon cancer	29190958

BCAR4	colorectal cancer	27197301
BCAR4	non-small cell lung carcinoma	28077810
BCAR4	osteosarcoma	27460090
BCAR4	stomach cancer	29028095
BCYRN1	breast cancer	9422992
BCYRN1	Aging	17553964
BCYRN1	Alzheimer's disease	1603265
BCYRN1	astrocytoma	25561975
BCYRN1	cancer	28651607
BCYRN1	cervical cancer	9422992
BCYRN1	colon cancer	29625226
BCYRN1	esophageal cancer	9422992
BCYRN1	Glioblastoma	25561975
BCYRN1	lung cancer	9422992
BCYRN1	malignant glioma	25561975
BCYRN1	microinvasive gastric cancer	29039538
BCYRN1	non-small cell lung carcinoma	25866480
BCYRN1	ovarian cancer	9422992
BCYRN1	parotid gland cancer	9422992
BCYRN1	squamous cell carcinoma	27143917
BCYRN1	tongue cancer	9422992
BCYRN1	asthma	28960519
CCAT1	breast cancer	26464701
CCAT1	acute myeloid leukemia	26923190
CCAT1	cancer	24594601
CCAT1	cervical cancer	28849215
CCAT1	colon cancer	29190961
CCAT1	colorectal cancer	23416875
CCAT1	endometrial cancer	27432114
CCAT1	esophageal squamous cell carcinoma	27956498
CCAT1	gallbladder cancer	25569100
CCAT1	Glioma	28475287
CCAT1	hepatocellular carcinoma	25884472
CCAT1	intrahepatic cholangiocarcinoma	28921383
CCAT1	laryngeal squamous cell carcinoma	28631575
CCAT1	lung adenocarcinoma	27566568
CCAT1	lung cancer	27212446
CCAT1	lung squamous cell carcinoma	28076325
CCAT1	malignant glioma	27765628

CCAT1	medulloblastoma	28777430
CCAT1	multiple myeloma	29228867
CCAT1	Nasopharyngeal carcinoma	28358263
CCAT1	non-small cell lung carcinoma	25129441
CCAT1	oral squamous cell carcinoma	28413645
CCAT1	osteosarcoma	28549102
CCAT1	ovarian cancer	24379988
CCAT1	Ovarian epithelial cancer	28754469
CCAT1	pancreatic cancer	28078015
CCAT1	renal cell carcinoma	28470345
CCAT1	retinoblastoma	28088735
CCAT1	stomach cancer	28535628
CCAT1	Neuralgia	29163801
DANCR	breast cancer	27716745
DANCR	astrocytoma	26252651
DANCR	bone disease	23438432
DANCR	brain cancer	29476310
DANCR	colorectal cancer	26617879
DANCR	hepatocellular carcinoma	25964079
DANCR	non-small cell lung carcinoma	29635134
DANCR	Osteoporosis, Postmenopausal	25660720
DANCR	osteosarcoma	26986815
DANCR	prostate cancer	23728290
DANCR	renal cell carcinoma	28765964
DANCR	stomach cancer	28618943
DANCR	Triple Negative Breast Neoplasms	28760736
DLEU2	astrocytoma	26252651
DLEU2	chronic lymphocytic leukemia	9395242
DLEU2	Laryngeal Neoplasms	29687850
DLEU2	Leukemia, Lymphoid	18562676
DLEU2	lymphoma	11072235
DLEU2	myeloid neoplasm	19591824
DLEU2	pancreatic cancer	26045769
DLEU7-AS1	colorectal cancer	29364477
DSCAM-AS1	breast cancer	12177779
DSCAM-AS1	idiopathic scoliosis	21216876
FGF14-AS2	breast cancer	26820525
GAS5	breast cancer	18836484
GAS5	astrocytoma	26252651

GAS5	B-cell lymphoma	24583225
GAS5	bladder carcinoma	29445179
GAS5	bladder transitional cell carcinomas	27878359
GAS5	bladder urothelial carcinoma	28060759
GAS5	cancer	22996375
GAS5	cervical cancer	22487937
GAS5	colorectal cancer	24926850
GAS5	coronary artery disease	29267258
GAS5	endometrial carcinoma	26511107
GAS5	esophageal cancer	29386089
GAS5	esophageal squamous cell carcinoma	29170131
GAS5	Glioblastoma	23726844
GAS5	Glioma	28666797
GAS5	head and neck cancer	26482616
GAS5	hepatocellular carcinoma	25120813
GAS5	hypersensitivity reaction type II disease	20124551
GAS5	hypertension	27432865
GAS5	inflammatory bowel disease	28722800
GAS5	kidney cancer	24373479
GAS5	leukemia	20421347
GAS5	liver cirrhosis	26446789
GAS5	LPS-induced inflammatory injury	29448248
GAS5	lung adenocarcinoma	25925741
GAS5	lung cancer	26634743
GAS5	lymphoma	18406879
GAS5	malignant glioma	26370254
GAS5	malignant pleural mesothelioma	24885398
GAS5	mantle cell lymphoma	24703244
GAS5	melanoma	18836484
GAS5	multiple myeloma	24583225
GAS5	Nasopharyngeal carcinoma	28977945
GAS5	neuroblastoma	28035057
GAS5	non-small cell lung carcinoma	24357161
GAS5	osteoarthritis	25196583
GAS5	osteosarcoma	28519068
GAS5	ovarian cancer	26503132
GAS5	pancreatic cancer	24026436
GAS5	polycystic ovary syndrome	29648472
GAS5	Prostate	24373479

GAS5	prostate cancer	18836484
GAS5	renal cell carcinoma	23621190
GAS5	stomach cancer	24884417
GAS5	T-cell leukemia	18354083
GAS5	thyroid cancer	28506768
GAS5	Thyroid cancer, papillary	29423063
GAS5	type 2 diabetes mellitus	26675493
GAS5	urinary bladder cancer	24069260
GATA3-AS1	renal cell carcinoma	24905231
H19	breast adenocarcinoma	9811352
H19	abdominal aortic aneurysm	29669788
H19	adenocarcinoma	8785513
H19	adrenocortical carcinoma	22019903
H19	aortic valve disease	27789555
H19	astrocytoma	25561975
H19	atherosclerosis	21954592
H19	Beckwith-Wiedemann syndrome	7987305
H19	bladder carcinoma	7589512
H19	breast cancer	12419837
H19	cancer	15618002
H19	cardiac fibroblast proliferation and fibrosis	27318893
H19	cardiomyocyte hypertrophy	27084844
H19	central nervous system disease	20380817
H19	cervical cancer	8570220
H19	cholangiocarcinoma	27809873
H19	cholestatic liver injury	29425397
H19	choriocarcinoma	8564957
H19	chronic myeloid leukemia	24685695
H19	colon cancer	15521051
H19	colon carcinoma	21489289
H19	colorectal cancer	8564957
H19	Congenital Hyperinsulinism	11395395
H19	coronary artery disease	25772106
H19	Diabetic Cardiomyopathies	27796346
H19	embryonal carcinoma	26415227
H19	endometrial cancer	27775072
H19	endometriosis	26089099
H19	esophageal cancer	8564957
H19	gallbladder cancer	27073719

H19	gastric adenocarcinoma	29479897
H19	gastric cardia adenocarcinoma	24414129
H19	gastrointestinal system cancer	27738631
H19	germ cell cancer	16001432
H19	gestational choriocarcinoma	8188082
H19	gestational trophoblastic neoplasm	12648595
H19	Glioblastoma	16707459
H19	Glioma	27981546
H19	growth restriction	20104244
H19	head and neck squamous cell carcinoma	27994496
H19	Heart Defects, Congenital	27035723
H19	heart disease	27895893
H19	Hematopoiesis	15645136
H19	hepatocellular carcinoma	15736456
H19	Hydatidiform Mole	12783848
H19	hyperhomocysteinemia	15899898
H19	hyperprolactinemia	15525575
H19	infertility	20042264
H19	intestinal epithelial barrier function	26884465
H19	Keloid	27698867
H19	kidney cancer	24373479
H19	laryngeal squamous cell carcinoma	26872375
H19	liver cancer	11175353
H19	lung adenocarcinoma	25758555
H19	lung cancer	8564957
H19	malignant glioma	20380817
H19	Marek Disease	10696440
H19	medulloblastoma	8957451
H19	melanoma	11437411
H19	meningioma	10738131
H19	Mullerian aplasia	21458801
H19	multiple myeloma	29273733
H19	myeloproliferative neoplasm	12682647
H19	Nasopharyngeal carcinoma	27040767
H19	nephroblastoma	16179496
H19	Neural Tube Defects	22234160
H19	neuroblastoma	23791884
H19	non-small cell lung carcinoma	26482621
H19	obesity	22341586

H19	oral squamous cell carcinoma	28975993
H19	osteoarthritis	25430712
H19	osteosarcoma	24141783
H19	ovarian cancer	19656414
H19	Ovarian epithelial cancer	10428315
H19	pancreatic cancer	24920070
H19	pancreatic ductal adenocarcinoma	24920070
H19	papillary thyroid carcinoma	29287713
H19	Parkinson's disease	27021022
H19	Pheochromocytoma	21937622
H19	pituitary adenoma	23791884
H19	pneumoconiosis	27626436
H19	polycythemia vera	10640993
H19	Prader-Willi syndrome	23791884
H19	pre-eclampsia	19570415
H19	Prostate	24373479
H19	prostate cancer	24063685
H19	renal cell carcinoma	25866221
H19	rheumatoid arthritis	12937131
H19	Silver-Russell syndrome	19066168
H19	squamous cell carcinoma	22996375
H19	stomach cancer	9570380
H19	thyroid cancer	27093644
H19	trophoblastic neoplasm	8188082
H19	ulcerative colitis	27661667
H19	urinary bladder cancer	10413100
HAGLR	castration-resistant prostate cancer	28487115
HAGLR	cervical cancer	29228418
HAGLR	Glioma	29341117
HAGLR	hepatocellular carcinoma	28724429
HAGLR	neuroblastoma	24555823
HAGLR	non-small cell lung carcinoma	28443464
HAGLR	ovarian cancer	29416930
HAGLR	Ovarian epithelial cancer	29239819
HAGLR	stomach cancer	28475004
HAGLR	Thyroid cancer, papillary	28817151
HAGLR	urinary bladder cancer	27328915
HAGLR	stomach cancer	28475004
HIF1A-AS2	breast cancer	22664915

HIF1A-AS2	breast carcinoma	14580258
HIF1A-AS2	Glioblastoma	27264189
HIF1A-AS2	kidney cancer	9923855
HIF1A-AS2	osteosarcoma	23466354
HIF1A-AS2	stomach cancer	25686741
HIF1A-AS2	urinary bladder cancer	27018306
HOTAIR	breast cancer	19182780
HOTAIR	Abortion, Habitual	28750739
HOTAIR	acute myeloid leukemia	25979172
HOTAIR	Asthenozoospermia	26823733
HOTAIR	astrocytoma	25085602
HOTAIR	atypical teratoid rhabdoid tumor	25085602
HOTAIR	B-cell lymphoma	24583225
HOTAIR	bladder carcinoma	29673865
HOTAIR	bladder urothelial carcinoma	26781446
HOTAIR	cancer	29463216
HOTAIR	cerebrovascular disease	27613094
HOTAIR	cervical cancer	22487937
HOTAIR	chronic myeloid leukemia	27875938
HOTAIR	colon cancer	24667321
HOTAIR	colorectal cancer	21862635
HOTAIR	congestive heart failure	27317124
HOTAIR	cutaneous squamous cell carcinoma	27067026
HOTAIR	diffuse large B-cell lymphoma	27550047
HOTAIR	embryonal cancer	25085602
HOTAIR	endometrial cancer	24285342
HOTAIR	endometrial carcinoma	29466670
HOTAIR	Ependymoma	25085602
HOTAIR	esophageal cancer	28441714
HOTAIR	esophageal squamous cell carcinoma	27810266
HOTAIR	esophagus squamous cell carcinoma	24022190
HOTAIR	functionless pituitary adenoma	24469926
HOTAIR	gallbladder cancer	24953832
HOTAIR	gastric adenocarcinoma	23888369
HOTAIR	gastric cardia adenocarcinoma	25476857
HOTAIR	gastrointestinal stromal tumor	27659532
HOTAIR	gastrointestinal system cancer	24667321
HOTAIR	Glioblastoma	24203894
HOTAIR	Glioma	28083786

HOTAIR	head and neck squamous cell carcinoma	26592246
HOTAIR	heart disease	24788418
HOTAIR	hepatitis C	27129296
HOTAIR	hepatocellular carcinoma	21327457
HOTAIR	kidney cancer	24616104
HOTAIR	laryngeal squamous cell carcinoma	23141928
HOTAIR	Lemierre's syndrome	26806307
HOTAIR	leukemia	27748863
HOTAIR	Leukemia, Lymphoid	29513085
HOTAIR	liver cancer	24667321
HOTAIR	liver cirrhosis	27979710
HOTAIR	lung adenocarcinoma	24155936
HOTAIR	lung cancer	23668363
HOTAIR	lung small cell carcinoma	24591352
HOTAIR	malignant glioma	24203894
HOTAIR	medulloblastoma	25085602
HOTAIR	melanoma	23862139
HOTAIR	multiple myeloma	24583225
HOTAIR	Nasopharyngeal carcinoma	23281836
HOTAIR	neuroblastoma	29603181
HOTAIR	non-small cell lung carcinoma	23743197
HOTAIR	osteoarthritis	25430712
HOTAIR	osteosarcoma	25728753
HOTAIR	ovarian cancer	23600210
HOTAIR	Ovarian epithelial cancer	24662839
HOTAIR	pancreatic cancer	22614017
HOTAIR	pancreatic carcinoma	24667321
HOTAIR	pancreatic ductal adenocarcinoma	26482614
HOTAIR	papillary thyroid carcinoma	25997963
HOTAIR	Parkinson's disease	26979073
HOTAIR	pituitary adenoma	24469926
HOTAIR	pre-eclampsia	25807808
HOTAIR	prostate cancer	20864820
HOTAIR	renal carcinoma	25149152
HOTAIR	renal cell carcinoma	24935377
HOTAIR	retinoblastoma	27966488
HOTAIR	rheumatoid arthritis	24722995
HOTAIR	sarcoma	23543869
HOTAIR	solid tumors	27333150

HOTAIR	sporadic thoracic aortic aneurysm	28757056
HOTAIR	squamous cell carcinoma	23717443
HOTAIR	stomach cancer	23847441
HOTAIR	thyroid cancer	28565838
HOTAIR	triple-receptor negative breast cancer	25996380
HOTAIR	urinary bladder cancer	25030736
HOTAIRM1	acute myeloid leukemia	26436590
HOTAIRM1	acute promyelocytic leukemia	24824789
HOTAIRM1	astrocytoma	25561975
HOTAIRM1	colorectal cancer	27307307
HOTAIRM1	Glioblastoma	26111795
HOTAIRM1	leukemia	28180285
HOTAIRM1	malignant glioma	22709987
HOTAIRM1	melanoma	27016304
HOTAIRM1	pancreatic ductal adenocarcinoma	26676849
HOXA-AS2	breast cancer	28545023
HOXA-AS2	acute promyelocytic leukemia	23649634
HOXA-AS2	colorectal cancer	28112720
HOXA-AS2	gallbladder carcinoma	28388535
HOXA-AS2	hepatocellular carcinoma	27855366
HOXA-AS2	malignant glioma	29310118
HOXA-AS2	melanoma	27016304
HOXA-AS2	stomach cancer	26384350
KCNK15-AS1	breast cancer	25929808
KCNK15-AS1	osteoarthritis	25430712
KRTAP5-AS1	astrocytoma	26252651
KRTAP5-AS1	hepatocellular carcinoma	26492393
LHX4-AS1	astrocytoma	26252651
LINC00472	breast cancer	25865225
LINC00472	colorectal cancer	29488624
LINC00472	lung adenocarcinoma	27826625
LINC00472	ovarian cancer	27667152
LINC00511	breast cancer	26929647
LINC00511	lung adenocarcinoma	27797003
LINC00511	non-small cell lung carcinoma	27845772
LINC00993	breast cancer	25996380
LINC01016	breast cancer	26426411
MALAT1	breast cancer	18006640
MALAT1	acute monocytic leukemia	28713913

MALAT1	acute myeloid leukemia	28713913
MALAT1	amyotrophic lateral sclerosis	27338628
MALAT1	astrocytoma	26252651
MALAT1	B-cell lymphoma	21489289
MALAT1	bladder carcinoma	28648755
MALAT1	bladder urothelial carcinoma	23153939
MALAT1	calcific aortic valve disease	28522163
MALAT1	cancer	20711585
MALAT1	cervical cancer	20213048
MALAT1	cholangiocarcinoma	28592124
MALAT1	choriocarcinoma	29096355
MALAT1	colon cancer	21489289
MALAT1	colorectal cancer	21503572
MALAT1	Congenital Microtia	26282502
MALAT1	decreased myogenesis	23485710
MALAT1	diabetes mellitus	24436191
MALAT1	diabetes mellitus	26512840
MALAT1	Diabetic Cardiomyopathies	26476026
MALAT1	Diabetic Nephropathies	27964927
MALAT1	endometrial adenocarcinoma	25085246
MALAT1	endometrial stromal sarcoma	16441420
MALAT1	esophageal cancer	27470544
MALAT1	esophageal squamous cell carcinoma	27935117
MALAT1	fatty liver disease	26935028
MALAT1	Fibroma	27101025
MALAT1	fibrosarcoma	22491206
MALAT1	Flavivirus Infections	26634309
MALAT1	Follicular and H¬ürthle Cell Thyroid Neoplasm	28660408
MALAT1	gallbladder cancer	24658096
MALAT1	gastrointestinal system cancer	27313790
MALAT1	Glioblastoma	25772239
MALAT1	Glioma	27313790
MALAT1	hepatocellular carcinoma	16878148
MALAT1	high glucose-induced podocyte injury	28444861
MALAT1	histiocytoid hemangioma	27709553
MALAT1	HIV	26139386
MALAT1	Hyperglycemia	25787249
MALAT1	ischemic stroke	28093478
MALAT1	kidney cancer	24373479

MALAT1	Klatskin's tumor	28059437
MALAT1	laryngeal squamous cell carcinoma	24817925
MALAT1	liver cancer	21489289
MALAT1	liver cirrhosis	26697839
MALAT1	lung adenocarcinoma	19690017
MALAT1	lung cancer	17270048
MALAT1	lung small cell carcinoma	22928560
MALAT1	lymph node metastasis	26989678
MALAT1	malignant glioma	24926466
MALAT1	mantle cell lymphoma	27998273
MALAT1	melanoma	19625619
MALAT1	multiple myeloma	24583225
MALAT1	Nasopharyngeal carcinoma	23688988
MALAT1	neuroblastoma	20149803
MALAT1	non-small cell lung carcinoma	12970751
MALAT1	oral squamous cell carcinoma	26522444
MALAT1	osteosarcoma	17660802
MALAT1	ovarian cancer	18006640
MALAT1	ovarian endometrial cancer	27446438
MALAT1	Ovarian epithelial cancer	28770968
MALAT1	pancreatic cancer	25269958
MALAT1	pancreatic carcinoma	22996375
MALAT1	pancreatic ductal adenocarcinoma	24815433
MALAT1	papillary thyroid carcinoma	25997963
MALAT1	Parkinson's disease	27021022
MALAT1	pituitary adenoma	24469926
MALAT1	pre-eclampsia	26722461
MALAT1	primary pulmonary hypertension	27362960
MALAT1	proliferative vitreoretinopathy	26241674
MALAT1	Prostate	22996375
MALAT1	prostate cancer	21489289
MALAT1	renal cell carcinoma	25600645
MALAT1	renal clear cell carcinoma	25480417
MALAT1	retinal degeneration	24436191
MALAT1	retinoblastoma	28550678
MALAT1	rheumatoid arthritis	28026003
MALAT1	Seizures	22960213
MALAT1	squamous cell carcinoma	25538231
MALAT1	stomach cancer	24857172

MALAT1	systemic lupus erythematosus	29100395
MALAT1	TDP-43 protein, human	23791884
MALAT1	thyroid cancer	27470543
MALAT1	thyroid medullary carcinoma	29107050
MALAT1	tongue cancer	28260102
MALAT1	tongue squamous cell carcinoma	27353727
MALAT1	Triple Negative Breast Neoplasms	28915533
MALAT1	triple-receptor negative breast cancer	25996380
MALAT1	urinary bladder cancer	22722759
MALAT1	uterine cancer	21489289
MALAT1	uterine corpus endometrial stromal sarcoma	19379481
MALAT1	uveal melanoma	27725873
MALAT1	vulva squamous cell carcinoma	27633334
MAPT-AS1	Triple Negative Breast Neoplasms	29441192
MAPT-AS1	Parkinson's disease	27336847
MCM3AP-AS1	Glioblastoma	27229531
MEG3	breast cancer	14602737
MEG3	acute myeloid leukemia	19595458
MEG3	bladder urothelial carcinoma	28060759
MEG3	cancer	21400503
MEG3	cerebrovascular disease	27651151
MEG3	cervical cancer	14602737
MEG3	chronic myeloid leukemia	14602737
MEG3	chronic obstructive pulmonary disease	27932875
MEG3	colon cancer	14602737
MEG3	colorectal cancer	25636452
MEG3	diabetes mellitus	26603935
MEG3	endometrial cancer	27470401
MEG3	endometrial carcinoma	29094270
MEG3	esophageal cancer	28539329
MEG3	esophageal squamous cell carcinoma	28405686
MEG3	esophagus squamous cell carcinoma	27778235
MEG3	functionless pituitary adenoma	15644399
MEG3	gallbladder cancer	26812694
MEG3	gastric cardia adenocarcinoma	28345805
MEG3	Glioblastoma	22234798
MEG3	Glioma	28276316
MEG3	hepatocellular carcinoma	21625215
MEG3	Heroin Dependence	21128942

MEG3	Hirschsprung's disease	29050236
MEG3	Huntington's disease	22202438
MEG3	kidney cancer	24373479
MEG3	liver cancer	29449541
MEG3	liver cirrhosis	25201080
MEG3	liver disease	27770549
MEG3	lung adenocarcinoma	25992654
MEG3	lung cancer	14602737
MEG3	lung squamous cell carcinoma	28076325
MEG3	malignant glioma	14602737
MEG3	melanoma	27016304
MEG3	meningioma	20179190
MEG3	metabolic syndrome X	26898430
MEG3	multiple myeloma	25753650
MEG3	myelodysplastic syndrome	19595458
MEG3	myelofibrosis	24707949
MEG3	Nasopharyngeal carcinoma	27597634
MEG3	nephroblastoma	15798773
MEG3	neuroblastoma	15798773
MEG3	non-small cell lung carcinoma	24098911
MEG3	oral squamous cell carcinoma	23292713
MEG3	osteoarthritis	26090403
MEG3	ovarian cancer	28175963
MEG3	Ovarian epithelial cancer	24859196
MEG3	pancreatic cancer	26850851
MEG3	pancreatic endocrine carcinoma	25565142
MEG3	papillary thyroid carcinoma	25997963
MEG3	phaeochromocytoma	15798773
MEG3	pituitary adenoma	14602737
MEG3	pituitary cancer	18628527
MEG3	Prostate	14602737
MEG3	prostate cancer	14602737
MEG3	Purpura, Thrombocytopenic	27522004
MEG3	renal clear cell carcinoma	26223924
MEG3	retinoblastoma	26662307
MEG3	stomach cancer	24006224
MEG3	testicular germ cell cancer	27158395
MEG3	tongue squamous cell carcinoma	24343426
MEG3	type 1 diabetes mellitus	19966805

MEG3	urinary bladder cancer	14602737
MEG3	vulva squamous cell carcinoma	27633334
MIR31HG	Hirschsprung's disease	29626357
MIR31HG	lung adenocarcinoma	27903974
MIR31HG	non-small cell lung carcinoma	28529576
MIR31HG	pancreatic ductal adenocarcinoma	26549028
MIR31HG	stomach cancer	26692098
MIR31HG	urinary bladder cancer	27434291
MIR99AHG	megakaryoblastic leukemia	25027842
MIR99AHG	myeloid leukemia	25027842
NCK1-AS1	astrocytoma	26252651
NNT-AS1	breast cancer	29710510
NNT-AS1	cervical cancer	28628975
NNT-AS1	colorectal cancer	27966450
NNT-AS1	hepatocellular carcinoma	29179477
NNT-AS1	osteosarcoma	29518771
NNT-AS1	ovarian cancer	28969062
PCAT6	triple-receptor negative breast cancer	25996380
PCAT6	lung cancer	27458097
PCAT6	non-small cell lung carcinoma	27322209
PCAT6	prostate cancer	23728290
PIK3CD-AS2	astrocytoma	26252651
PTPRG-AS1	breast cancer	26409453
PVT1	breast cancer	17908964
PVT1	stomach cancer	27986464
PVT1	acute promyelocytic leukemia	26545364
PVT1	asthma	27484035
PVT1	astrocytoma	26252651
PVT1	B-cell lymphoma	23547836
PVT1	bladder urothelial carcinoma	28969069
PVT1	Burkitt lymphoma	17503467
PVT1	cancer	2725491
PVT1	Cardiomegaly	26045764
PVT1	cervical cancer	27232880
PVT1	clear cell renal cell carcinoma	29081406
PVT1	cleft lip	19270707
PVT1	colon cancer	25043044
PVT1	colorectal cancer	24196785
PVT1	diabetes mellitus	26971628

PVT1	Diabetic Nephropathies	21526116
PVT1	esophageal cancer	27698800
PVT1	esophageal squamous cell carcinoma	28404954
PVT1	Glioma	28351322
PVT1	hematologic cancer	26458445
PVT1	hepatocellular carcinoma	25624916
PVT1	Hodgkin's lymphoma	21037568
PVT1	kidney cancer	17881614
PVT1	lung squamous cell carcinoma	26928440
PVT1	lymph node metastasis	26882847
PVT1	lymphoma	2470097
PVT1	malignant glioma	27282637
PVT1	malignant pleural mesothelioma	24926545
PVT1	melanoma	28265576
PVT1	multiple myeloma	22869583
PVT1	Nasopharyngeal carcinoma	29445147
PVT1	non-small cell lung carcinoma	25400777
PVT1	osteosarcoma	28602700
PVT1	ovarian cancer	17908964
PVT1	pancreatic cancer	21316338
PVT1	pancreatic ductal adenocarcinoma	25668599
PVT1	papillary thyroid carcinoma	29280051
PVT1	plasmacytoma	17503467
PVT1	prostate cancer	21814516
PVT1	renal carcinoma	27366943
PVT1	renal cell carcinoma	26878386
PVT1	renal cell carcinoma	29152119
PVT1	stomach cancer	25258543
PVT1	thyroid cancer	26427660
PVT1	type 1 diabetes mellitus	21526116
PVT1	type 2 diabetes mellitus	17395743
PVT1	urinary bladder cancer	26517688
RAMP2-AS1	astrocytoma	26252651
RAMP2-AS1	Glioblastoma	27784795
RHPN1-AS1	uveal melanoma	28124977
RMST	breast cancer	27380926
RMST	melanoma	27016304
RMST	rhabdomyosarcoma	12082533
RMST	Triple Negative Breast Neoplasms	29215701

SNHG14	Angelman syndrome	23781896
SNHG14	Prader-Willi syndrome	23781896
SNHG14	stomach cancer	29667771
SNHG17	colorectal cancer	28933484
SNHG3	Alzheimer's disease	21961160
SNHG3	colorectal cancer	28731158
SNHG3	hepatocellular carcinoma	26373735
ST8SIA6-AS1	breast cancer	26929647
TINCR	astrocytoma	26252651
TINCR	chronic obstructive pulmonary disease	27932875
TINCR	colorectal cancer	27009809
TINCR	esophagus squamous cell carcinoma	26833746
TINCR	lung cancer	29324317
TINCR	non-small cell lung carcinoma	29427662
TINCR	squamous cell carcinoma	24115003
TINCR	stomach cancer	25728677
TINCR	urinary bladder cancer	27586866
TMPO-AS1	astrocytoma	26252651
TUSC8	cervical cancer	24667250
UCA1	breast cancer	16914571
UCA1	acute myeloid leukemia	26053097
UCA1	acute myocardial infarction	26949706
UCA1	astrocytoma	26252651
UCA1	bladder adenocarcinoma	25123267
UCA1	bladder carcinoma	29113184
UCA1	cancer	24457952
UCA1	cervical cancer	16914571
UCA1	cholangiocarcinoma	29221199
UCA1	chronic myeloid leukemia	27854515
UCA1	colon cancer	26885155
UCA1	colon carcinoma	16914571
UCA1	endometrial cancer	27540975
UCA1	esophageal cancer	16914571
UCA1	gallbladder cancer	28624787
UCA1	glandular cystitis	16914571
UCA1	Glioma	28105536
UCA1	hepatocellular carcinoma	25760077
UCA1	hypopharyngeal squamous cell carcinoma	28327194
UCA1	Lithiasis	16914571

UCA1	liver cancer	16914571
UCA1	lung cancer	26380024
UCA1	melanoma	24892958
UCA1	multiple myeloma	28543758
UCA1	muscle-invasive bladder cancer	27863388
UCA1	non-small cell lung carcinoma	26160838
UCA1	non-small cell lung carcinoma	27329842
UCA1	oral squamous cell carcinoma	23292713
UCA1	osteosarcoma	27335776
UCA1	osteosarcoma	28239821
UCA1	ovarian cancer	24379988
UCA1	pancreatic cancer	21593646
UCA1	pancreatic ductal adenocarcinoma	27628540
UCA1	prostate cancer	23728290
UCA1	Prostatic Hyperplasia	16914571
UCA1	renal cell carcinoma	16914571
UCA1	rheumatoid arthritis	29509238
UCA1	squamous cell carcinoma	17416635
UCA1	stomach cancer	16914571
UCA1	temporal lobe epilepsy	25552301
UCA1	thyroid cancer	16914571
UCA1	tongue squamous cell carcinoma	24332332
UCA1	urinary bladder cancer	16914571
VPS9D1-AS1	stomach cancer	29036784
XIST	breast cancer	17545591
XIST	acute lymphocytic leukemia	27535859
XIST	bladder carcinoma	29212249
XIST	cancer	23660942
XIST	cervical squamous cell carcinoma	27899965
XIST	collecting duct carcinoma	19154479
XIST	colon cancer	29679755
XIST	colorectal cancer	17143621
XIST	denatured dermis	28771809
XIST	esophageal squamous cell carcinoma	29100288
XIST	Glioblastoma	25578780
XIST	hematologic cancer	23415223
XIST	hepatocellular carcinoma	27100897
XIST	Klinefelter's syndrome	18854511
XIST	malignant glioma	25578780

XIST	melanoma	27016304
XIST	microinvasive gastric cancer	29039538
XIST	Nasopharyngeal carcinoma	27461945
XIST	neurodegenerative disease	22312272
XIST	non-small cell lung carcinoma	26339353
XIST	osteosarcoma	28409547
XIST	ovarian cancer	12492109
XIST	pancreatic cancer	28295543
XIST	prostate cancer	16261845
XIST	stomach cancer	27620004
XIST	testicular germ cell cancer	12629412
XIST	urinary bladder cancer	24373479
ZFAS1	breast cancer	21460236
ZFAS1	bladder carcinoma	29653362
ZFAS1	cancer	29137442
ZFAS1	Carcinoma, Ductal	21460236
ZFAS1	Colonic Neoplasms	27862275
ZFAS1	colorectal cancer	26506418
ZFAS1	Glioma	28081466
ZFAS1	hepatocellular carcinoma	26069248
ZFAS1	non-small cell lung carcinoma	28051258
ZFAS1	ovarian cancer	28099946
ZFAS1	Ovarian epithelial cancer	28099946
ZFAS1	prostate cancer	29416676
ZFAS1	rheumatoid arthritis	28721682
ZFAS1	stomach cancer	27246976
ZNF667-AS1	breast cancer	28690657
ZNF667-AS1	cervical cancer	29243775

Table S4: List of known lncRNAs associated with breast cancer and other diseases (*n* = 25).

lncRNA	**Disease Name**	**PubMed ID**
BCAR4	breast cancer	16778085
BCAR4	cervical cancer	28112728
BCAR4	chondrosarcoma	28399646
BCAR4	colon cancer	29190958
BCAR4	colorectal cancer	27197301
BCAR4	non-small cell lung carcinoma	28077810

BCAR4	osteosarcoma	27460090
BCAR4	stomach cancer	29028095
BCYRN1	breast cancer	9422992
BCYRN1	Aging	17553964
BCYRN1	Alzheimer's disease	1603265
BCYRN1	astrocytoma	25561975
BCYRN1	cancer	28651607
BCYRN1	cervical cancer	9422992
BCYRN1	colon cancer	29625226
BCYRN1	esophageal cancer	9422992
BCYRN1	Glioblastoma	25561975
BCYRN1	lung cancer	9422992
BCYRN1	malignant glioma	25561975
BCYRN1	microinvasive gastric cancer	29039538
BCYRN1	non-small cell lung carcinoma	25866480
BCYRN1	ovarian cancer	9422992
BCYRN1	parotid gland cancer	9422992
BCYRN1	squamous cell carcinoma	27143917
BCYRN1	tongue cancer	9422992
BCYRN1	asthma	28960519
CCAT1	breast cancer	26464701
CCAT1	acute myeloid leukemia	26923190
CCAT1	cancer	24594601
CCAT1	cervical cancer	28849215
CCAT1	colon cancer	29190961
CCAT1	colorectal cancer	23416875
CCAT1	endometrial cancer	27432114
CCAT1	esophageal squamous cell carcinoma	27956498
CCAT1	gallbladder cancer	25569100
CCAT1	Glioma	28475287
CCAT1	hepatocellular carcinoma	25884472
CCAT1	intrahepatic cholangiocarcinoma	28921383
CCAT1	laryngeal squamous cell carcinoma	28631575
CCAT1	lung adenocarcinoma	27566568
CCAT1	lung cancer	27212446
CCAT1	lung squamous cell carcinoma	28076325
CCAT1	malignant glioma	27765628
CCAT1	medulloblastoma	28777430
CCAT1	multiple myeloma	29228867

CCAT1	Nasopharyngeal carcinoma	28358263
CCAT1	non-small cell lung carcinoma	25129441
CCAT1	oral squamous cell carcinoma	28413645
CCAT1	osteosarcoma	28549102
CCAT1	ovarian cancer	24379988
CCAT1	Ovarian epithelial cancer	28754469
CCAT1	pancreatic cancer	28078015
CCAT1	renal cell carcinoma	28470345
CCAT1	retinoblastoma	28088735
CCAT1	stomach cancer	28535628
CCAT1	Neuralgia	29163801
DANCR	breast cancer	27716745
DANCR	astrocytoma	26252651
DANCR	bone disease	23438432
DANCR	brain cancer	29476310
DANCR	colorectal cancer	26617879
DANCR	hepatocellular carcinoma	25964079
DANCR	non-small cell lung carcinoma	29635134
DANCR	Osteoporosis, Postmenopausal	25660720
DANCR	osteosarcoma	26986815
DANCR	prostate cancer	23728290
DANCR	renal cell carcinoma	28765964
DANCR	stomach cancer	28618943
DANCR	Triple Negative Breast Neoplasms	28760736
DSCAM-AS1	breast cancer	12177779
DSCAM-AS1	idiopathic scoliosis	21216876
GAS5	breast cancer	18836484
GAS5	astrocytoma	26252651
GAS5	B-cell lymphoma	24583225
GAS5	bladder carcinoma	29445179
GAS5	bladder transitional cell carcinomas	27878359
GAS5	bladder urothelial carcinoma	28060759
GAS5	cancer	22996375
GAS5	cervical cancer	22487937
GAS5	colorectal cancer	24926850
GAS5	coronary artery disease	29267258
GAS5	endometrial carcinoma	26511107
GAS5	esophageal cancer	29386089
GAS5	esophageal squamous cell carcinoma	29170131

GAS5	Glioblastoma	23726844
GAS5	Glioma	28666797
GAS5	head and neck cancer	26482616
GAS5	hepatocellular carcinoma	25120813
GAS5	hypersensitivity reaction type II disease	20124551
GAS5	hypertension	27432865
GAS5	inflammatory bowel disease	28722800
GAS5	kidney cancer	24373479
GAS5	leukemia	20421347
GAS5	liver cirrhosis	26446789
GAS5	LPS-induced inflammatory injury	29448248
GAS5	lung adenocarcinoma	25925741
GAS5	lung cancer	26634743
GAS5	lymphoma	18406879
GAS5	malignant glioma	26370254
GAS5	malignant pleural mesothelioma	24885398
GAS5	mantle cell lymphoma	24703244
GAS5	melanoma	18836484
GAS5	multiple myeloma	24583225
GAS5	Nasopharyngeal carcinoma	28977945
GAS5	neuroblastoma	28035057
GAS5	non-small cell lung carcinoma	24357161
GAS5	osteoarthritis	25196583
GAS5	osteosarcoma	28519068
GAS5	ovarian cancer	26503132
GAS5	pancreatic cancer	24026436
GAS5	polycystic ovary syndrome	29648472
GAS5	Prostate	24373479
GAS5	prostate cancer	18836484
GAS5	renal cell carcinoma	23621190
GAS5	stomach cancer	24884417
GAS5	T-cell leukemia	18354083
GAS5	thyroid cancer	28506768
GAS5	Thyroid cancer, papillary	29423063
GAS5	type 2 diabetes mellitus	26675493
GAS5	urinary bladder cancer	24069260
H19	breast adenocarcinoma	9811352
H19	abdominal aortic aneurysm	29669788
H19	adenocarcinoma	8785513

H19	adrenocortical carcinoma	22019903
H19	aortic valve disease	27789555
H19	astrocytoma	25561975
H19	atherosclerosis	21954592
H19	Beckwith-Wiedemann syndrome	7987305
H19	bladder carcinoma	7589512
H19	breast cancer	12419837
H19	cancer	15618002
H19	cardiac fibroblast proliferation and fibrosis	27318893
H19	cardiomyocyte hypertrophy	27084844
H19	central nervous system disease	20380817
H19	cervical cancer	8570220
H19	cholangiocarcinoma	27809873
H19	cholestatic liver injury	29425397
H19	choriocarcinoma	8564957
H19	chronic myeloid leukemia	24685695
H19	colon cancer	15521051
H19	colon carcinoma	21489289
H19	colorectal cancer	8564957
H19	Congenital Hyperinsulinism	11395395
H19	coronary artery disease	25772106
H19	Diabetic Cardiomyopathies	27796346
H19	embryonal carcinoma	26415227
H19	endometrial cancer	27775072
H19	endometriosis	26089099
H19	esophageal cancer	8564957
H19	gallbladder cancer	27073719
H19	gastric adenocarcinoma	29479897
H19	gastric cardia adenocarcinoma	24414129
H19	gastrointestinal system cancer	27738631
H19	germ cell cancer	16001432
H19	gestational choriocarcinoma	8188082
H19	gestational trophoblastic neoplasm	12648595
H19	Glioblastoma	16707459
H19	Glioma	27981546
H19	growth restriction	20104244
H19	head and neck squamous cell carcinoma	27994496
H19	Heart Defects, Congenital	27035723
H19	heart disease	27895893

H19	Hematopoiesis	15645136
H19	hepatocellular carcinoma	15736456
H19	Hydatidiform Mole	12783848
H19	hyperhomocysteinemia	15899898
H19	hyperprolactinemia	15525575
H19	infertility	20042264
H19	intestinal epithelial barrier function	26884465
H19	Keloid	27698867
H19	kidney cancer	24373479
H19	laryngeal squamous cell carcinoma	26872375
H19	liver cancer	11175353
H19	lung adenocarcinoma	25758555
H19	lung cancer	8564957
H19	malignant glioma	20380817
H19	Marek Disease	10696440
H19	medulloblastoma	8957451
H19	melanoma	11437411
H19	meningioma	10738131
H19	Mullerian aplasia	21458801
H19	multiple myeloma	29273733
H19	myeloproliferative neoplasm	12682647
H19	Nasopharyngeal carcinoma	27040767
H19	nephroblastoma	16179496
H19	Neural Tube Defects	22234160
H19	neuroblastoma	23791884
H19	non-small cell lung carcinoma	26482621
H19	obesity	22341586
H19	oral squamous cell carcinoma	28975993
H19	osteoarthritis	25430712
H19	osteosarcoma	24141783
H19	ovarian cancer	19656414
H19	Ovarian epithelial cancer	10428315
H19	pancreatic cancer	24920070
H19	pancreatic ductal adenocarcinoma	24920070
H19	papillary thyroid carcinoma	29287713
H19	Parkinson's disease	27021022
H19	Pheochromocytoma	21937622
H19	pituitary adenoma	23791884
H19	pneumoconiosis	27626436

H19	polycythemia vera	10640993
H19	Prader-Willi syndrome	23791884
H19	pre-eclampsia	19570415
H19	Prostate	24373479
H19	prostate cancer	24063685
H19	renal cell carcinoma	25866221
H19	rheumatoid arthritis	12937131
H19	Silver-Russell syndrome	19066168
H19	squamous cell carcinoma	22996375
H19	stomach cancer	9570380
H19	thyroid cancer	27093644
H19	trophoblastic neoplasm	8188082
H19	ulcerative colitis	27661667
H19	urinary bladder cancer	10413100
HIF1A-AS2	breast cancer	22664915
HIF1A-AS2	breast carcinoma	14580258
HIF1A-AS2	Glioblastoma	27264189
HIF1A-AS2	kidney cancer	9923855
HIF1A-AS2	osteosarcoma	23466354
HIF1A-AS2	stomach cancer	25686741
HIF1A-AS2	urinary bladder cancer	27018306
HOTAIR	breast cancer	19182780
HOTAIR	Abortion, Habitual	28750739
HOTAIR	acute myeloid leukemia	25979172
HOTAIR	Asthenozoospermia	26823733
HOTAIR	astrocytoma	25085602
HOTAIR	atypical teratoid rhabdoid tumor	25085602
HOTAIR	B-cell lymphoma	24583225
HOTAIR	bladder carcinoma	29673865
HOTAIR	bladder urothelial carcinoma	26781446
HOTAIR	cancer	29463216
HOTAIR	cerebrovascular disease	27613094
HOTAIR	cervical cancer	22487937
HOTAIR	chronic myeloid leukemia	27875938
HOTAIR	colon cancer	24667321
HOTAIR	colorectal cancer	21862635
HOTAIR	congestive heart failure	27317124
HOTAIR	cutaneous squamous cell carcinoma	27067026
HOTAIR	diffuse large B-cell lymphoma	27550047

HOTAIR	embryonal cancer	25085602
HOTAIR	endometrial cancer	24285342
HOTAIR	endometrial carcinoma	29466670
HOTAIR	Ependymoma	25085602
HOTAIR	esophageal cancer	28441714
HOTAIR	esophageal squamous cell carcinoma	27810266
HOTAIR	esophagus squamous cell carcinoma	24022190
HOTAIR	functionless pituitary adenoma	24469926
HOTAIR	gallbladder cancer	24953832
HOTAIR	gastric adenocarcinoma	23888369
HOTAIR	gastric cardia adenocarcinoma	25476857
HOTAIR	gastrointestinal stromal tumor	27659532
HOTAIR	gastrointestinal system cancer	24667321
HOTAIR	Glioblastoma	24203894
HOTAIR	Glioma	28083786
HOTAIR	head and neck squamous cell carcinoma	26592246
HOTAIR	heart disease	24788418
HOTAIR	hepatitis C	27129296
HOTAIR	hepatocellular carcinoma	21327457
HOTAIR	kidney cancer	24616104
HOTAIR	laryngeal squamous cell carcinoma	23141928
HOTAIR	Lemierre's syndrome	26806307
HOTAIR	leukemia	27748863
HOTAIR	Leukemia, Lymphoid	29513085
HOTAIR	liver cancer	24667321
HOTAIR	liver cirrhosis	27979710
HOTAIR	lung adenocarcinoma	24155936
HOTAIR	lung cancer	23668363
HOTAIR	lung small cell carcinoma	24591352
HOTAIR	malignant glioma	24203894
HOTAIR	medulloblastoma	25085602
HOTAIR	melanoma	23862139
HOTAIR	multiple myeloma	24583225
HOTAIR	Nasopharyngeal carcinoma	23281836
HOTAIR	neuroblastoma	29603181
HOTAIR	non-small cell lung carcinoma	23743197
HOTAIR	osteoarthritis	25430712
HOTAIR	osteosarcoma	25728753
HOTAIR	ovarian cancer	23600210

HOTAIR	Ovarian epithelial cancer	24662839
HOTAIR	pancreatic cancer	22614017
HOTAIR	pancreatic carcinoma	24667321
HOTAIR	pancreatic ductal adenocarcinoma	26482614
HOTAIR	papillary thyroid carcinoma	25997963
HOTAIR	Parkinson's disease	26979073
HOTAIR	pituitary adenoma	24469926
HOTAIR	pre-eclampsia	25807808
HOTAIR	prostate cancer	20864820
HOTAIR	renal carcinoma	25149152
HOTAIR	renal cell carcinoma	24935377
HOTAIR	retinoblastoma	27966488
HOTAIR	rheumatoid arthritis	24722995
HOTAIR	sarcoma	23543869
HOTAIR	solid tumors	27333150
HOTAIR	sporadic thoracic aortic aneurysm	28757056
HOTAIR	squamous cell carcinoma	23717443
HOTAIR	stomach cancer	23847441
HOTAIR	thyroid cancer	28565838
HOTAIR	triple-receptor negative breast cancer	25996380
HOTAIR	urinary bladder cancer	25030736
HOXA-AS2	breast cancer	28545023
HOXA-AS2	acute promyelocytic leukemia	23649634
HOXA-AS2	colorectal cancer	28112720
HOXA-AS2	gallbladder carcinoma	28388535
HOXA-AS2	hepatocellular carcinoma	27855366
HOXA-AS2	malignant glioma	29310118
HOXA-AS2	melanoma	27016304
HOXA-AS2	stomach cancer	26384350
KCNK15-AS1	breast cancer	25929808
KCNK15-AS1	osteoarthritis	25430712
LINC00472	breast cancer	25865225
LINC00472	colorectal cancer	29488624
LINC00472	lung adenocarcinoma	27826625
LINC00472	ovarian cancer	27667152
LINC00511	breast cancer	26929647
LINC00511	lung adenocarcinoma	27797003
LINC00511	non-small cell lung carcinoma	27845772
MALAT1	breast cancer	18006640

MALAT1	acute monocytic leukemia	28713913
MALAT1	acute myeloid leukemia	28713913
MALAT1	amyotrophic lateral sclerosis	27338628
MALAT1	astrocytoma	26252651
MALAT1	B-cell lymphoma	21489289
MALAT1	bladder carcinoma	28648755
MALAT1	bladder urothelial carcinoma	23153939
MALAT1	calcific aortic valve disease	28522163
MALAT1	cancer	20711585
MALAT1	cervical cancer	20213048
MALAT1	cholangiocarcinoma	28592124
MALAT1	choriocarcinoma	29096355
MALAT1	colon cancer	21489289
MALAT1	colorectal cancer	21503572
MALAT1	Congenital Microtia	26282502
MALAT1	decreased myogenesis	23485710
MALAT1	diabetes mellitus	24436191
MALAT1	diabetes mellitus	26512840
MALAT1	Diabetic Cardiomyopathies	26476026
MALAT1	Diabetic Nephropathies	27964927
MALAT1	endometrial adenocarcinoma	25085246
MALAT1	endometrial stromal sarcoma	16441420
MALAT1	esophageal cancer	27470544
MALAT1	esophageal squamous cell carcinoma	27935117
MALAT1	fatty liver disease	26935028
MALAT1	Fibroma	27101025
MALAT1	fibrosarcoma	22491206
MALAT1	Flavivirus Infections	26634309
MALAT1	Follicular and H¬ürthle Cell Thyroid Neoplasm	28660408
MALAT1	gallbladder cancer	24658096
MALAT1	gastrointestinal system cancer	27313790
MALAT1	Glioblastoma	25772239
MALAT1	Glioma	27313790
MALAT1	hepatocellular carcinoma	16878148
MALAT1	high glucose-induced podocyte injury	28444861
MALAT1	histiocytoid hemangioma	27709553
MALAT1	HIV	26139386
MALAT1	Hyperglycemia	25787249
MALAT1	ischemic stroke	28093478

MALAT1	kidney cancer	24373479
MALAT1	Klatskin's tumor	28059437
MALAT1	laryngeal squamous cell carcinoma	24817925
MALAT1	liver cancer	21489289
MALAT1	liver cirrhosis	26697839
MALAT1	lung adenocarcinoma	19690017
MALAT1	lung cancer	17270048
MALAT1	lung small cell carcinoma	22928560
MALAT1	lymph node metastasis	26989678
MALAT1	malignant glioma	24926466
MALAT1	mantle cell lymphoma	27998273
MALAT1	melanoma	19625619
MALAT1	multiple myeloma	24583225
MALAT1	Nasopharyngeal carcinoma	23688988
MALAT1	neuroblastoma	20149803
MALAT1	non-small cell lung carcinoma	12970751
MALAT1	oral squamous cell carcinoma	26522444
MALAT1	osteosarcoma	17660802
MALAT1	ovarian cancer	18006640
MALAT1	ovarian endometrial cancer	27446438
MALAT1	Ovarian epithelial cancer	28770968
MALAT1	pancreatic cancer	25269958
MALAT1	pancreatic carcinoma	22996375
MALAT1	pancreatic ductal adenocarcinoma	24815433
MALAT1	papillary thyroid carcinoma	25997963
MALAT1	Parkinson's disease	27021022
MALAT1	pituitary adenoma	24469926
MALAT1	pre-eclampsia	26722461
MALAT1	primary pulmonary hypertension	27362960
MALAT1	proliferative vitreoretinopathy	26241674
MALAT1	Prostate	22996375
MALAT1	prostate cancer	21489289
MALAT1	renal cell carcinoma	25600645
MALAT1	renal clear cell carcinoma	25480417
MALAT1	retinal degeneration	24436191
MALAT1	retinoblastoma	28550678
MALAT1	rheumatoid arthritis	28026003
MALAT1	Seizures	22960213
MALAT1	squamous cell carcinoma	25538231

MALAT1	stomach cancer	24857172
MALAT1	systemic lupus erythematosus	29100395
MALAT1	TDP-43 protein, human	23791884
MALAT1	thyroid cancer	27470543
MALAT1	thyroid medullary carcinoma	29107050
MALAT1	tongue cancer	28260102
MALAT1	tongue squamous cell carcinoma	27353727
MALAT1	Triple Negative Breast Neoplasms	28915533
MALAT1	triple-receptor negative breast cancer	25996380
MALAT1	urinary bladder cancer	22722759
MALAT1	uterine cancer	21489289
MALAT1	uterine corpus endometrial stromal sarcoma	19379481
MALAT1	uveal melanoma	27725873
MALAT1	vulva squamous cell carcinoma	27633334
MAPT-AS1	Triple Negative Breast Neoplasms	29441192
MAPT-AS1	Parkinson's disease	27336847
MEG3	breast cancer	14602737
MEG3	acute myeloid leukemia	19595458
MEG3	bladder urothelial carcinoma	28060759
MEG3	cancer	21400503
MEG3	cerebrovascular disease	27651151
MEG3	cervical cancer	14602737
MEG3	chronic myeloid leukemia	14602737
MEG3	chronic obstructive pulmonary disease	27932875
MEG3	colon cancer	14602737
MEG3	colorectal cancer	25636452
MEG3	diabetes mellitus	26603935
MEG3	endometrial cancer	27470401
MEG3	endometrial carcinoma	29094270
MEG3	esophageal cancer	28539329
MEG3	esophageal squamous cell carcinoma	28405686
MEG3	esophagus squamous cell carcinoma	27778235
MEG3	functionless pituitary adenoma	15644399
MEG3	gallbladder cancer	26812694
MEG3	gastric cardia adenocarcinoma	28345805
MEG3	Glioblastoma	22234798
MEG3	Glioma	28276316
MEG3	hepatocellular carcinoma	21625215
MEG3	Heroin Dependence	21128942

MEG3	Hirschsprung's disease	29050236
MEG3	Huntington's disease	22202438
MEG3	kidney cancer	24373479
MEG3	liver cancer	29449541
MEG3	liver cirrhosis	25201080
MEG3	liver disease	27770549
MEG3	lung adenocarcinoma	25992654
MEG3	lung cancer	14602737
MEG3	lung squamous cell carcinoma	28076325
MEG3	malignant glioma	14602737
MEG3	melanoma	27016304
MEG3	meningioma	20179190
MEG3	metabolic syndrome X	26898430
MEG3	multiple myeloma	25753650
MEG3	myelodysplastic syndrome	19595458
MEG3	myelofibrosis	24707949
MEG3	Nasopharyngeal carcinoma	27597634
MEG3	nephroblastoma	15798773
MEG3	neuroblastoma	15798773
MEG3	non-small cell lung carcinoma	24098911
MEG3	oral squamous cell carcinoma	23292713
MEG3	osteoarthritis	26090403
MEG3	ovarian cancer	28175963
MEG3	Ovarian epithelial cancer	24859196
MEG3	pancreatic cancer	26850851
MEG3	pancreatic endocrine carcinoma	25565142
MEG3	papillary thyroid carcinoma	25997963
MEG3	phaeochromocytoma	15798773
MEG3	pituitary adenoma	14602737
MEG3	pituitary cancer	18628527
MEG3	Prostate	14602737
MEG3	prostate cancer	14602737
MEG3	Purpura, Thrombocytopenic	27522004
MEG3	renal clear cell carcinoma	26223924
MEG3	retinoblastoma	26662307
MEG3	stomach cancer	24006224
MEG3	testicular germ cell cancer	27158395
MEG3	tongue squamous cell carcinoma	24343426
MEG3	type 1 diabetes mellitus	19966805

MEG3	urinary bladder cancer	14602737
MEG3	vulva squamous cell carcinoma	27633334
NNT-AS1	breast cancer	29710510
NNT-AS1	cervical cancer	28628975
NNT-AS1	colorectal cancer	27966450
NNT-AS1	hepatocellular carcinoma	29179477
NNT-AS1	osteosarcoma	29518771
NNT-AS1	ovarian cancer	28969062
PCAT6	triple-receptor negative breast cancer	25996380
PCAT6	lung cancer	27458097
PCAT6	non-small cell lung carcinoma	27322209
PCAT6	prostate cancer	23728290
PIK3CD-AS2	astrocytoma	26252651
PVT1	breast cancer	17908964
PVT1	stomach cancer	27986464
PVT1	acute promyelocytic leukemia	26545364
PVT1	asthma	27484035
PVT1	astrocytoma	26252651
PVT1	B-cell lymphoma	23547836
PVT1	bladder urothelial carcinoma	28969069
PVT1	Burkitt lymphoma	17503467
PVT1	cancer	2725491
PVT1	Cardiomegaly	26045764
PVT1	cervical cancer	27232880
PVT1	clear cell renal cell carcinoma	29081406
PVT1	cleft lip	19270707
PVT1	colon cancer	25043044
PVT1	colorectal cancer	24196785
PVT1	diabetes mellitus	26971628
PVT1	Diabetic Nephropathies	21526116
PVT1	esophageal cancer	27698800
PVT1	esophageal squamous cell carcinoma	28404954
PVT1	Glioma	28351322
PVT1	hematologic cancer	26458445
PVT1	hepatocellular carcinoma	25624916
PVT1	Hodgkin's lymphoma	21037568
PVT1	kidney cancer	17881614
PVT1	lung squamous cell carcinoma	26928440
PVT1	lymph node metastasis	26882847

PVT1	lymphoma	2470097
PVT1	malignant glioma	27282637
PVT1	malignant pleural mesothelioma	24926545
PVT1	melanoma	28265576
PVT1	multiple myeloma	22869583
PVT1	Nasopharyngeal carcinoma	29445147
PVT1	non-small cell lung carcinoma	25400777
PVT1	osteosarcoma	28602700
PVT1	ovarian cancer	17908964
PVT1	pancreatic cancer	21316338
PVT1	pancreatic ductal adenocarcinoma	25668599
PVT1	papillary thyroid carcinoma	29280051
PVT1	plasmacytoma	17503467
PVT1	prostate cancer	21814516
PVT1	renal carcinoma	27366943
PVT1	renal cell carcinoma	26878386
PVT1	renal cell carcinoma	29152119
PVT1	stomach cancer	25258543
PVT1	thyroid cancer	26427660
PVT1	type 1 diabetes mellitus	21526116
PVT1	type 2 diabetes mellitus	17395743
PVT1	urinary bladder cancer	26517688
RMST	breast cancer	27380926
RMST	melanoma	27016304
RMST	rhabdomyosarcoma	12082533
RMST	Triple Negative Breast Neoplasms	29215701
UCA1	breast cancer	16914571
UCA1	acute myeloid leukemia	26053097
UCA1	acute myocardial infarction	26949706
UCA1	astrocytoma	26252651
UCA1	bladder adenocarcinoma	25123267
UCA1	bladder carcinoma	29113184
UCA1	cancer	24457952
UCA1	cervical cancer	16914571
UCA1	cholangiocarcinoma	29221199
UCA1	chronic myeloid leukemia	27854515
UCA1	colon cancer	26885155
UCA1	colon carcinoma	16914571
UCA1	endometrial cancer	27540975

UCA1	esophageal cancer	16914571
UCA1	gallbladder cancer	28624787
UCA1	glandular cystitis	16914571
UCA1	Glioma	28105536
UCA1	hepatocellular carcinoma	25760077
UCA1	hypopharyngeal squamous cell carcinoma	28327194
UCA1	Lithiasis	16914571
UCA1	liver cancer	16914571
UCA1	lung cancer	26380024
UCA1	melanoma	24892958
UCA1	multiple myeloma	28543758
UCA1	muscle-invasive bladder cancer	27863388
UCA1	non-small cell lung carcinoma	26160838
UCA1	non-small cell lung carcinoma	27329842
UCA1	oral squamous cell carcinoma	23292713
UCA1	osteosarcoma	27335776
UCA1	osteosarcoma	28239821
UCA1	ovarian cancer	24379988
UCA1	pancreatic cancer	21593646
UCA1	pancreatic ductal adenocarcinoma	27628540
UCA1	prostate cancer	23728290
UCA1	Prostatic Hyperplasia	16914571
UCA1	renal cell carcinoma	16914571
UCA1	rheumatoid arthritis	29509238
UCA1	squamous cell carcinoma	17416635
UCA1	stomach cancer	16914571
UCA1	temporal lobe epilepsy	25552301
UCA1	thyroid cancer	16914571
UCA1	tongue squamous cell carcinoma	24332332
UCA1	urinary bladder cancer	16914571
XIST	breast cancer	17545591
XIST	acute lymphocytic leukemia	27535859
XIST	bladder carcinoma	29212249
XIST	cancer	23660942
XIST	cervical squamous cell carcinoma	27899965
XIST	collecting duct carcinoma	19154479
XIST	colon cancer	29679755
XIST	colorectal cancer	17143621
XIST	denatured dermis	28771809

XIST	esophageal squamous cell carcinoma	29100288
XIST	Glioblastoma	25578780
XIST	hematologic cancer	23415223
XIST	hepatocellular carcinoma	27100897
XIST	Klinefelter's syndrome	18854511
XIST	malignant glioma	25578780
XIST	melanoma	27016304
XIST	microinvasive gastric cancer	29039538
XIST	Nasopharyngeal carcinoma	27461945
XIST	neurodegenerative disease	22312272
XIST	non-small cell lung carcinoma	26339353
XIST	osteosarcoma	28409547
XIST	ovarian cancer	12492109
XIST	pancreatic cancer	28295543
XIST	prostate cancer	16261845
XIST	stomach cancer	27620004
XIST	testicular germ cell cancer	12629412
XIST	urinary bladder cancer	24373479
ZFAS1	breast cancer	21460236
ZFAS1	bladder carcinoma	29653362
ZFAS1	cancer	29137442
ZFAS1	Carcinoma, Ductal	21460236
ZFAS1	Colonic Neoplasms	27862275
ZFAS1	colorectal cancer	26506418
ZFAS1	Glioma	28081466
ZFAS1	hepatocellular carcinoma	26069248
ZFAS1	non-small cell lung carcinoma	28051258
ZFAS1	ovarian cancer	28099946
ZFAS1	Ovarian epithelial cancer	28099946
ZFAS1	prostate cancer	29416676
ZFAS1	rheumatoid arthritis	28721682
ZFAS1	stomach cancer	27246976
ZNF667-AS1	breast cancer	28690657
ZNF667-AS1	cervical cancer	29243775

Table S5: List of known lncRNAs associated with only breast cancer (*n* = 6).

lncRNA	Disease Name	PubMed ID
AC008268.1	breast cancer	26910840
FGF14-AS2	breast cancer	26820525
LINC00993	breast cancer	25996380
LINC01016	breast cancer	26426411
PTPRG-AS1	breast cancer	26409453
ST8SIA6-AS1	breast cancer	26929647

Table S6: List of common lncRNAs between novel (*n* = 38) and overlapping in all three methods (*n* = 21) (see Figure 3. Venn diagram) (*n* = 6)

LncRNA	Sub-type	Chrom	Start	End
AC005152.3	Basal	chr17	72021851	72034092
TTC39A-AS1	Basal	chr1	51329654	51335324
SEMA3B-AS1	LumA	chr3	50266641	50267371
RARA-AS1	LumB	chr17	40340867	40343136
TPTEP1	LumB	chr22	16601887	16698742
CTB-51J22.1	Normal-like	chr7	74059576	74062284

Table S7: Survival analysis for 91 key lncRNAs with respect to whole cohort as well as subtype-specific cohort. There are unique 44 unique lncRNAs found prognostically significant.

lncRNA	P-value	Hazard Ratio	95% CI Low	95% CI High	Cohort
LINC00152	0.0071	0.354	0.168	0.748	Basal
PTPRG-AS1	0.0240	2.425	1.138	5.166	Basal
TTC39A-AS1	0.0046	0.313	0.149	0.655	Basal
AC016735.2	0.0453	0.359	0.134	0.962	HER2
AC087491.2	0.0431	3.027	1.138	8.048	HER2
DLEU2	0.0051	0.229	0.085	0.621	HER2
ELOVL2-AS1	0.0272	0.338	0.122	0.936	HER2
GATA3-AS1	0.0387	0.323	0.122	0.860	HER2
HOTAIRM1	0.0172	3.321	1.223	9.016	HER2
KRTAP5-AS1	0.0497	2.747	1.027	7.349	HER2
LINC00504	0.0336	0.314	0.118	0.835	HER2
PRKAG2-AS1	0.0128	3.739	1.398	10.003	HER2

RP11-281O15.4	0.0113	3.539	1.295	9.676	HER2
SEMA3B-AS1	0.0149	3.697	1.383	9.884	HER2
CTD-2015G9.2	0.0280	1.719	1.073	2.753	Luminal A
HOTAIRM1	0.0480	1.611	1.009	2.572	Luminal A
LINC00992	0.0258	0.583	0.365	0.932	Luminal A
LINC01016	0.0263	1.689	1.055	2.705	Luminal A
MALAT1	0.0026	0.485	0.303	0.775	Luminal A
MAPT-AS1	0.0131	1.820	1.139	2.908	Luminal A
MCM3AP-AS1	0.0131	0.550	0.344	0.879	Luminal A
NNT-AS1	0.0441	0.615	0.385	0.981	Luminal A
RHPN1-AS1	0.0381	0.605	0.379	0.966	Luminal A
STK4-AS1	0.0095	0.538	0.336	0.861	Luminal A
UCA1	0.0264	1.704	1.067	2.723	Luminal A
XIST	0.0096	0.537	0.336	0.858	Luminal A
ZFAS1	0.0285	1.683	1.051	2.696	Luminal A
AC005152.3	0.0066	2.599	1.324	5.104	Luminal B
BCAR4	0.0149	0.430	0.219	0.846	Luminal B
CTD-2015G9.2	0.0034	2.856	1.457	5.601	Luminal B
CTD-2284J15.1	0.0262	2.252	1.151	4.407	Luminal B
DOCK9-AS2	0.0248	2.157	1.095	4.248	Luminal B
ELOVL2-AS1	0.0013	3.350	1.713	6.554	Luminal B
LINC01016	0.0117	2.349	1.187	4.649	Luminal B
MCM3AP-AS1	0.0484	1.959	0.980	3.916	Luminal B
RP11-28F1.2	0.0361	2.073	1.059	4.060	Luminal B
XIST	0.0378	2.026	1.031	3.980	Luminal B
LINC00152	0.0349	0.143	0.032	0.628	Normal-Like
LINC01272	0.0071	0.099	0.021	0.457	Normal-Like
AC005152.3	0.0414	1.385	1.013	1.893	Whole
BCAR4	0.0109	0.665	0.486	0.909	Whole
CTB-33O18.1	0.0284	1.424	1.041	1.948	Whole
CTB-51J22.1	0.0329	1.414	1.034	1.933	Whole
CTD-2015G9.2	0.0077	1.530	1.119	2.093	Whole
ELOVL2-AS1	0.0014	1.679	1.229	2.295	Whole
FGF14-AS2	0.0172	1.460	1.067	1.997	Whole

LINC00472	0.0105	1.503	1.100	2.055	Whole
LINC01016	0.0258	1.426	1.043	1.951	Whole
MAPT-AS1	0.0019	1.646	1.204	2.251	Whole
MIR205HG	0.0041	1.571	1.147	2.153	Whole
MIR31HG	0.0425	1.384	1.013	1.892	Whole
PIK3CD-AS2	0.0215	1.440	1.052	1.969	Whole
RHPN1-AS1	0.0075	0.650	0.475	0.888	Whole
RMST	0.0010	1.692	1.238	2.314	Whole
RP11-21L23.2	0.0160	0.679	0.497	0.927	Whole
RP11-28F1.2	0.0006	1.744	1.276	2.384	Whole
RP1-232P20.1	0.0324	1.403	1.025	1.919	Whole
SYN2	0.0365	1.397	1.022	1.910	Whole
VPS9D1-AS1	0.0263	1.426	1.042	1.952	Whole

8.8 Appendix B

8.8.1 Appendix B.1: Hyperparameter Tuning

In the decoder part of the CAE, different node values from 240 to 340, with a step size of 10 was tested to tune the number of nodes in the decoder layer. It was found the 300 nodes would yield the highest accuracy, as evident in Figure 1. So the number of nodes was selected as 300.

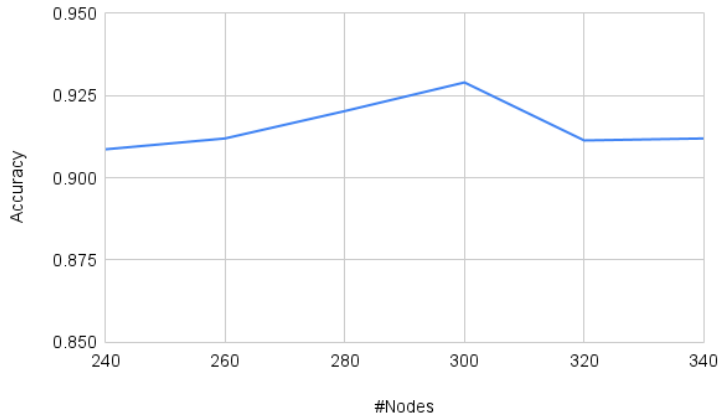

Figure S7: Tuning number of nodes in the decoder. For 300 nodes, it yields the highest accuracy.

We did a random search of parameters from a range of values to tune the number of epochs and learning rate. For the number of epochs, the used values are - 200, 300, 500, 1000, 1500, 2000, 2500, 3000. Similarly, for learning rate, the values are - 0.001, 0.002, 0.005, 0.0005, 0.01 and 0.05. Table 1 contains the accuracy of all different combinations of epoch and learning rate. The highest value of accuracy was 0.9507, which is found for the epoch of 300 and learning rate of 0.002.

Table S8: Summary of hyperparameter tuning for epoch and learning rates. For different values of epoch and learning rate, there are different accuracies for the SVM model and by the features selected by CAE.

Epoch	Learning Rate	Accuracy	Epoch	Learning Rate	Accuracy
200	0.0005	0.9256	500	0.005	0.9371
300	0.0005	0.9203	1000	0.005	0.9403
1000	0.0005	0.9340	1500	0.005	0.9266
1500	0.0005	0.9308	2000	0.005	0.9224
2000	0.0005	0.9277	2500	0.005	0.9308
2500	0.0005	0.9340	2500	0.005	0.9235
3000	0.0005	0.9434	3000	0.005	0.9224
200	0.001	0.9256	200	0.01	0.2296
300	0.001	0.9382	300	0.01	0.2180
500	0.001	0.9361	500	0.01	0.2317
1000	0.001	0.9444	1000	0.01	0.3071
1500	0.001	0.9476	1500	0.01	0.2453
2000	0.001	0.9340	2000	0.01	0.4130
2500	0.001	0.9497	2500	0.01	0.2233
3000	0.001	0.9413	3000	0.01	0.2914
200	0.002	0.9266	200	0.05	0.2421
300	0.002	0.9507	300	0.05	0.2411
500	0.002	0.9444	500	0.05	0.2254
1000	0.002	0.9486	1000	0.05	0.2222
1500	0.002	0.9392	1500	0.05	0.2379
2000	0.002	0.9434	2000	0.05	0.2254
2500	0.002	0.9319	2500	0.05	0.2285
3000	0.002	0.9403	2500	0.05	0.2159
200	0.005	0.9361	3000	0.05	0.2170

8.8.2 Appendix B.2: Comparing tumor features with normal features

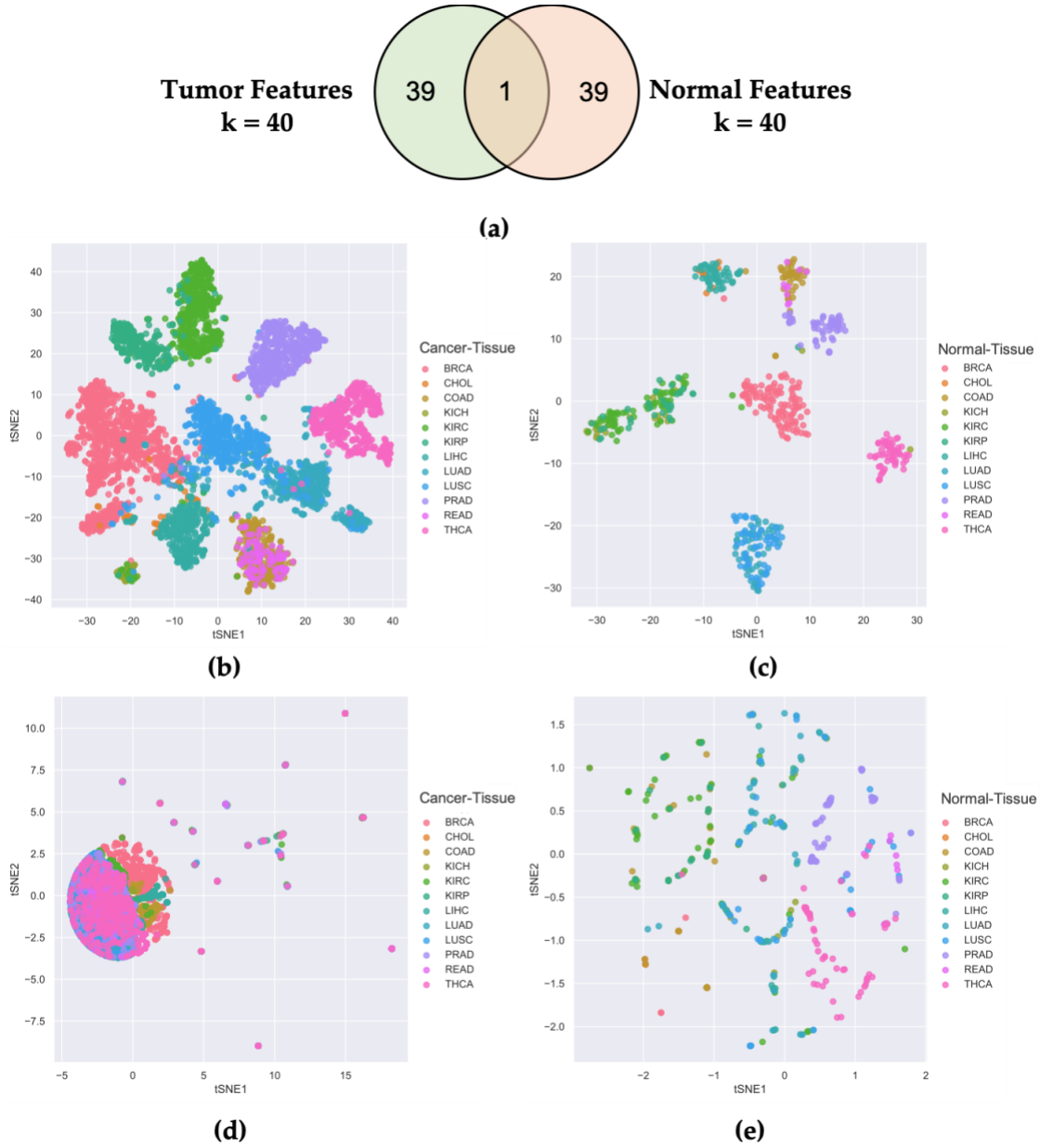

Figure S8: Comparing tumor features with normal features. a) Venn diagram of top 40 tumor features and top 40 normal features derived from CAE; b) t-SNE plot of tumor samples using tumor features; c) t-SNE plot of normal samples using normal features; d) t-SNE plot of tumor samples using normal features; e) t-SNE plot of normal samples using tumor features.

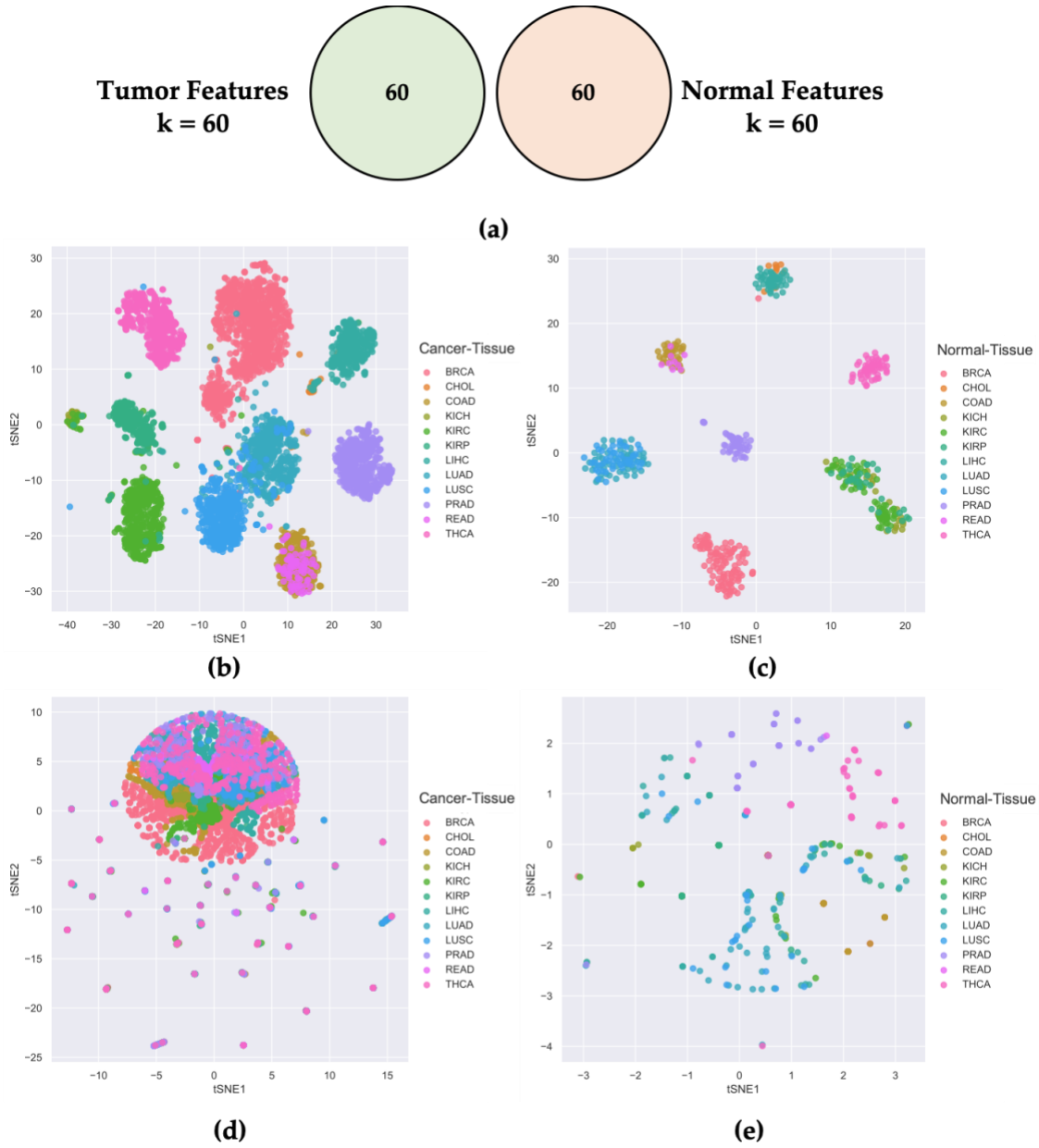

Figure S9: Comparing tumor features with normal features. a) Venn diagram of top 60 tumor features and top 60 normal features derived from CAE; b) t-SNE plot of tumor samples using tumor features; c) t-SNE plot of normal samples using normal features; d) t-SNE plot of tumor samples using normal features; e) t-SNE plot of normal samples using tumor features.

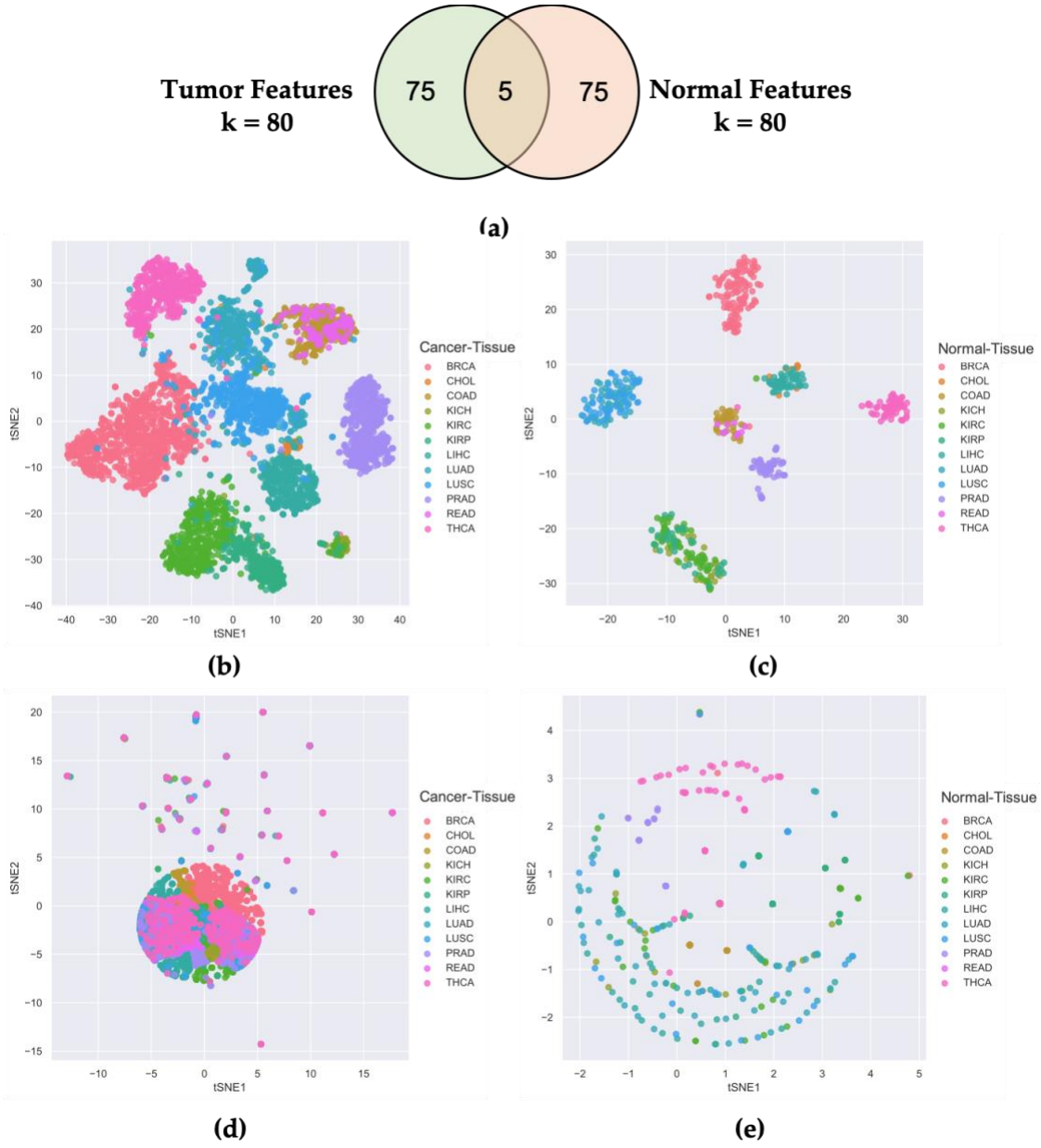

Figure S10: Comparing tumor features with normal features. a) Venn diagram of top 80 tumor features and top 80 normal features derived from CAE; b) t-SNE plot of tumor samples using tumor features; c) t-SNE plot of normal samples using normal features; d) t-SNE plot of tumor samples using normal features; e) t-SNE plot of normal samples using tumor features.

8.8.3 Appendix B.3: Top 128 lncRNAs

Table S9: Top 128 lncRNAs

id	gene	chrom	chromStart	chromEnd	strand
ENSG00000250522.1	AC004066.3	chr4	1.06E+08	1.06E+08	-
ENSG00000230918.1	AC008063.2	chr2	1.62E+08	1.62E+08	+
ENSG00000235584.2	AC008268.1	chr2	95666084	95668715	+
ENSG00000267968.1	AC011523.2	chr19	50830530	50851089	+
ENSG00000231013.1	AC013275.2	chr2	1.19E+08	1.19E+08	+
ENSG00000232153.2	AC073218.3	chr2	34734975	34737118	-
ENSG00000223914.1	AC079630.2	chr12	40156239	40167707	+
ENSG00000225342.2	AC079630.4	chr12	40186009	40224915	-
ENSG00000233850.1	AC103563.8	chr2	95025193	95026709	-
ENSG00000232555.1	AC104088.1	chr2	1.72E+08	1.72E+08	+
ENSG00000235688.1	AC116614.1	chr2	949634	950274	-
ENSG00000229380.1	AC147651.5	chr7	561958	565619	+
ENSG00000241158.4	ADAMTS9-AS1	chr3	64561322	64592757	+
ENSG00000178803.9	ADORA2A-AS1	chr22	24429206	24495074	-
ENSG00000237609.1	AF064858.10	chr21	39028536	39029128	-
ENSG00000235888.1	AF064858.8	chr21	38988707	39006153	-
ENSG00000255020.1	AF131216.5	chr8	11345748	11347502	-
ENSG00000232855.5	AF131217.1	chr21	28439346	28674848	-
ENSG00000228923.1	AP000355.2	chr22	24516508	24518386	+
ENSG00000257002.1	AP000438.2	chr11	62909546	62918361	+
ENSG00000255774.1	AP000439.3	chr11	69477133	69479940	-
ENSG00000229719.3	AP001187.9	chr11	64889560	64893449	-
ENSG00000236304.1	AP001189.4	chr11	76657056	76663866	+
ENSG00000236935.1	AP003774.1	chr11	64325050	64329504	-
ENSG00000223400.1	AP006748.1	chr21	41576135	41581319	-
ENSG00000249599.1	BMPR1B-AS1	chr4	94743800	94757533	-
ENSG00000203709.8	C1orf132	chr1	2.08E+08	2.08E+08	-
ENSG00000260581.1	CTB-113P19.4	chr5	1.52E+08	1.52E+08	+
ENSG00000253315.1	CTB-11I22.2	chr5	1.59E+08	1.59E+08	+
ENSG00000253768.1	CTB-33O18.1	chr5	1.74E+08	1.74E+08	+
ENSG00000269486.2	CTC-360G5.9	chr19	38935297	38938632	-
ENSG00000248268.1	CTC-499J9.1	chr5	1.11E+08	1.11E+08	-
ENSG00000251532.1	CTD-2245E15.3	chr5	1544107	1551710	-
ENSG00000268416.1	CTD-2626G11.2	chr19	20746923	20755250	-

ENSG00000259725.1	CTD-3032H12.1	chr16	54937786	54938671	-
ENSG00000273032.1	DGCR9	chr22	19017834	19020248	+
ENSG00000231651.1	DLG3-AS1	chrX	70452958	70455994	-
ENSG00000230316.5	FEZF1-AS1	chr7	1.22E+08	1.22E+08	+
ENSG00000197308.7	GATA3-AS1	chr10	8050450	8053484	-
ENSG00000266010.1	GATA6-AS1	chr18	22166898	22168968	-
ENSG00000258667.1	HIF1A-AS2	chr14	61715558	61751097	-
ENSG00000241388.4	HNF1A-AS1	chr12	1.21E+08	1.21E+08	-
ENSG00000272733.1	KB-208E9.1	chr22	23580880	23583859	-
ENSG00000135253.12	KCP	chr7	1.29E+08	1.29E+08	-
ENSG00000261399.1	LA16c-329F2.1	chr16	1713527	1714208	-
ENSG00000259840.1	LA16c-380A1.1	chr16	710746	711277	-
ENSG00000167117.7	LINC00483	chr17	50761029	50767557	-
ENSG00000248360.6	LINC00504	chr4	14470465	14888169	-
ENSG00000258955.1	LINC00519	chr14	51304416	51328386	-
ENSG00000213373.6	LINC00671	chr17	42874670	42898704	-
ENSG00000177133.9	LINC00982	chr1	3059615	3068437	-
ENSG00000224559.2	LINC01087	chr2	1.32E+08	1.32E+08	+
ENSG00000249601.2	LINC01187	chr5	1.7E+08	1.7E+08	-
ENSG00000244541.4	LINC01213	chr3	1.5E+08	1.5E+08	+
ENSG00000231210.2	LINC01510	chr7	1.17E+08	1.17E+08	-
ENSG00000253563.2	NKX2-1-AS1	chr14	36519278	36523016	+
ENSG00000152931.7	PART1	chr5	60487713	60547657	+
ENSG00000225937.1	PCA3	chr9	76764436	76787569	+
ENSG00000265369.3	PCAT18	chr18	26687621	26703638	-
ENSG00000255794.5	RMST	chr12	97431653	97565035	+
ENSG00000253508.1	RP1-170O19.14	chr7	27186573	27193448	-
ENSG00000224961.1	RP1-278O22.1	chr20	10753090	10753966	+
ENSG00000232412.1	RP1-315G1.3	chrX	1.24E+08	1.24E+08	-
ENSG00000269894.1	RP11-1020A11.1	chr3	9935706	9936258	+
ENSG00000258919.1	RP11-1029J19.4	chr14	1.02E+08	1.02E+08	-
ENSG00000214797.3	RP11-1036E20.9	chr11	59268876	59284033	-
ENSG00000273209.1	RP11-107N15.1	chr2	2.02E+08	2.02E+08	-
ENSG00000246640.1	RP11-1094H24.4	chr17	50050349	50055739	-
ENSG00000273001.1	RP11-118K6.3	chr10	3065424	3066001	-
ENSG00000251637.5	RP11-119D9.1	chr11	67886477	67906350	+
ENSG00000225472.1	RP11-120J1.1	chr9	14317085	14357908	+
ENSG00000224842.2	RP11-123K19.1	chr9	1.27E+08	1.27E+08	-

ENSG00000269707.1	RP11-13J10.1	chr2	1.05E+08	1.05E+08	+
ENSG00000232110.6	RP11-149I23.3	chr10	89283765	89292125	+
ENSG00000248554.1	RP11-159F24.6	chr5	43511058	43521811	+
ENSG00000255847.4	RP11-167N4.2	chr11	73963657	73970287	-
ENSG00000271850.1	RP11-16D22.2	chr13	34348043	34614170	+
ENSG00000234396.3	RP11-181G12.4	chr1	2212523	2220738	+
ENSG00000238102.1	RP11-19D2.1	chr20	7256580	7258214	-
ENSG00000271830.1	RP11-1C8.7	chr8	1.03E+08	1.03E+08	-
ENSG00000273066.4	RP11-216L13.19	chr9	1.37E+08	1.37E+08	+
ENSG00000255474.1	RP11-234B24.2	chr12	4700417	4720102	-
ENSG00000260618.1	RP11-23N2.4	chr15	52577842	52598709	+
ENSG00000174171.5	RP11-23P13.6	chr15	41892793	41898575	+
ENSG00000272205.1	RP11-277B15.3	chr1	1.67E+08	1.67E+08	-
ENSG00000255746.1	RP11-283I3.4	chr12	253442	257299	-
ENSG00000271996.1	RP11-337N6.1	chr2	1.77E+08	1.77E+08	+
ENSG00000258414.1	RP11-356O9.1	chr14	37564047	37579125	+
ENSG00000265408.1	RP11-361L15.4	chr16	66942712	66963256	+
ENSG00000271387.1	RP11-382D12.2	chr1	1.84E+08	1.84E+08	-
ENSG00000236066.4	RP11-389O22.1	chr1	1.13E+08	1.13E+08	+
ENSG00000267284.1	RP11-397A16.1	chr18	55721063	55788761	+
ENSG00000273248.1	RP11-399K21.13	chr10	75408973	75409326	-
ENSG00000259793.1	RP11-400N9.1	chr2	2.33E+08	2.33E+08	-
ENSG00000273388.1	RP11-401O9.4	chr17	10291820	10317926	+
ENSG00000273153.1	RP11-406H21.2	chr10	17137336	17137585	-
ENSG00000271631.1	RP11-408O19.5	chr9	1.13E+08	1.13E+08	+
ENSG00000214733.7	RP11-429J17.8	chr8	1.44E+08	1.44E+08	+
ENSG00000224251.5	RP11-499O7.7	chr10	4995488	4997380	+
ENSG00000251141.4	RP11-53O19.1	chr5	44744900	44808777	-
ENSG00000248779.1	RP11-53O19.2	chr5	44752949	44765744	+
ENSG00000227947.1	RP11-543D5.1	chr1	47688463	47703383	+
ENSG00000248429.4	RP11-597D13.9	chr4	1.58E+08	1.58E+08	+
ENSG00000263427.1	RP11-599B13.3	chr17	8056225	8057621	-
ENSG00000250740.1	RP11-710F7.2	chr4	1.06E+08	1.06E+08	-
ENSG00000254528.6	RP11-728F11.4	chr11	1.18E+08	1.18E+08	+
ENSG00000259367.1	RP11-815J21.4	chr15	85619623	85670948	-
ENSG00000266441.1	RP11-91I8.3	chr18	6728821	6729862	-
ENSG00000259887.1	RP11-923I11.5	chr12	51848223	51852729	+
ENSG00000250643.1	RP11-93K22.6	chr3	1.3E+08	1.3E+08	-
ENSG00000269489.1	RP11-98D18.17	chr1	1.52E+08	1.52E+08	+
ENSG00000254872.3	RP13-870H17.3	chr11	1049880	1055749	+

ENSG00000227066.1	RP3-340N1.2	chr1	20154338	20160568	+
ENSG00000231628.1	RP3-355L5.4	chr6	1.05E+08	1.05E+08	+
ENSG00000255202.1	RP4-541C22.5	chr11	33665220	33696701	-
ENSG00000261786.1	RP4-555D20.2	chr3	44117299	44122365	+
ENSG00000258586.1	RP5-1021I20.2	chr14	73822559	73830135	-
ENSG00000231566.1	RP5-1158E12.3	chrX	45848074	45851490	-
ENSG00000236772.1	RP5-1184F4.5	chr20	32449755	32453607	+
ENSG00000226812.2	RP5-881L22.5	chr20	44347552	44355185	-
ENSG00000234184.4	RP5-887A10.1	chr1	80535755	80646788	+
ENSG00000229591.1	RP5-981O7.2	chr7	1.52E+08	1.52E+08	-
ENSG00000233705.5	SLC26A4-AS1	chr7	1.08E+08	1.08E+08	-
ENSG00000232803.1	SLCO4A1-AS1	chr20	62663019	62666724	-
ENSG00000242808.6	SOX2-OT	chr3	1.81E+08	1.82E+08	+
ENSG00000227640.2	SOX21-AS1	chr13	94712716	94716246	+
ENSG00000232504.4	ST3GAL5-AS1	chr2	85889280	85890980	+
ENSG00000224490.4	TTC21B-AS1	chr2	1.66E+08	1.66E+08	+

8.8.4 Appendix B.4: Survival Analysis (Forest Plots)

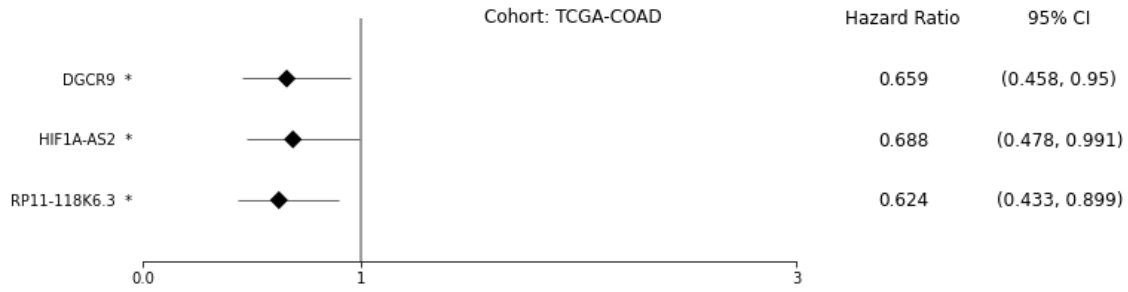

Figure S11: Forest plot of survival analysis with significant lncRNAs on TCGA-COAD cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

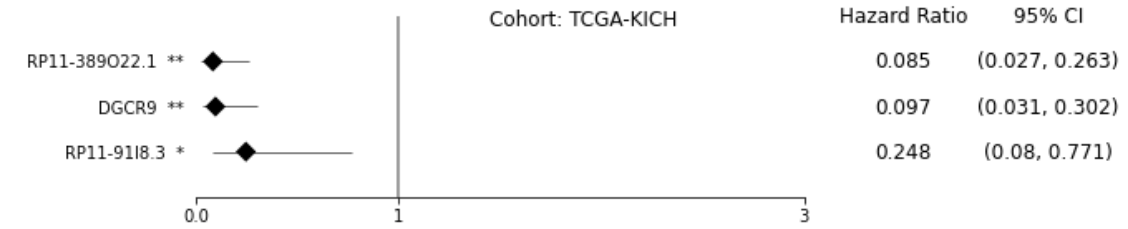

Figure S12: Forest plot of survival analysis with significant lncRNAs on TCGA-KICH cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

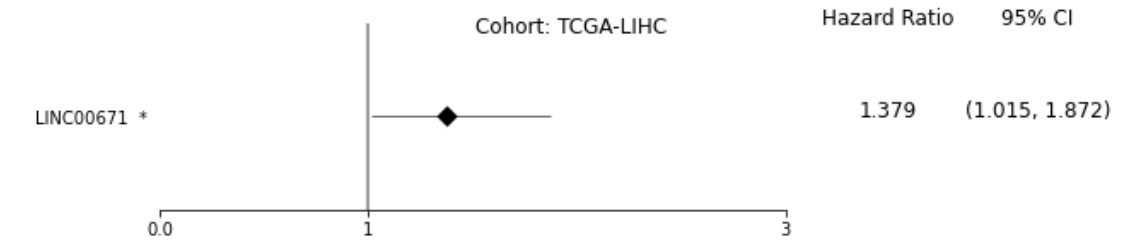

Figure S13: Forest plot of survival analysis with significant lncRNAs on TCGA-LIHC cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

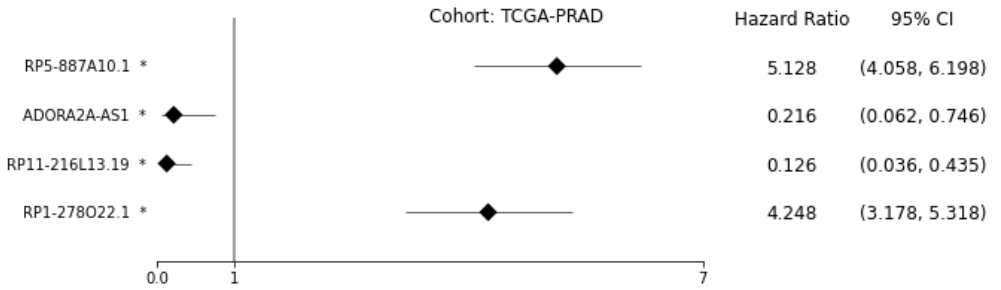

Figure S14: Forest plot of survival analysis with significant lncRNAs on TCGA-PRAD cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

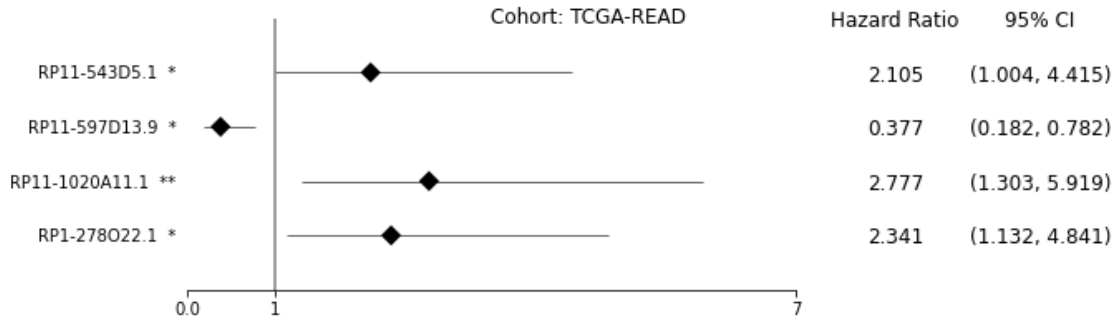

Figure S15: Forest plot of survival analysis with significant lncRNAs on TCGA-READ cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

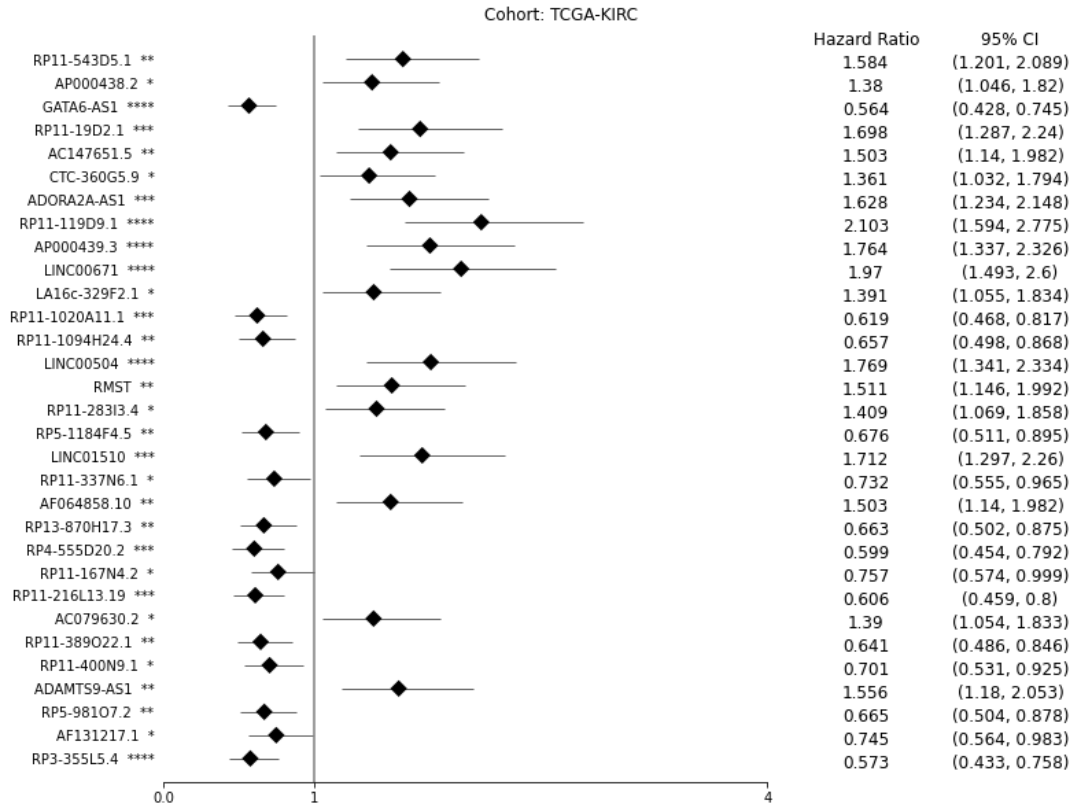

Figure S16: Forest plot of survival analysis with significant lncRNAs on TCGA-KIRC cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

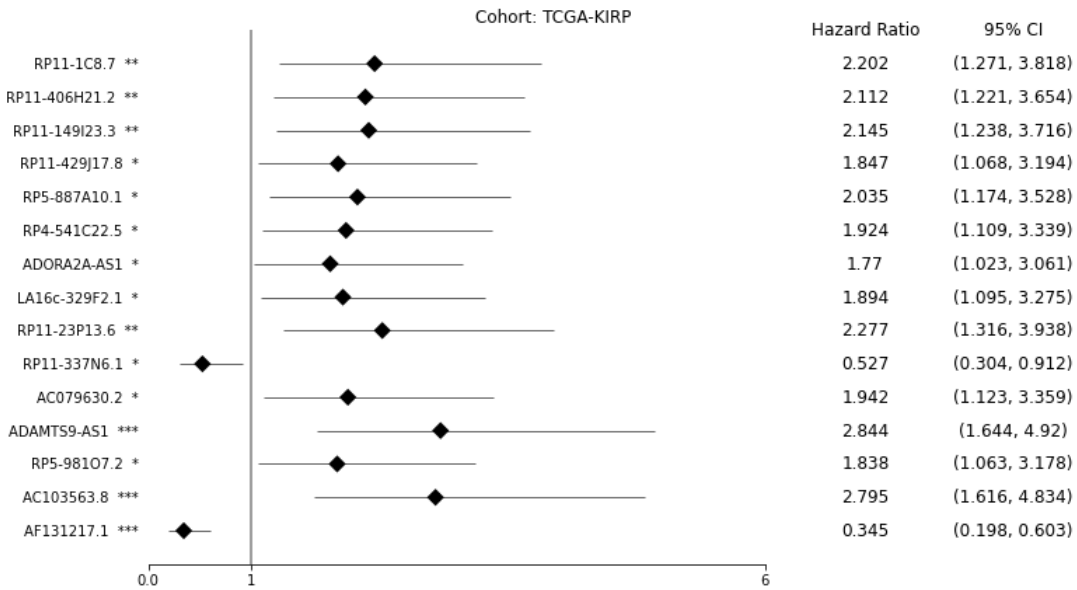

Figure S17: Forest plot of survival analysis with significant lncRNAs on TCGA-KIRP cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

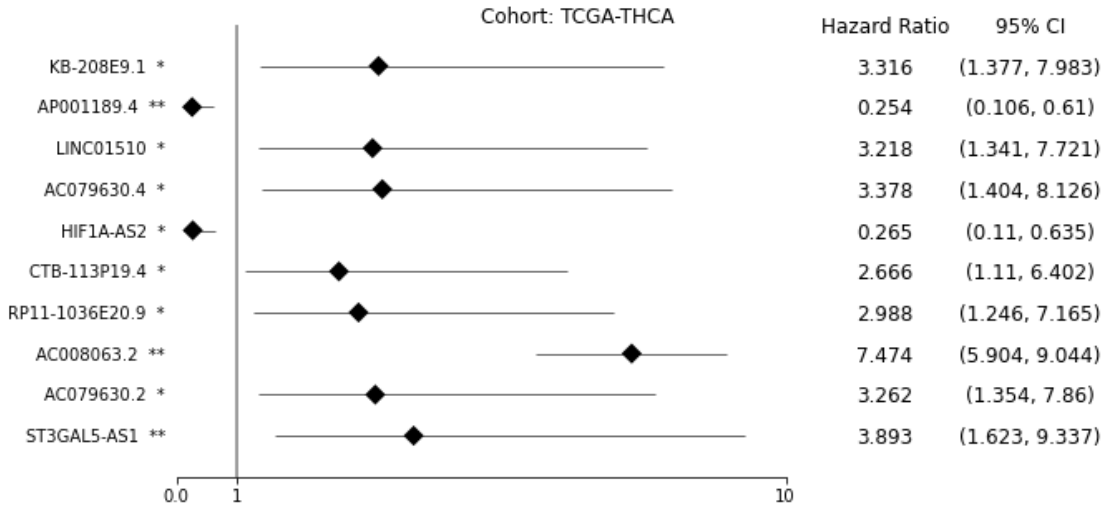

Figure S18: Forest plot of survival analysis with significant lncRNAs on TCGA-THCA cohort. The asterisks represent the Log-rank P-values: (* - $P \leq 0.05$, ** - $P \leq 0.01$, *** - $P \leq 0.001$, **** - $P \leq 0.0001$)

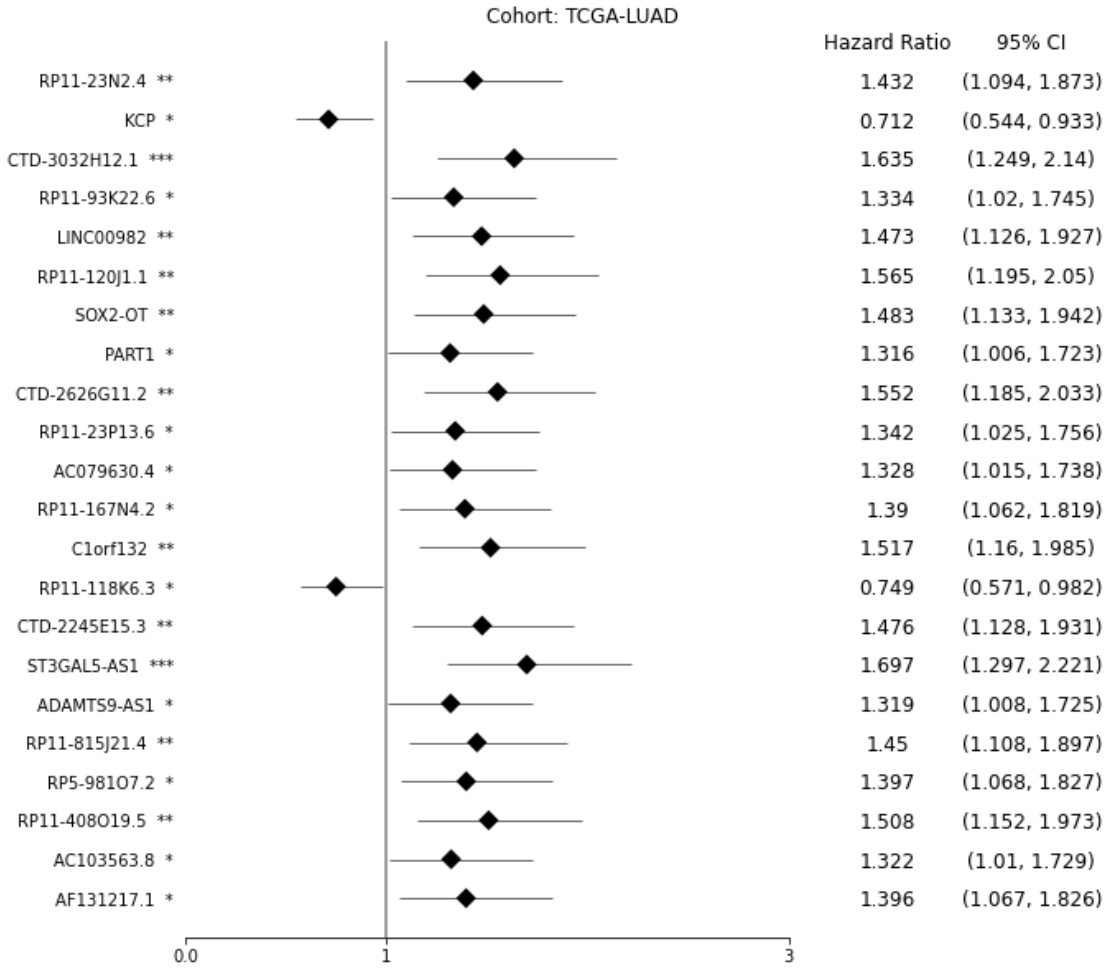

Figure S19: Forest plot of survival analysis with significant lncRNAs on TCGA-LUAD cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

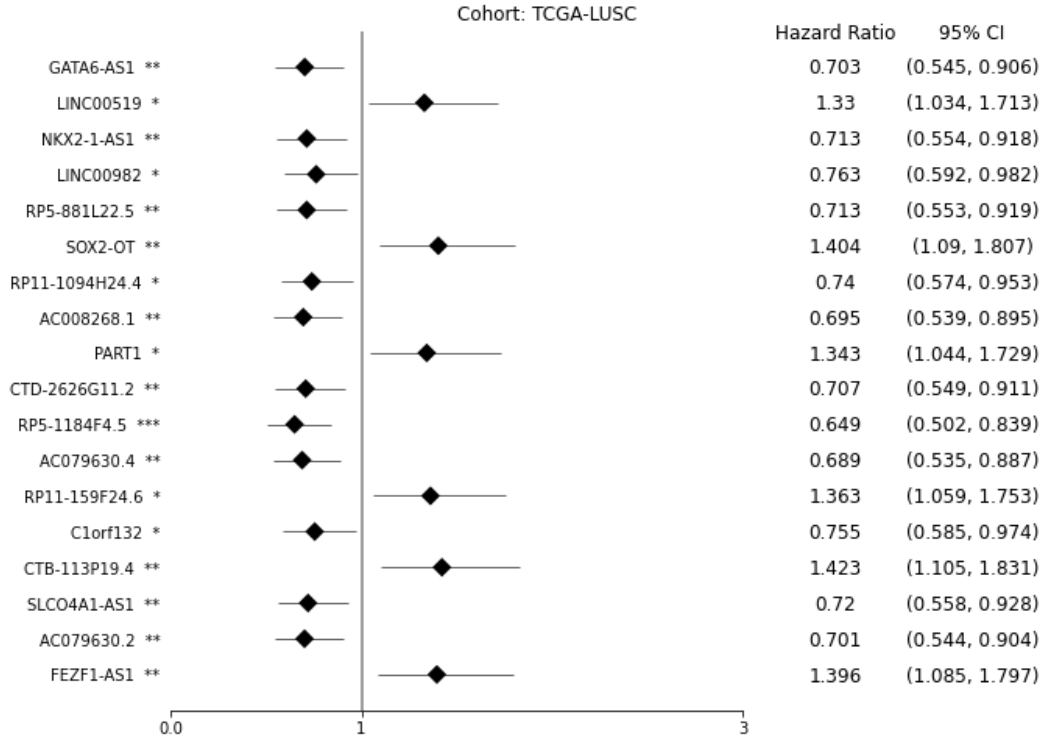

Figure S20: Forest plot of survival analysis with significant lncRNAs on TCGA-LUSC cohort. The asterisks represent the Log-rank P-values: (* - P ≤ 0.05, ** - P ≤ 0.01, *** - P ≤ 0.001, **** - P ≤ 0.0001)

VITA

MD ABDULLAH AL MAMUN

2008-2012	B.S., Computer Science and Engineering Dhaka University of Engineering and Technology Bangladesh
2014-2017	M.S., Computer Engineering King Fahd University of Petroleum and Minerals Saudi Arabia
2017-2018	Teaching Assistant Qatar University Doha, Qatar
2018-2022	Graduate Assistant Florida International University Miami, Florid, USA

PUBLICATIONS AND PRESENTATIONS

Peer-reviewed Journal Articles

Al Mamun, A., Tanvir, R.B., Sobhan, M., Mathee, K., Narasimhan, G., Holt, G.E. and Mondal, A.M., 2021. Multi-run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers. International Journal of Molecular Sciences, 22(21), p.11919. (Impact Factor 5.92)

Tanvir, R.B., Aqila, T., Maharjan, M., Mamun, A.A. and Mondal, A.M., 2019. Graph theoretic and pearson correlation-based discovery of network biomarkers for cancer. Data, 4(2), p.81. (Impact Factor 3.50)

Al Mamun, A., Mohammad Tariq Nasir, and Ahmad Khayyat. ``Embedded System for Motion Control of an Omnidirectional Mobile Robot." IEEE Access, vol. 6, pp. 6722-6739, Jan 2018. (Impact Factor 3.36)

Peer-reviewed Conferences Proceedings

Al Mamun, A., M. Sobhan, R. B. Tanvir, C. J. Dimitroff and A. M. Mondal,``Deep Learning to Discover Cancer Glycome Genes Signifying the Origins of Cancer," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 2425-2431, doi: 10.1109/BIBM49941.2020.9313450.

Al Mamun, A., W. Duan and A. M. Mondal, ``Pan-cancer Feature Selection and Classification Reveals Important Long Non-coding RNAs," 2020 IEEE International

Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 2417-2424, doi: 10.1109/BIBM49941.2020.9313332.

M. Sobhan, A. A. Mamun, R. B. Tanvir, M. J. Alfonso, P. Valle and A. M. Mondal, "Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020, pp. 2990-2992, doi: 10.1109/BIBM49941.2020.9313426.

Al Mamun, A., and Ananda Mohan Mondal. "Long non-coding RNA based cancer classification using deep neural networks." Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2019.

Al Mamun, A. and A. M. Mondal, "Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 2825-2831, doi: 10.1109/BIBM47256.2019.8983413.

T. Aqila, A. A. Mamun, and A. M. Mondal, ``Pseudotime Based Discovery of Breast Cancer Heterogeneity," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 2049-2054, doi: 10.1109/BIBM47256.2019.8983300.

M. H. Al-Meer and A. A. Mamun ``Deep Learning in Classifying Sleep Stages," 2018 Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, 2018, pp. 12-17.

AlSaad, Rawan, Somaya Al-Máadeed, A. A. Mamun, and Sabri Boughorbel. ``A Deep Learning Based Automatic Severity Detector for Diabetic Retinopathy." In International Conference on Machine Learning and Data Mining in Pattern Recognition, pp. 64-76. Springer, Cham, 2018.

Al Mamun, A., Abdullah Al Mamun, and Abdullatif Shikfa. ``Challenges and Mitigation of Cyber Threat in Automated Vehicle: An Integrated Approach." In 2018 International Conference of Electrical and Electronic Technologies for Automotive, pp. 1-6. IEEE, 2018.

A. Al Mamun, K. Salah, S. Al-maadeed and T. R. Sheltami, "BigCrypt for big data encryption," 2017 Fourth International Conference on Software Defined Systems (SDS), 2017, pp. 93-99, doi: 10.1109/SDS.2017.7939147.

Al Mamun, A., Fahim Djatmiko, and Mridul Kanti Das. ``Binary multi-objective PSO and GA for adding new features into an existing product line." In 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 581-585. IEEE, 2016.

Almadani, B., A. A. Mamun, and Khayyat, A., 2015. Real-Time QoS-Aware Vehicle Tracking: An Experimental and Comparative Study. Procedia Computer Science, 56, pp.349-356.