Analysis for RNA based liquid biopsy for various cancers

by
Vikash Dabas

Under the supervision of
Dr. Vibhor Kumar

Center for Computational Biology Indraprastha Institute of Information Technology - Delhi May, 2022

Analysis for RNA based liquid biopsy for various cancers

by
Vikash Dabas

Under the supervision of
Dr. Vibhor Kumar

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, Computational Biology

Center for Computational Biology
Indraprastha Institute of Information Technology - Delhi May, 2022

Certificate

This is to certify that the thesis titled *"Analysis for RNA based liquid biopsy for various cancers"* being submitted by **Vikash Dabas** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May,2022

Dr Vibhor Kumar
Department of ComputationalBiology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

For this entire work, I would like to thank my supervisor Dr. Vibhor Kumar. His advice has helped make this work reach unprecedented heights and has also molded me into a critical thinker. I also admire him for the patience that he has shown while working with me, considering the Covid scenario.

Also, whenever I felt low or stuck with some error, my IIITD batch mates always came forward to help and provide me with good advice. I would also like to extend my sincere gratitude to my teachers, seniors, and brilliant minds working at IIIT-D. I would also like to thank the Department of Biotechnology, the Government of India, for the student fellowship and support to the MTech (Computational Biology) program.

I am also grateful to every research freely available with the data like the TCGA and GTEX databases; without their data, the work could not be done.I also would like to express my gratitude to the COVID warriors that served our nation in time of need; without them, the situation could not have been controlled. Also, I am very grateful to my family, who have always supported me in every decision in my life.

Abstract

Due to the advent of next-generation sequencing techniques, the sequencing cost is reduced with increasing accuracy. We can use this as an advantage in liquid biopsy to identify and monitor cancer at different stages, which has many advantages like being non-invasive and help in disease monitoring at different stages over conventional techniques like tissue extraction. We can use the RNA detected in the liquid biopsy and can help in the prognosis and diagnosis of cancer.

The challenge with the liquid biopsy is that the data is minimal and not publicly accessible. To overcome this, we have tried to simulate liquid biopsy data using publicly available TCGA pan-cancer and GTEX whole blood data at different dilutions. Later this simulated data was used to train the machine learning models, and then we identified the best markers present in six types of cancers.

We have identified multiple marker genes from six type of simulated cancer and enhancer RNA in BRCA.validated on already available markers and literature. Further analysis of these markers will give us more insights. We believe that our work can help in designing cancer gene panels for the detection and prognosis of cancer .

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For the past few decades, cancer has been detected using very invasive methods that are very painful or may have harmful effects on the body. However, in today's world, when informatics has progressed, the concept of personalized medicine or treatment is getting popular. Therefore, it is necessary to introduce novel ways to cut down the diagnosis cost of cancer and provide reliable results to the patient.

This problem can be solved by taking the blood samples from the patient and identifying the changes that occur in the patient blood when compared with an average healthy person. However, problem is that the data is scarce and only available for some conditions. We have tried to solve this problem by simulating liquid biopsy data from normal blood samples mixed with cancer and the results are promising.

In this thesis, we are generalizing for different types of RNA that are present in the blood to find the best potential biomarkers found in the liquid biopsies that include mRNAs,miRNAs,snRNAs,snoRNAs, and lncRNAs from the simulated data from TGCA and GTEX databases.

1.1 Division of thesis

The thesis is divided into four chapters. The first chapter is the introduction that gives a brief overview of how the liquid biopsy can be helpful for the cancer prognosis and diagnosis. The second chapter explains the methodology that we have used to simulate the blood data from various databases and the machine learning techniques that are used to find potential markers for six different types of cancer.

Chapter three describes the relevance and inferences from the results obtained, and we have used literature and other resources to validate the results. Fourth chapter is the conclusion that lists the functional and mechanistic insights based on the results obtained.

1.2 Liquid Biopsy

Liquid biopsy is a technique in which the blood is extracted from the patient body which includes cell DNA and RNA along with the proteomic data[1]. Liquid biopsy data is widely used for prognosis and screening of multiple cancer to provide an personalized treatment to the patients. Although tissue biopsy is considered best for tumor screening, it also poses risks like being invasive, hard to reach some locations, and post-operative complications[2] that may occur in patients. Also, tumors are a heterogeneous mixture of cells so, they undergo various genetic and epigenetic changes rapidly, so it is not practical to perform the tissue biopsy frequently.

These problems can be fixed by using the liquid biopsy samples as they can be frequently sampled and continuously monitored for disease progression[3].Considering the advantages of liquid biopsy over the traditional tissue biopsy multiple studies have been conducted for the liquid biopsy taking circulating tumor DNA[4][5],RNA and other fluids for providing generic[6],proteomic,transcriptomic and epigenetic[7] insights.

1.2.1 RNA based liquid biopsy

This thesis will focus mainly on RNA for the prognosis and diagnosis of cancers. Different types of RNA molecules are present in the liquid biopsy sample. Also, searching for the biomarkers in the extracellular body fluids gives a better treatment plan for the patient. There are freely circulating

RNA molecules in the bodily fluid or packaged in the microvesicles. As a result, they are highly stable and can be stored and sampled several times.

Both coding and noncoding RNA species can be used as cancer markers from liquid biopsies. However, the most popular among all RNAs are the microRNAs which are classified as biomarkers for multiple cancers. Nowadays, multiple studies also include other types of RNAs, including Piwi-interacting RNAs, small noncoding RNAs, and circular RNAs that can also be used for the prognosis and prediction of cancer.[8]

Cirulating transcriptome

The Cirulating transcriptome is the different RNA that are present in blood plasma which is released as metabolites by the body as a result of different processes that occurs in the cells. The major part of the transcriptome consists of noncoding RNA, which has a variety of practical uses, which including ncRNAs,miRNAs,tiRNAs,snoRNAs, and large ncRNAs together. They are commonly abbreviated as circulation transcriptome. The circulating cell-free RNA was firstly found in Epstein Barr virus in the plasma of cancer patient[9]. Very few of the cfRNAs are stable and a large number of cfRNA are degraded in blood by RNAse[10].

MiRNA in liquid biopsy

Micro RNA (miRNA) are small molecules with 20-22 nucleotides of length which can act on post transcription processes to facilitate the protein coding gene expressions by activation binding on target genes[11]. Recently discovered miRNA are very stable when compared with the cfRNA they are stable even after freeze and thaw cycles and also are ph insensitive, which makes them ideal for cancer biomarkers[12]. These miRNA are found to have a tissue-specific expression pattern and also in various diseases, including cancers[13]. Micro RNA can be obtained from different extracellular fluids like plasma [14], saliva[15] and other body fluids as well.

The miRNAs can target cells as they are free circulating molecules associated with different proteins or can be present in vesicles like exosomes and apoptotic bodies, which can act like signaling molecules for various cellular activities in cancers[16]. The amount of research done for miRNAs is very large due to their stability; various miRNAs are discovered in literature, which can be used as biomarkers. The miRNAs miR-155 and miR-210

were first reported in 2008 as DLBCL cancer biomarkers[17]. Many studies show the miRNAs are dysregulated and are found in multiple cancer types, leading to behave like oncogenes(tumor-causing) or tumor-suppressing genes. Some of the miRNAs are highly specific to one type of tumor; for example, miR-127-3p was found only in gynecologic cancers, including breast and ovarian body fluids[18]. Whereas some are found in large no of tumor samples like miR-21-5p and know oncogene was found in 20 cancers body fluids and tissue[4].In another study they find prognostic markers miR-25-3p,mir-130a and mir-27a for the colorectal cancer [19]and miR-486-5p, miR-938 for the pancreatic cancer[3].

SnRNA and SnoRNAs in liquid biopsy

Small nucleolar RNAs are a type of noncoding RNAs which are omnipresent in the body and are involved in the modification and processing of rRNA(Ribosomal RNA). Small nuclear and small nucleolar RNAs are a part of ncRNAs; their length varies between 50-400 nucleotide bases, and they can be transcribed for the introns region present in coding genes[20]. When the spliceosome binds with the precursor mRNA and specifically cleaves the intronic regions to facilitate creating of mature mRNA, snRNAs are present in RNA - RNA complex and RNA - Protein complex, which make sure the precursor mRNA binds correctly with the spliceosome complex. Therefore changes in the structure of snRNA will have functional consequences and can make the gene oncogenic[21].

Studies reported that snRNA which are present in liquid biopsies can act as cancer biomarkers. One of the snRNAs that is found predominantly is U2 which is found in the body fluids. There are altered expression patterns in parts of RNU2 which have been reported in plasma of colorectal, ovarian, and lung cancer[22][23]. For pancreatic cancer, SNORA74A, SNORA25[24], RNU2[22] are reported as promising biomarkers which were able to distinguish between cancer and control samples. Also, SnoRD33 and snoRD76 are found upregulated in the plasma of epithelial lung cancer[25].

LncRNAs in liquid biopsy

Long noncoding RNAs are noncoding RNAs with more than 200 nucleotides in length that can also be involved in a covalently bonded circular form called ac circRNAs[26]. There are more than 200,000 lncRNA that have

been identified, and they play a crucial part in the splicing of RNA and post transcriptions processing[27].

One of the most studied lncRNA is MALAT1 involved in metastasis of lung adenocarcinoma transcript[28]. A later study also found high expression of lncRNAs which include HOXAS2, HYMA1, and HOTAIR in the serum and LINC00455 and HUMA1 in the urinary exosome[28].In another study on the plasma of esophageal squamous cell carcinoma, lncRNAs POU3f3, SPRY4-IT1, and HNF1A-AS1 were found to be highly upregulated, which can be used as bio markers[29].

Enhancer RNA

Enhancer RNA(eRNA) transcribed to form the enhancer region is a noncoding RNA. Enhancer RNAs are shown to have essential roles in mediating the transcriptional cycle for activation of genes[30]. They are part of long noncoding RNAs.

Also, in the case of cancer, they are involved in the activation of oncogenic pathways. For example, the ESR1 activation is shown to increase the eRNA transcription in breast tumor[31]. It can also be induced in tumor suppressors; like when TP53, which is a tumor-suppressing gene, is induced, it increases eRNAs, which leads to activation of pathways that cause cell cycle arrest in different cancer cells lines[32].

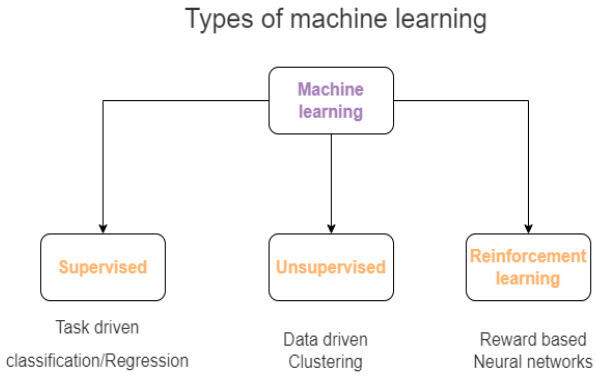

Figure 1.1: Types of machine learning

1.3 Machine learning

Machine learning is a very powerful technique that gives the ability to the computer to learn without being explicitly programmed. Machine learning and data science is the most booming field in the biological domain due to the enormous available data sets. Machine learning works by building an algorithm for the given problem statement. It uses training data for the learning and then finds patterns or statistical relevance in the data and uses that to predict the output in the testing data set. Machine learning is also used extensively in the field of bio-informatics, like cancer detection, drug discovery, vaccine modeling, and many more.

Machine learning is divided mainly into three categories supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, we provide the model with the labels that classify data as positive or negative; then, the model is built, and a prediction is made. In the case of unsupervised learning, labels are not provided, so the machine has to cluster the data into groups and find patterns to make predictions. Reinforcement learning is similar to how our brain works that is reward-based machine learning, so in this case, the labels are not provided for all the inputs and model trains in order to maximize the reward that is correct prediction. In this thesis work, I have primarily used the supervised learning models with the label provided. The model we have used is mainly for binary classification, including Support vector machine, logistic regression, random forest, and K-Nearest neighbor.

1.3.1 Selected Models

1.3.2 Support vector machines

SVM is a type of supervised learning technique. Support vector machine (SVM) is one of the methods used for binary or multiclass classification, which classifies by defining separating hyperplanes in the dataset. A hyperplane can be considered a line in different spaces separating the input. Moreover, if we replace the log loss with some loos function, the solution will be a sparse vector, and the predictions on the training subset will be called support vectors. When we combine the support vectors with the kernel trick and a customized loss function, we have SVM[33]. SVM classifies based on the decision plane separating the data into different classes. We try to maximize the margins of the decision planes so that the classification can be done correctly. We have used the RBF kernel for our model training and testing combined with 5-fold cross-validation.

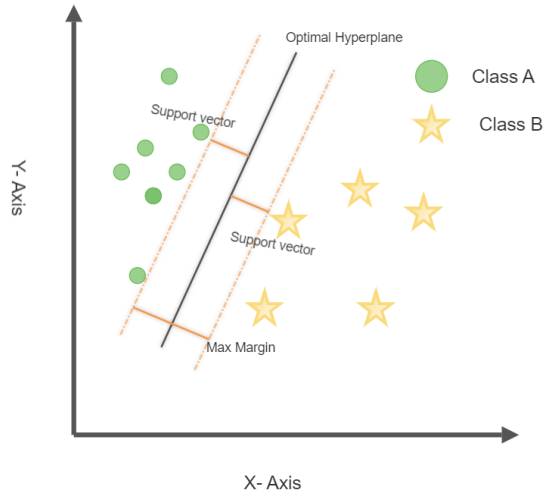

Figure 1.2: SVM Model

1.3.3 Logistic Regression

Logistic regression is also an example of supervised learning. It is a type of classification model rather than a regression model. Logistic regression is majorly used for the binary classification model, which gives good accuracy, although it can also be used in the multi class-classification. Logistic regression uses a logistic function to model binary output data ranging between 0 and 1[34]. For our work, we have used the limited memory Royden Fletcher Goldfarb Shanno algorithm(lbfgs) for the logistic regression solver, which works well for large datasets[35].

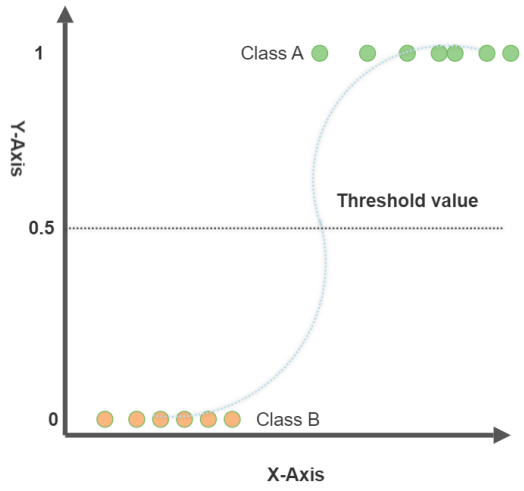

Figure 1.3: A simple logistic regression model

1.3.4 K-Nearnest Neighbors

KNN is a type of lazy learning that combines both classifications and regression[36]. This works by searching the entire training set for new instances of K's closest instance. Then the distance is calculated between the new instances also; different weights are given depending upon the closeness. Fewer weights are assigned to neighbors who are distant.KNN classification uses k as input and classifies the unlabeled vector depending upon the frequency and distance of the neighbor. KNN can be used for either binary or multiclass classification. The main drawback of this algorithm is the computation power it draws as it uses the brute force technique and tries all pairs with different combinations. We used the KNN with ten neighbors, which gave us optimal results. Below shown is a two-class classification with the closet neighbor marked in hyperplane represented by the dotted line.

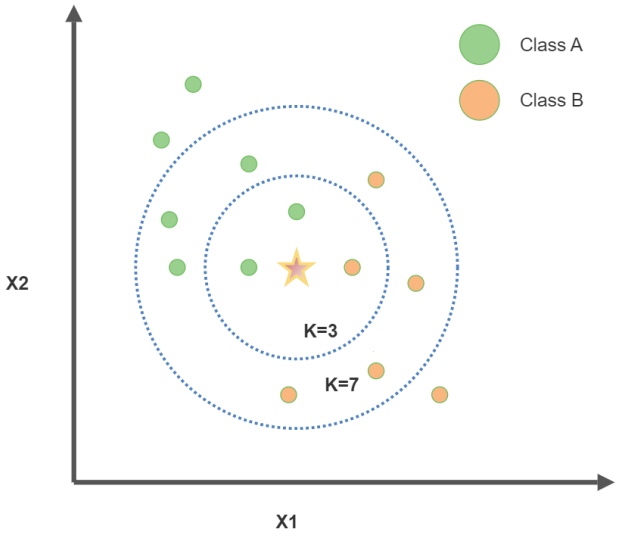

Figure 1.4: A KNN represent

1.3.5 Random Forest

A random forest is a combination of multiple decision trees, an excellent classification technique. The random forest has a start node as a root, and every node is defined with its conditions at the branches. Random forest is also an example of a supervised learning technique. In this, the data is divided into multiple smaller datasets, and random sets of features are used to build the decision tree depending upon the number of trees we are using[37]. Furthermore, whenever new data needs to be tested, the random forest uses the best sets of instructions made using decision trees and makes a prediction. A random forest is a type of ensemble model which combines a decision tree and makes the algorithm better. The run time complexity of the random forest depends upon the number of trees used for our study; we have used up to 10 decision trees.

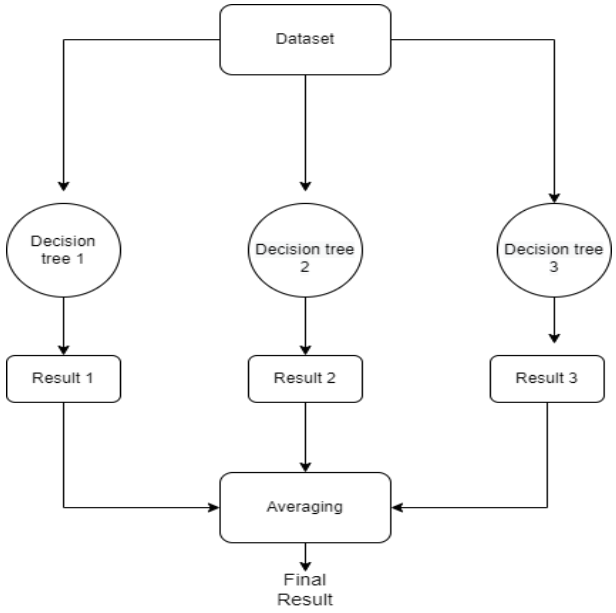

Figure 1.5: A simple Random forest example

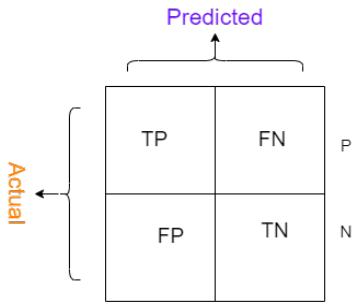

Figure 1.6: Confusion Matrix

1.3.6 Evaluation Metrics

After the training and predictions done by the model, evaluation of the model performance is very important to improve the algorithm or for parameter tuning. So first, data is split into train data set and a testing data set, then the model is trained, and predictions are made afterward. We compare the predicted labels with the actual levels and create evaluation metrics with critical elements like precision, recall, false positives, true positives, false negatives, and true negatives. Using the confusion matrix, which comprises four different combinations of actual vs. predicted values, precision, recall, and accuracy are calculated.

Terminologies used in confusion matrix and evaluation of the model.
True Positive(TP): These are the values that the machine learning model predicted as positive, and they belong to the positive class.
True Negative(TN): These are the values that the machine learning model has predicted as negatives, and they belong to the negative class.
False Positive(FP): These are values that the machine learning model has predicted to be positive, but they belong to the negative class.
False Negative(FN): These are the value that the machine learning model has predicted as negative, but they belong to the positive class.
Precision: Precision is the ratio of true positives with the total no of positives predicted.
Recall: Recall is the ratio between the accurate positive predictions and all of the predicted results.
F1 score: F1 score is the harmonic mean between recall and precision.
Accuracy: Accuracy is the ratio between correctly predicted values and all values present in the data.

1.4 Challenges

Although the liquid biopsy method for the prognosis ands diagnois of cancer has its merits, it also has some challenges. The primary is the availability of freely accessible data for any cancer. The data is not available publicly, so small-scale researchers with limited resources cannot have the data for further analysis.

The second challenge is validating the bio-markers that have been detected for the prognosis or diagnosis of cancer.To overcome this, there is a need to develop kits of few markers that will be readily available for some cancer which can directly work on the liquid biopsy extracted RNA for cancer detection.We believe that the work we have done can help with this problem.

Chapter 2

Methodology

2.1 Data

In our study, we have used different data sets, including data from TCGA and GTEX databases. For cancer data, the NCBI GEO (BioSample: SAMN03164102; GEO: GSM1536837) Present in the humans, which contains 9264 samples belonging to the different tumors, they are covered under bio project PRJNA266377, which can be accessible publicly[38]. This data set has TCGA Pan-Cancer expression values which are used for differential gene analysis, data simulation, and machine learning in this study.

GTEx data released in public in version 8, namely GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2) is used to extract whole blood data. We have used the RNA-seq data containing gene read counts (GTEx Analysis 2017-06-05v8 RNASeQCv1.1.9 gene reads. gct)[39]. It contains more than 20,000 Genes and 16,000 samples from every part of the body. We extracted the whole blood samples from this dataset and used them for data simulation and machine learning in our study.

For enhancer RNA(eRNA), data is taken from two datasets which are TCGA eRNA data eRiC(Enhancer RNA in cancers)[40], and for the normal blood expression for eRNA, we have used HeRA(Human enhancers RNA atlas) [41]database to get the eRNA expression matrix.These datasets is used to find the most important features after removing the whole blood genes from the HeRA database with the TCGA eRNA data described below in detail.

2.2 Data processing

2.2.1 TCGA RNA

As a part of the analysis in the biological fields, we have to deal with very large datasets where information is in abundance, so there is a need to get the relevant information from the dataset for our usage. Therefore a lot of pre-processing needs to be done to tailor the data to our requirements. For the processing of the data R[42] and Python was used extensively. Steps to process the data are briefly explained below .

The TCGA pan-cancer expression values[38] contained data from 24 different cancers. Here 10,746 samples were present, out of which 1,482 samples were normal, and the rest 9,264 samples contained cancer expression data. In total, 22,000 genes expression is present in this data. We first take the data into the R and have used the package dplyr[43] for processing along with the basic R commands. We first transposed the dataset and added the labels for machine learning. The cancer samples were assigned 1, and the normal samples were assigned 0. After processing and using differential gene analysis, we have dropped 16,000 features (genes) as they were not significantly expressed in the dataset. As the data is a pan-cancer data which contains all of the samples from different cancer in a single file, We first identify the TCGA barcode and split the data into different cancer according to the sample ids.

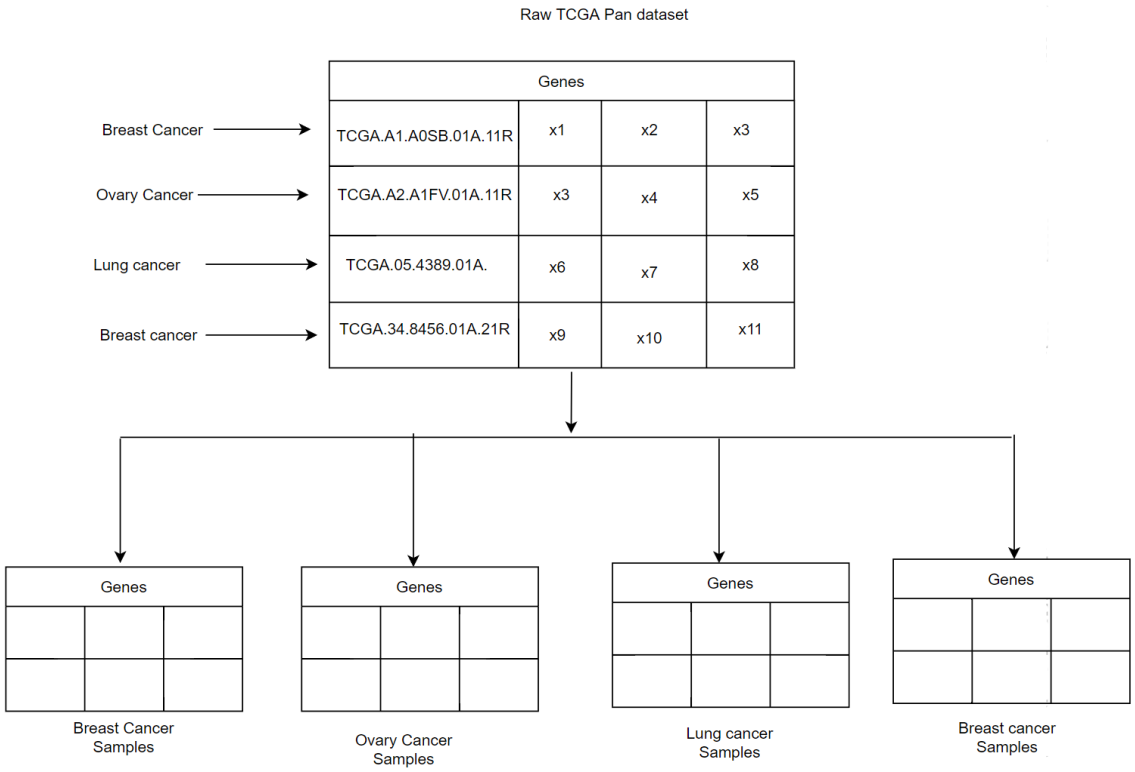

Figure 2.1: TCGA Data splitting cancer wise

2.2.2 GTEx RNA

GTEx RNA expression data [39] which had over 20,000 genes and 16,000 samples. In this, we only needed the blood expression data, we split the large file, which was approx 3 GB, into three smaller files so that they can be processed in R. Then, we have selected all the genes with the whole blood expression by using the annotation file. Then the ids which we got were intersected with the GTEx RNA expression, and blood expression data were extracted using R and python. Out of 16,000 samples, only 755 samples were selected, which had the whole blood RNA expression. Then the labels are added for machine learning and further processing.

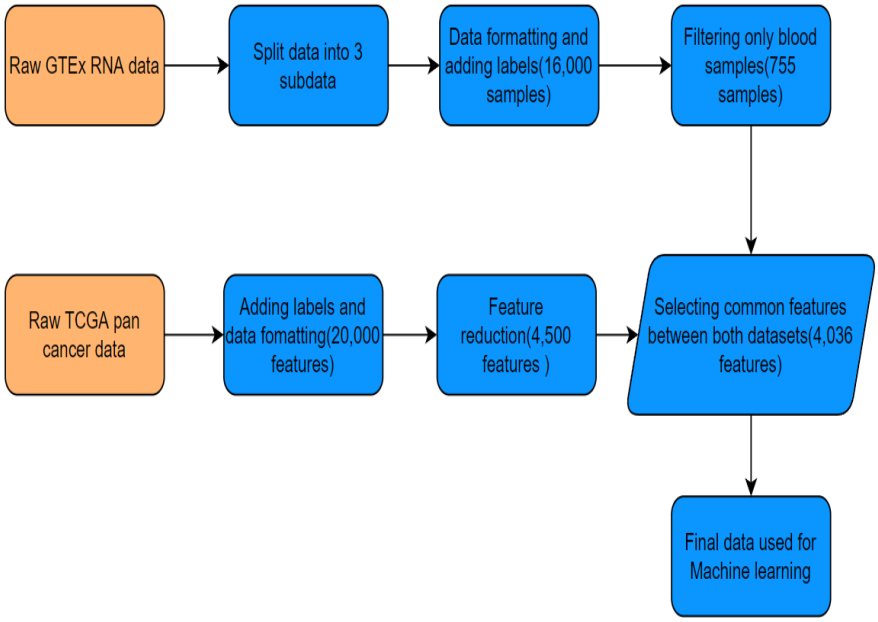

Figure 2.2: TCGA and GTEx RNA data processing

As we want to work with TCGA and GTEx RNA read counts data simultaneously, we had to make sure that the features which are genes are arranged in the correct order in both data sets.TCGA data, after filtering, had 4,500 genes, and the GTEx data had 20,000 genes, so we intersected the two data with R and have used only those genes which are shared between the both that lead to 4,036 features in the both TCGA and GTEx RNA dataset in the same order.

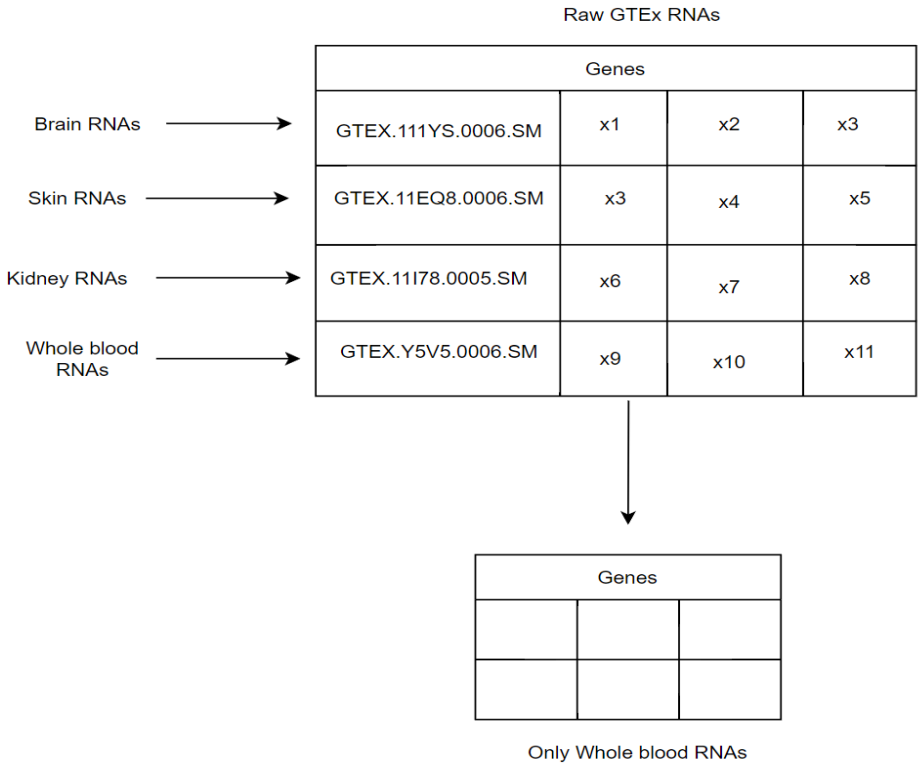

Figure 2.3: GTEx Whole blood samples extraction

2.2.3 Enhancer RNA

Two datasets are used for the enhancer RNAs. The TCGA eRNA expression data, which contains the cancer samples and normal samples for 24 types of cancer, and another dataset which contains the normal eRNA expression in different body parts. The TCGA eRNA expression data for cancers can be downloaded from eRiC database[44]. The other data which contains eRNA expression in normal body regions is downloaded from HeRA database[45].

We have taken the whole blood data from the HeRA and removed all of the common genes between HeRA and the TCGA eRNA to find the markers that will be present in cancer. The limitation of this dataset is that the number of normal samples was very low in the TCGA eRNA cancers. For example, in bladder cancer, only four samples were normal, and 430 samples were cancerous, so it is impossible to use conventional machine learning techniques.So we have to run the machine learning models on the limited can-

cers.Afterward labels are added to the samples id depending upon cancer(1) or normal(0) using R.

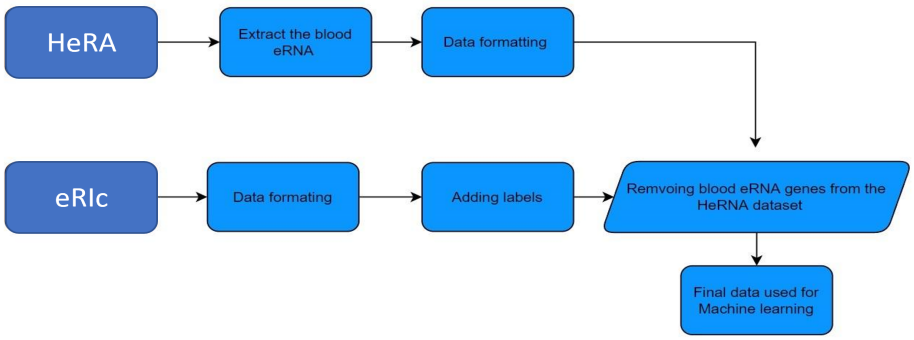

Figure 2.4: Enhancer RNA data processing

2.3 Enhancer RNA machine learning

For the Enhancer RNA we have done data processing for all 20 types of cancers that was present in the eRic database[45]. After doing the data processing, there were very few normal samples to run the machine learning models on the data. To overcome this, we used different techniques to fix the data imbalance, which included oversampling from the smaller data set, SMOTE[46] and undersampling from the larger dataset. However, these techniques to fix the imbalance also have their limitation like if there are no normal samples at all, these techniques will not work; for example, the cancers ACC, DLBC, LGG, MESO, OV, SKCM, TGCT, UCA, UVM had no normal samples at all, so these cancers were dropped. Other cancers, which include BLCA, Cesc, CHOL, COAD, GBM, HNSC, KICH, KIRC, LUAD,PRAD, and THCA, had very few normal samples for machine learning models to train and run.

Therefore, we are left with BRCA, which had 100 normal samples and 1250 cancerous samples on which we run the machine learning models. We could not run on the direct data as there is a class imbalance. Hence, we split the data into eight smaller datasets, each having the same normal samples

but different cancerous samples to have the same number of positive and negative features. We then ran four classification models on our eight datasets for every data set. We calculated feature importance using the Random Forest classifier and then observed which features were repeatedly coming in every sub-data. Five-fold cross-validation is done to make sure the model is performing well.

Results for BRCA on eRNAs

Enhancer RNA(Ensemble ID)	Frequency
ENSR00000043065	7
ENSR00000147309	7
ENSR00000156405	6
ENSR00000342021	5
ENSR00000115682	5
ENSR00000031042	5
ENSR00000148472	5
ENSR00000205366	4
ENSR00000293601	4
ENSR00000238462	4
ENSR00000326489	4
ENSR00000012874	4
ENSR00000201430	4
ENSR00000303726	4
ENSR00000043150	4
ENSR00000198506	4
ENSR00000038731	3
ENSR00000023843	3
ENSR00000323747	3
ENSR00000205365	3
ENSR00000293600	3
ENSR00000154577	3
ENSR00000123178	3
ENSR00000298505	3
ENSR00000126014	3
ENSR00000151513	3
ENSR00000318620	3
ENSR00000258012	3
ENSR00000258113	3
ENSR00000264769	3
ENSR00000025108	3
ENSR00000123179	3
ENSR00000022371	3
ENSR00000285930	3
ENSR00000202880	3
ENSR00000243983	3
ENSR00000122295	2
ENSR00000251619	2
ENSR00000251618	2
ENSR00000235397	2

Table 2.1: Best BRCA eRNA features.

Logistic Regression

The row of data contains the eRNA gene IDS, and the columns have the sample ID and the feature labels for the machine learning. We have used the sklearn package in Python to implement the Logistic Regression. We have used the libilinear kernel[47] as it works well for large datasets. The data is split into 80-20 ratios,80% for training and 20% for testing using the sklearn. We have obtained an overall accuracy after five fold cross-validation is 94% and an F1 score of 95% using the testing dataset shown in Table:2.2. Also AUC plot is shown in Fig2.3 with AUC score of 99%.

Model	Logistic Regression
Best Hyper parameter	Kernel= libilinear,max_iter=1000
Best accuracy	0.94
F1- Score	0.94
5-Fold cross Val	0.94 accuracy with 0.01 SD

Table 2.2: Logistic Regression results for BRCA eRNA

Support vector machine

We have used the sklearn package for implementing the SVM model.We have used the "RBF"kernel[48] for SVM model.The training and testing are done on the eRIc Dataset[45] for BRCA cancer. The data is split into 80-20 ratios,80% for training and 20% for testing using the sklearn. We have obtained an overall accuracy after 5 fold cross validation is 75% and an F1 score of 96% using the testing dataset shown in Table:2.3. Also, the AUC plot is shown in Fig2.3 with an AUC score of 95%.

Model	SVC
Best Hyper parameter	kernel= rbf
Best accuracy	0.76
F1- Score	0.96
5-Fold cross Val	0.75 accuracy and 0.02 SD

Table 2.3: Support vector results

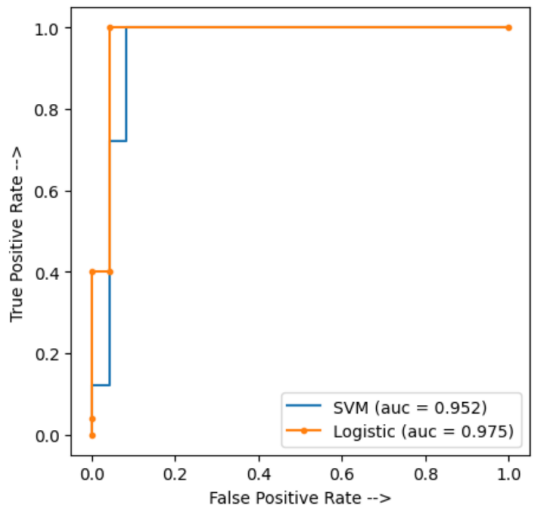

Figure 2.5: AUC curve for Logistic Regression and Support vector machine for BRCA eRNA

Random Forest

We have used sklearn package Random forest classifier in python to implement random forest. The training and testing are done on the eRIc Dataset[45] for BRCA cancer. The data is split into 80-20 ratios,80% for training and 20% for testing using the sklearn. We have obtained an overall accuracy after 5-fold cross-validation is 96% and an F1 score of 98% using the testing dataset shown in Table:2.4. Also, feature importance was calculated for eight different sub-datasets using random forest. The best features are selected for every sub dataset, and the most occurring features are shown in Table:2.1 for the BRCA.

Model	Random Forest
Best Hyper parameter	Model=Randomforestclassifier,nestimators=6
Best accuracy	0.96
F1- Score	0.98
5-Fold cross Val	0.95 accuracy with 0.01 SD

Table 2.4: Random Forest results for BRCA eRNA

K-nearest neighbour

For KNN, we have used the package KNeighborsClassifier from the sklearn in Python. The training and testing are done on the eRIc Dataset[45] for BRCA cancer. The data is split into 80-20 ratios,80% for training and 20% for testing using the sklearn. We have obtained an overall accuracy after 5 fold cross-validation is 74% and an F1 score of 73% using the testing dataset shown in Table:2.5.

Model	KNN
Best Hyper parameter	Model=KneighborsClassifier, MaxNeighbors=10
Best accuracy	0.76
F1- Score	0.73
5-Fold cross Val	0.73 accuracy with 0.05 SD

Table 2.5: KNN results for BRCA eRNA

2.4 Differential gene analysis

While doing machine learning, having many features can add up to run time and space complexity. To prevent this, we need to make the feature reduction and keep only those significant features. In order to reduce bad features, we perform differential gene analysis on the TGCA pan-cancer RNA expression data. We had 22,829 features in total, which are present across more than 24 types of cancer. So there is a need to remove the high number of features using feature reduction techniques.

We used a technique to reduce the number of features through differential gene analysis. We have used the deseq2[49] which is a package in R which helps in the normalization, visualization, and differential analysis of genes. We reduce the number of features from 22,829 to 4,750 by using a filter that log2Foldchange greater than one and value less than 0.1. Then we added labels to those genes; cancerous one was marker one, and normal was marked o. Also shown below is the plot that demonstrates the defferentially expressed genes in the entire dataset.

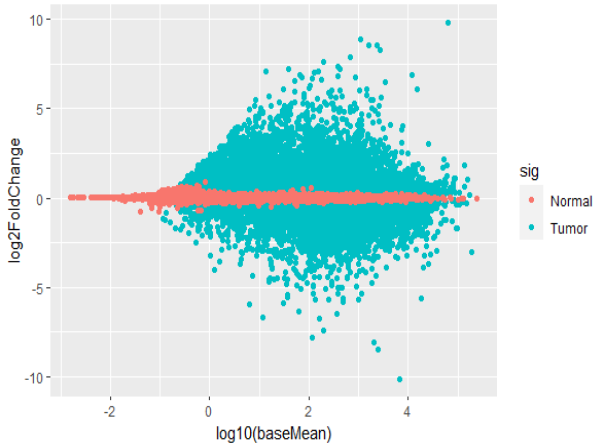

Figure 2.6: Differentially Expressing genes TCGA RNAs

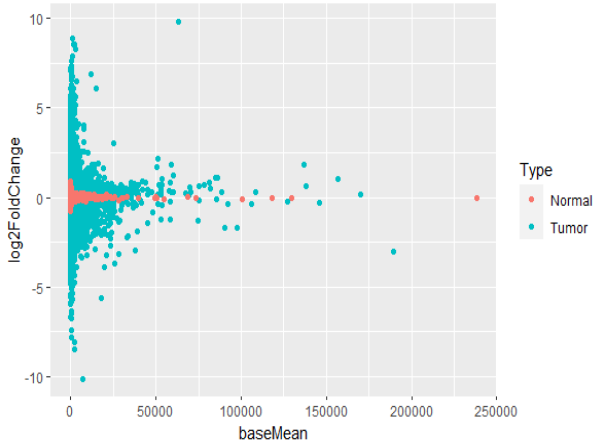

Figure 2.7: Differentially Expressing genes TCGA RNAs basemean vs log2fold

2.5 Stimulating liquid biopsy data

For simulating liquid biopsy data TCGA pan-cancer dataset and GTEx whole blood data is used to simulate data for liquid biopsy. In TCGA dataset, 24 cancer types that contain over 10,000 samples in total and 22,000 genes. Out of these samples, 9,000 samples were for different types of solid tumors and 1,000 normal solid tissue samples. Out of these 1,000 normal samples, only 5-10 samples contain expression from the blood samples. The liquid biopsy data for the normal samples is not available, and the available data is not accessible publicly.

Due to these constraints, we had to simulate the liquid biopsy data using our own approach, which is described below. We have used the GTEx normal whole blood data RNAs expression data to simulate the liquid biopsy cancer samples on different dilutions. In the figure:2.8, we have shown the

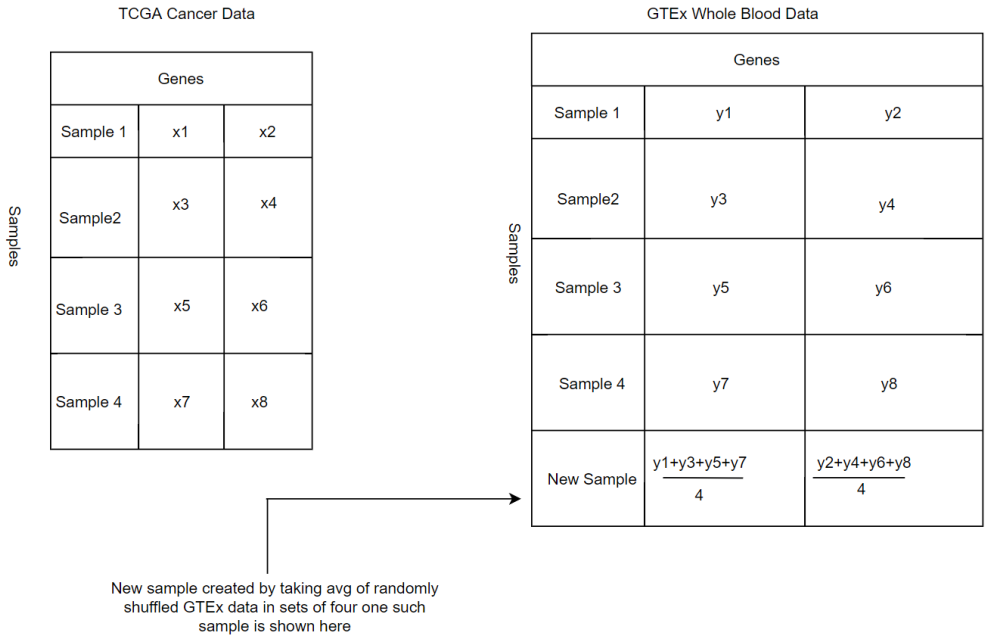

Figure 2.8: TCGA and GTEx RNAs data layout

data layout for the TCGA pan-cancer and GTEx whole blood RNAs. The columns contain the genes ids, whereas the rows have the samples. In order to stimulate the new samples by mixing blood, we need to process the whole blood data RNAs file. The first operation that was done on the GTEx whole blood data was that we randomly sampled the GTEX file and made pairs of four samples. Afterward, the new samples are created by taking the average expression value in four randomly sampled samples in the GTEx RNAs.Now we create a new file with those randomly averaged samples and mix it with the TCGA cancer samples in different ratios. So we have taken different dilutions. For example, if we want to simulate liquid biopsy data at 98% blood dilution, that means the new samples will have 98% blood data, and the rest 2% will be taken from TCGA cancer data. As demonstrated in Fig:2.9 for GTEx avg of randomly sampled four pairs of genes, the Z1-Z8 represents the average expression values in 4 GTEx whole blood data, and in the other data in which we have the TCGA samples, there expression values are shown with X1-X9, in order to simulate the new samples at 98% dilution we multiply

each row in the GTEx data with 0.98 and in the TCGA samples each row is multiplied by 0.02 then we add the corresponding samples in their gene order and create new diluted liquid biopsy samples which will be used for the machine learning process.

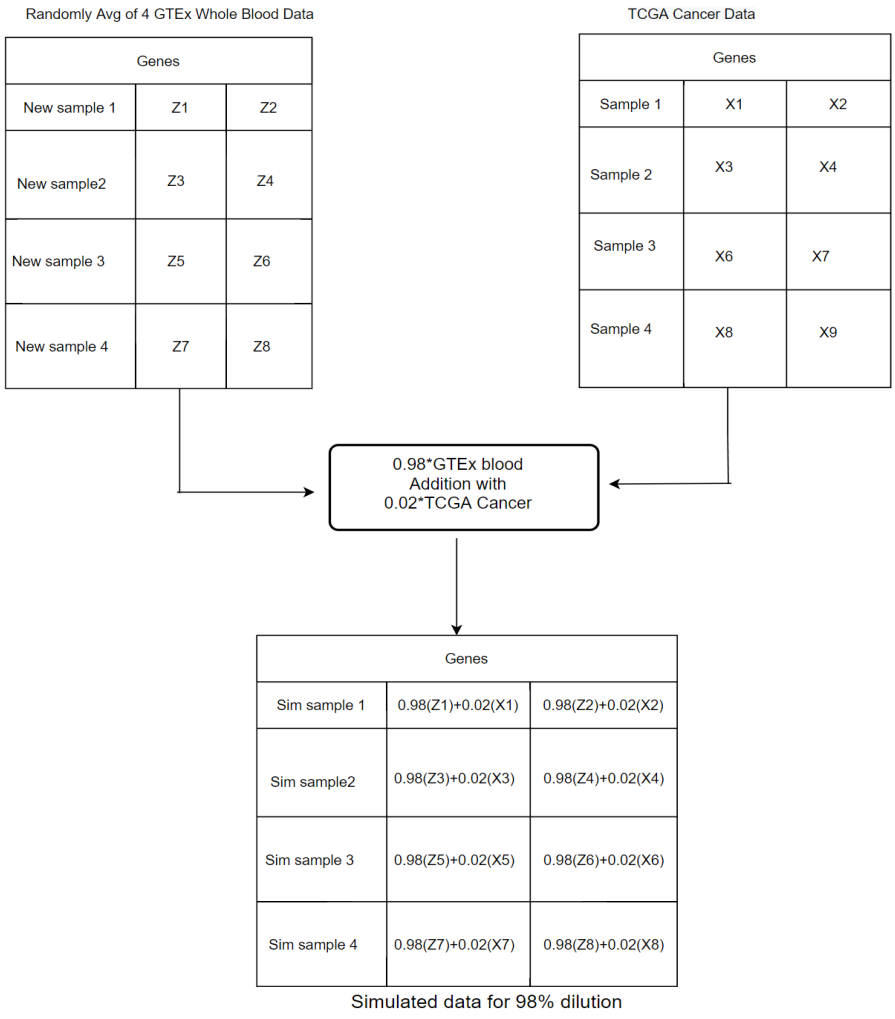

Figure 2.9: Data simulation for 98% Blood Dilution

2.6 Machine Learning

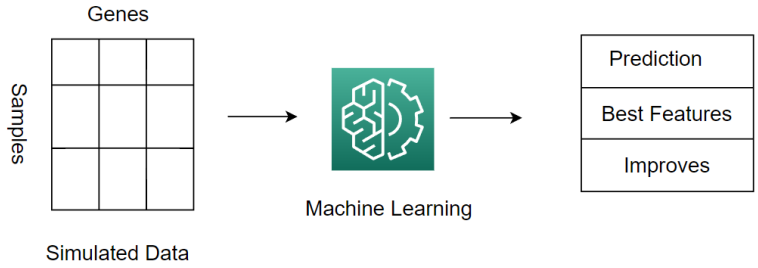

Figure 2.10: Machine learning overview

We have simulated liquid biopsy data using TCGA pan-cancer and GTEx whole blood RNAs described above. Now we have used this data for our machine learning models to find out the best features that are giving us the predictions. We have created three types of data models to further refine our features and ensure that some features are not because of the bias from the normal blood data we are getting. So the first model has the Stimulated cancer samples at 95% and 98% dilution vs. the Normal GTEx whole blood RNAs.This is our approach's primary model, and we have used the random forest technique to find the best features in this data. The second and the third model that we have is to ensure that we do not have any features which is there because of some bias in the dataset. So the second model has the Cancer TCGA samples(Without simulation) vs. the normal GTEx whole blood samples. The third model that we created is to determine if any feature is coming because of the bias in the GTEx whole blood dataset. This is the null model, which does not have any cancerous data. It has the randomly sampled in sets of four GTEx samples after taking their mean vs. the normal GTEx whole blood samples.

Here, now that we have all the data present and processed for the machine learning, we run the models on six types of cancer samples: breast cancer, prostate cancer, lung cancer, ovarian cancer, liver cancer, and colorectal cancer. We have used four machine learning models: logistic regression, support vector machine, random forest, and k-nearest neighbor. We tested the three models of the dataset for six different types of cancers.

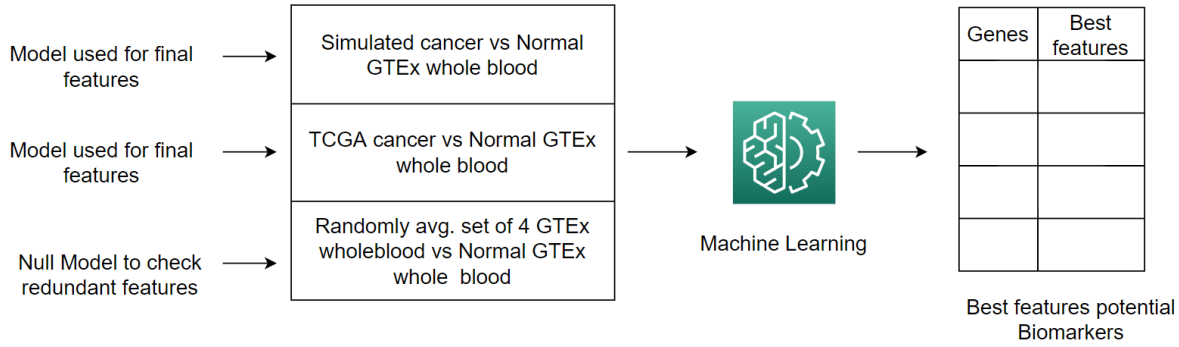

Figure 2.11: Machine learning workflow

2.7 Results for simulated data

2.7.1 Breast cancer

For the simulated cancer at 98% dilution vs the GTEx normal whole blood data we have taken 172 simulated samples and 172 normal whole blood samples for the machine learning models.Feature importance's are calculated using the random forest techniques the results are shown below which includes different machine learning model performance and the best features that we have obtained.

Table 2.6: Machine Learning scores for Breast cancer

Model	Logistic Regression	SVM	Random Forest	KNN
Best Hyper parameter	Kernel= lbfgs	kernel= rbf	n_estimators=6	MaxNeighbors=10
Best accuracy	0.99	0.98	0.98	0.81
F1- Score	0.99	0.98	0.99	0.79
5-Fold cross Val	0.99 accu and SD= 0.01	0.98 acc and SD=0.02	0.98 accu and SD= 0.01	0.80 accu and SD=0.06

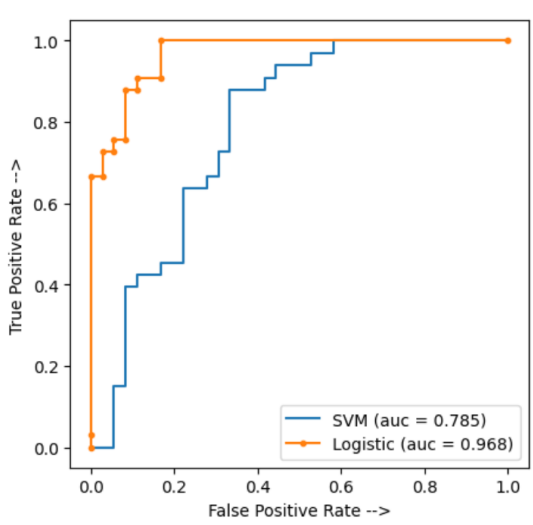

Figure 2.12: AUC for simulated breast cancer data comparison between SVM and logistic regression

Table 2.7: Best breast cancer markers

Best Breast cancer markers
POSTN
AZGP1
COL3A1
REP15
PODXL
SCGB2A1
COL1A2
CYP2F1
CST1
KRT9
MTTP
PLCL1
DMRTA2
LAMC2
SERPINE2
BAALC
GABRG1

2.7.2 Colorectal cancer

For the simulated cancer at 98% dilution vs the GTEx normal whole blood data we have taken 178 simulated samples and 178 normal whole blood samples for the machine learning models.Feature importance's are calculated using the random forest techniques the results are shown below which includes different machine learning model performance and the best features that we have obtained.

Table 2.8: Machine Learning scores for Colorectal cancer

Model	Logistic Regression	SVM	Random Forest	KNN
Best Hyper parameter	Kernel= lbfgs	kernel= rbf	n_estimators=6	MaxNeighbors=10
Best accuracy	0.98	0.97	0.99	0.71
F1- Score	0.99	0.97	0.98	0.69
5-Fold cross Val	0.97 accu and SD= 0.02	0.96 accu and SD= 0.03	0.99 accu and SD= 0.01	0.65 accu and SD= 0.02

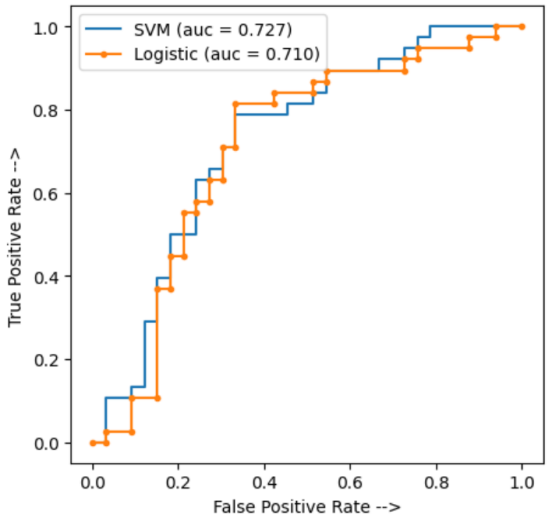

Figure 2.13: AUC for simulated colorectal cancer data comparison between SVM and logistic regression

Table 2.9: Best marker genes for colorectal cancer

Best Colorectal Cancer Markers
REP15
SCARNA10
RETNLB
SI
ELK2AP
FAM216B
RPPH1
LAMC2
CST1
SCGB2A1
CYP27C1
SYT6
CFI
CDH16
MIR4746
MYO1A
SCARNA16
CDCA7
ACER1
HSPA2
MIR1205
FAM181B
WDR72
PMEL

2.7.3 Liver Cancer

For the simulated cancer at 98% dilution vs the GTEx normal whole blood data we have taken 359 simulated samples and 359 normal whole blood samples for the machine learning models.Feature importance's are calculated using the random forest techniques the results are shown below which includes different machine learning model performance and the best features that we have obtained.

Table 2.10: Machine Learning scores for Liver cancer on simulated data

Model	Logistic Regression	SVM	Random Forest	KNN
Best Hyper parameter	Kernel= lbfgs	kernel= rbf	n_estimators=6	MaxNeighbors=10
Best accuracy	0.99	0.98	1	0.79
F1- Score	0.99	0.98	0.99	0.76
5-Fold cross Val	0.98 accu and SD= 0.01	0.98 acc and SD=0.01	1 accu and SD= 0.01	0.79 accu and SD=0.06

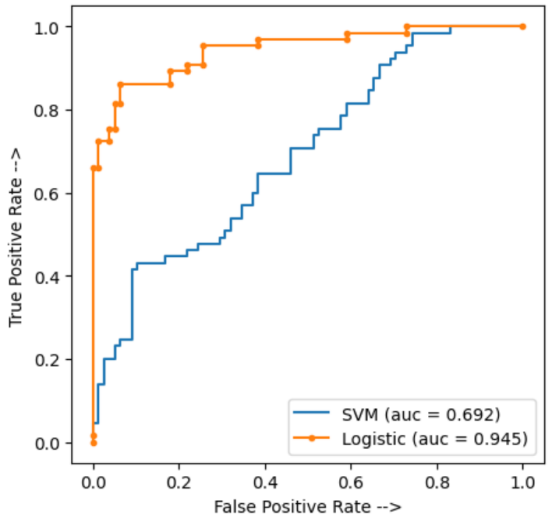

Figure 2.14: AUC for simulated liver data comparison between SVM and logistic regression

Table 2.11: Best marker genes for liver cancer

Best Liver cancer markers
ACOT6
CFHR4
AZGP1
SLC17A1
RPPH1
HMGCS2
NOXO1
SLC38A3
DCAF8L1
AKR7A2P1
CFH
UBD
APOB
APOC3
MTBP
TFCP2L1
CYP27C1
KCNJ8
CDH20
CNGB1
MGC27382
HIST3H2A
ADRB1

2.7.4 Lung Cancer

For the simulated cancer at 98% dilution vs the GTEx normal whole blood data we have taken 1,000 simulated samples and 1,000 normal whole blood samples for the machine learning models.Feature importance's are calculated using the random forest techniques the results are shown below which includes different machine learning model performance and the best features that we have obtained.

Table 2.12: Machine learning results for simulated lung cancer

Model	Logistic Regression	SVM	Random Forest	KNN
Best Hyper parameter	Kernel= lbfgs	kernel= rbf	n_estimators=6	MaxNeighbors=10
Best accuracy	0.97	0.98	1	0.81
F1- Score	0.98	0.97	0.99	0.79
5-Fold cross Val	0.94 accu and SD= 0.01	0.98 acc and SD=0.01	0.99 accu and SD= 0.01	0.80 accu and SD=0.02

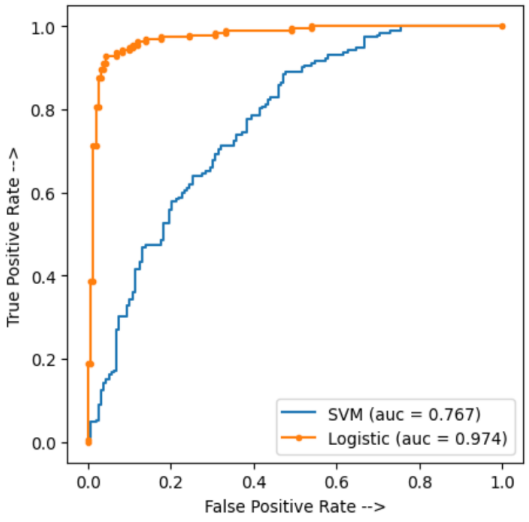

Figure 2.15: AUC for simulated lung data comparison between SVM and logistic regression

Table 2.13: Best marker genes for lung cancer simulated data

Best Lung cancer markers
LAMC2
ELK2AP
REP15
COL3A1
FAM216B
DMP1
DCAF8L1
UBD
NOXO1
AKR7A2P1
RPPH1
SCGB1A1
MS4A10
FAM71A
SLC13A5
CST2
CDSN
HIGD1B
SPINK6
AWAT2
PDYN
FAM19A3
SLC7A9
PDZK1P1
APOB
KIF4B
GRM8
ATP6V0A4
FOXB1
TRIM63
GPR17
KRT16P3
PECR
CELA3A
GSX2
A2ML1

2.7.5 Ovarian Cancer

For the simulated cancer at 98% dilution vs the GTEx normal whole blood data we have taken 391 simulated samples and 391 normal whole blood samples for the machine learning models.Feature importance's are calculated using the random forest techniques the results are shown below which includes different machine learning model performance and the best features that we have obtained.

Table 2.14: Machine learning score for simulated Ovarian cancer

Model	Logistic Regression	SVM	Random Forest	KNN
Best Hyper parameter	Kernel= lbfgs	kernel= rbf	n_estimators=6	MaxNeighbors=10
Best accuracy	0.94	0.98	1	0.8
F1- Score	0.95	0.97	0.99	0.79
5-Fold cross Val	0.92 accu and SD= 0.02	0.98 acc and SD=0.01	1 accu and SD= 0.01	0.78 accu and SD=0.03

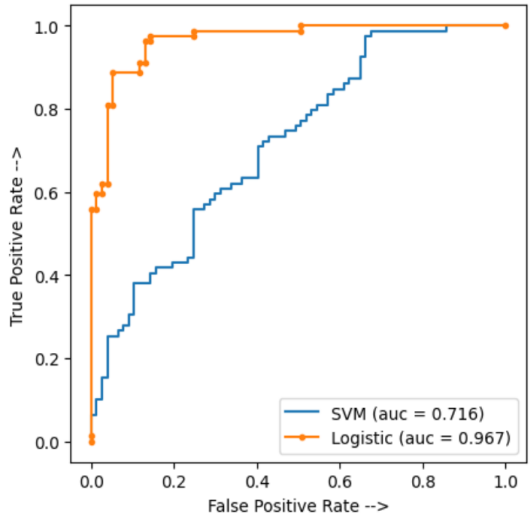

Figure 2.16: AUC for simulated Ovarian data comparison between SVM and logistic regression

Table 2.15: Best marker genes for ovarian cancer

Best marker Ovarian cancer
SOX17
SOX3
STRA8
SCGB2A1
NTS
LYPD1
ACPT
DMRT1
NRG3
MC4R
CST1
PYCR1
RPPH1
FAM181A
SLC22A24
CDSN
BMP3
COL22A1
QPCT
C19orf45
BBC3
RNF113B

2.7.6 Prostate Cancer

For the simulated cancer at 98% dilution vs the GTEx normal whole blood data we have taken 420 simulated samples and 420 normal whole blood samples for the machine learning models.Feature importance's are calculated using the random forest techniques the results are shown below which includes different machine learning model performance and the best features that we have obtained.

Table 2.16: Machine Learning score for the simulated prostate cancer

Model	Logistic Regression	SVM	Random Forest	KNN
Best Hyper parameter	Kernel= lbfgs	kernel= rbf	n_estimators=6	MaxNeighbors=10
Best accuracy	0.99	0.98	1	0.75
F1- Score	0.99	0.97	0.99	0.74
5-Fold cross Val	0.98 accu and SD= 0.02	0.98 acc and SD=0.01	1 accu and SD= 0.01	0.72 accu and SD=0.02

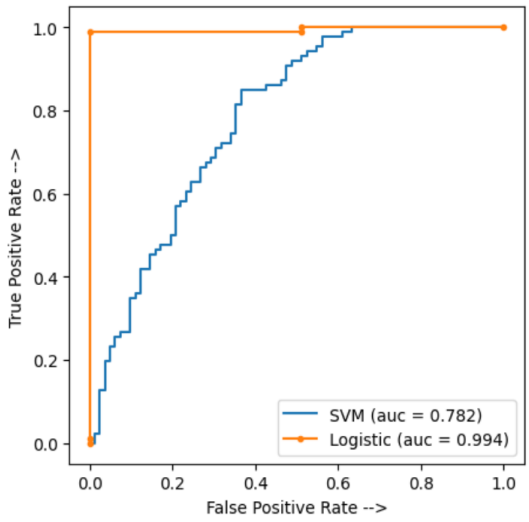

Figure 2.17: AUC for simulated Prostate cancer data comparison between SVM and logistic regression

Table 2.17: Best marker genes for simulated prostate cancer

Best markers for Prostate cancer
SLC9A2
BMPR1B
REP15
RAG2
THRB
MSI1
RPPH1
KLK2
LAGE3
CTGF
FOXF2
AKR7A2P1
PCA3
TMEM196
OSR2
SCN7A
MMP10
NPHS1
ARGFX
LIG1
OR1F2P
MIR4466
C1QTNF2
ELK2AP
GPD1
SCGB2A1
C1orf64
RUNDC3B

Chapter 3

Validation and inferences

3.1 Validation

3.1.1 Enhancer RNA inferences from results

The dataset which we have used for the analysis is the enhancer RNA in caners[44]. This data set has annotated the enhancers based on ENCODE taken from Ensembl 87. Then they mapped the eRNAs regions with the RNA-seq data and calculated the reads per million expression levels in different cancer samples. Also, the detectable eRNA level is those samples that have RPM greater than 1.

As we have shown in the results, we have calculated the best repetitive features from the 8 data sets, which we have created by splitting the eRiC eRNA dataset[44] for BRCA cancer. There are minimal studies on the enhancer RNA and their relation to cancer. Below we have shown the best markers that were found in BRCA with their mean expression shown, which is taken from the eRiC database[40].In our study, we find that 23 enhancers are up-regulated, and 4 are down-regulated, which can be used for targeting cancer or as a prognostic marker. ENSR00000258113 is the most differentially expressed gene with fold change 93 that is found in our study, which could act as a potential biomarker. Another interesting finding is ENSR00000023843(NET1e), found in three sub-data showed the most significant expression alteration between tumor and normal reported by [40] which is also present in our findings. This further suggests that the NET1e can be a promising target for BRCA eRNA cancer therapy. Also, ENSR00000122295, which is found in 2 sub-data and shown in the table below, is found to

have high expression in the BRCA-basal subtype, which putatively targets the BRCA-basal marker gene EN1[40]. - We observe that our markers are

Enhancer RNA(Ensemble ID)	Frequency	Fold change
ENSR00000258113	3	93
ENSR00000298505	3	24.83
ENSR00000285930	3	16
ENSR00000147309	7	10.8
ENSR00000342021	5	9.75
ENSR00000258012	3	8.1
ENSR00000122295	2	7.02
ENSR00000043065	7	7
ENSR00000023843	3	5.8
ENSR00000251619	2	4.54
ENSR00000251618	2	4.44
ENSR00000323747	3	4.06
ENSR00000205365	3	4.03
ENSR00000205366	4	3.8
ENSR00000148472	5	3.5
ENSR00000123178	3	3.03
ENSR00000293601	4	3.02
ENSR00000293600	3	3.02
ENSR00000123179	3	2.96
ENSR00000243983	3	2.7
ENSR00000202880	3	2.37
ENSR00000154577	3	1.86
ENSR00000012874	4	1.8
ENSR00000022371	3	1.29
ENSR00000025108	3	1.26
ENSR00000326489	4	1.1
ENSR00000264769	3	1.07
ENSR00000198506	4	1
ENSR00000235397	2	-1.0231
ENSR00000043150	4	-1.1
ENSR00000126014	3	-1.12
ENSR00000303726	4	-1.16
ENSR00000038731	3	-1.2
ENSR00000201430	4	-1.3
ENSR00000318620	3	-1.34
ENSR00000238462	4	-1.4
ENSR00000115682	5	-1.43
ENSR00000151513	3	-1.62
ENSR00000031042	5	-1.63
ENSR00000156405	6	-3.73

Table 3.1: BRCA eRNA best features comparison with the expression score in eRiC database

relevant and needs to be further tested for more validation but this results seems promising and can help for selecting the markers while designing the

experiment.

3.1.2 Simulated Breast cancer inferences from results

We have calculated the top marker genes using the simulated samples from TCGA and GTEx data. Now to see the differential expressed gene, we used the BBcancer[50] database, which is the only database that had blood data analysis for some of the cancers. We downloaded the ctDNA differential gene analysis data from the BBcancer. We compared it with the genes that we were getting as top markers. 12 genes had a very good-adjusted p-value greater than 0.05, which suggests that these are indeed good differential expressing genes. For some genes like DMRTA2, data was not given in the BBcancer.Also, fold change is higher than 1.5 in almost every sample, which suggests that these genes are up-regulated in breast cancer.

We also searched the current status of these genes in relation to breast cancer to check our top markers. **SCGB2A1** is found in peripheral blood of patients(7%) with breast cancer whereas absent in normal healthy donor[51].**BAALC** is found significantly up-regulated in a study done on the triple-negative breast cancer tissues and cells[52].**AZGP1** is found to mainly express in the epithelial cells in the breast and gastrointestinal organs. In a study, they found that AZGP1, along with three other genes, can act as four marker panels for the prognostics in breast carcinoma[53].**SERPINE2** was found to promote breast cancer metastasis and is also found to be highly overexpressed in breast cancer[54]. textbfCOL1A2 is detected as up-regulated gene in patients with breast cancer[55].**POSTN** is shown to express in statistically significant in different breast cancer cell lines at mRNA level[56].**PLC1** mRNA relates to PIK3CA in 2 breast cancer chohorts[57].**Col3A1** is found to be associated with the P4HA2 in breast cancer progression[58].**PODXL** is found to enhance the tumor growth and metastasis in breast cancer[59].**MTTP** is found to be deferentially expressed in human breast cancer metastasis[60].**CYP2E1** is found to play an important role in breast cancer in advanced stages[61].**CST1** is found to be negatively correlated with breast cancer patients survival and it also promotes metastasis in breast cancer.CST1 is also reported as a novel prognostic marker for breast cancer[62].

Table 3.2: Differential Gene comparison of Simulated top breast cancer with the BBcancer data breast cancer ctDNA

Best Markers	FoldChange	baseMean	p-value	adj-pvalue
SCGB2A1	21.1179	14.57913	0.000913	0.002986
BAALC	19.83894	5.592534	0.006057	1
AZGP1	10.50533	342.3223	0.003714	0.010932
SERPINE2	8.564719	44.47146	0.05812	0.112137
COL1A2	4.900622	2.095357	0.484637	1
LAMC2	4.513945	1.16157	0.533955	1
POSTN	3.620567	1.962247	0.630473	1
KRT9	3.138762	0.089556	0.776993	1
PLCL1	2.80692	0.289152	0.800043	1
REP15	2.568354	0.022271	0.816723	1
COL3A1	2.566398	0.392626	0.78911	1
PODXL	2.556754	0.021642	0.817536	1
GABRG1	2.419006	0.46944	0.827208	1
MTTP	2.419006	0.235539	0.827208	1
CYP2F1	2.408108	0.017982	0.827974	1
CST1	2.379397	0.004173	0.829994	1
DMRTA2	NA	NA	NA	NA

3.1.3 Simulated Colorectal cancer inferences from results

We have calculated the top marker genes by using the simulated samples from TCGA and GTEx data. Now we for see the differential expressed gene we used the BBcancer database, which is the only database that had blood data analysis for some of the cancers. We downloaded the blood mRNA differential gene analysis data from the BBcancer[50] and compared it with the genes that we were getting as top markers. 14 genes were differentially expressed by comparing with the BBcancer blood mRNA colorectal cancer genes. The data was not available for some of the genes marked with NA in the table below.

We also did the literature search for our top markers to know the relevance of the results.**REP15** is detected as differentially expressed gene in colorectal cancer cells and is able to effectively distinguish patients prognosis[63].**LAMC2**

is overexpressing in colorectal cancer and promotes cell growth[64].**CST1** is tumor biomarker which give use full insights in diagnosis of colorectal cancer[65].**CDCA7** is highly expressed in colorectal cancer is useful for predicting tumor stages[66].**HSPA2** is detected as a novel target for the colorectal cancer which is causing cancer proliferation[67].**SCARNA10** is found to be differentially expressing in colorectal cancer and can be used to compare pathological tumor node metastasis stages[68].**RPPH1** is promoting colorectal cancer cells metastasis in vivo by binding with the TUBB3[69].**SCGB2A1** is found as a novel prognostic bio marker for the colorectal cancer which is related with chemo resistance[70].**MIR-4746** is found to be inhibiting the colorectal cancer cells proliferation by inhibition of CCND1[71].**MYO1A** is identified as a tumor suppressor gene for the colorectal cancer[72] .**SCARNA16** is know for proliferation of colon cancer[73].**MIR1205** works with Circ0082182 to activate the wnt pathway and suppress colorectal cancer[74].

Table 3.3: Differential Gene comparison of Simulated top Colorectal cancer with the BBcancer data for the blood RNAs

Best Markers	FoldChange	baseMean	p-value	adj-pvalue
WDR72	-0.04276	2.064186	0.906937	0.990729
RETNLB	-0.03566	3.861078	0.885221	0.989071
REP15	-0.0725	2.629236	0.765894	0.974622
CFI	-0.10849	2.285237	0.662734	0.961186
CDH16	-0.15786	3.432986	0.635085	0.958602
LAMC2	-0.16982	2.443644	0.520307	0.943752
SYT6	0.335421	3.779011	0.51462	0.943752
ACER1	0.194122	3.631744	0.507675	0.943604
FAM216B	0.261808	3.034209	0.474733	0.936718
CST1	0.214287	3.149597	0.428822	0.931202
FAM181B	0.266648	7.413065	0.417119	0.927108
CDCA7	-0.39921	3.984698	0.401072	0.923338
HSPA2	1.381254	4.604732	0.030717	0.883148
PMEL	-1.21949	4.98533	0.060694	0.883148
SCARNA10	NA	NA	NA	NA
SI	NA	NA	NA	NA
ELK2AP	NA	NA	NA	NA
RPPH1	NA	NA	NA	NA
SCGB2A1	NA	NA	NA	NA
CYP27C1	NA	NA	NA	NA
MIR4746	NA	NA	NA	NA
MYO1A	NA	NA	NA	NA
SCARNA16	NA	NA	NA	NA
MIR1205	NA	NA	NA	NA

3.1.4 Simulated Liver cancer inferences from results

We have calculated the top marker genes by using the simulated samples from TCGA and GTEx data. To see the differential expressed gene, we used the BBcancer database, which is the only database with blood data analysis for some of the cancers. We downloaded the blood mRNA differential gene analysis data from the BBcancer[50] and compared it with the genes that we were getting as top markers. 24 genes were differentially expressed by

comparing with the BBcancer tissue mRNA liver cancer differential genes. The data was not available for some of the genes, which are marked with NA in the table below.

We also did the literature search for our top markers to know the relevance of the results.**AZGP1** is downregulated in the hepatocellular cancer and can act as a prognostic marker for HCC[75].**HMGCS2** regulates proliferation of hepatocellular caricinoma[76].**NOX1** has role in controlling the growth of the liver tumor cells by up-regulating the EFGR pathways[77].**CFH** help in regulating the stemness of the liver cancer cell by LSf-1[78].**UbD** is observed to be up-regulated in liver and can act as a potential preneoplastic marker[79].**ABOP** can be an important prognostic marker for HCC patients who underwent liver surgery[80].**APOC3** is likely to can be a robust diagnostic marker for hepatitis B virus-related HCC[81].**MTBp** promotes the HCC metastasis by promoting the degradation of E cadherin[82].**ADRA1A** promoter methylation is has association with hepatocellular cancer, which is shown in a study conducted on 160 patient's[83].

Table 3.4: Differential gene comparison of simulated top liver cancer genes with the BBcancer Liver cancer tumor

Best Markers	FoldChange	baseMean	p-value	adj-pvalue
MTBP	2.303848847	166.6053583	0.203764	11.30648
HIST3H2A	3.864954351	53.18145149	0.503969	7.669032
CYP27C1	2.386536639	43.62489926	0.45087	5.293181
DCAF8L1	5.987174443	16.77971733	1.963572	3.049124
UBD	2.161450958	1.877897271	0.876216	2.466803
APOB	0.972314978	6.644180044	0.410201	2.370339
CDH20	1.208525284	2.522237898	0.533347	2.265926
CNGB1	0.402941279	1.906703674	0.608052	0.662676
TFCP2L1	0.223149581	45.01047989	0.368855	0.604979
ACOT6	-0.139450645	9.18440206	0.387531	-0.35984
NOXO1	-1.438605187	0.444361975	0.962101	-1.49528
ADRB1	-0.897817593	37.41202322	0.541231	-1.65884
KCNJ8	-0.767417941	1786.678232	0.399663	-1.92016
SLC38A3	-0.472608602	13496.05362	0.245996	-1.92121
CFH	-0.749889869	67135.35938	0.245858	-3.05009
CFHR4	-1.483163839	2044.455967	0.387833	-3.82423
HMGCS2	-1.543397787	69596.36748	0.305432	-5.05316
SLC17A1	-1.456027583	1084.238898	0.284156	-5.12405
AZGP1	-1.750365254	36039.28155	0.28804	-6.07681
APOC3	-2.007780548	89406.88637	0.328684	-6.10855
RPPH1	NA	NA	NA	NA
AKR7A2P1	NA	NA	NA	NA
MGC27382	NA	NA	NA	NA

3.1.5 Simulated Lung cancer inferences from result

We have calculated the top marker genes by using the simulated samples from TCGA and GTEx data. Now, let us see the differential expressed gene; we used the BBcancer database, which is the only database that had blood data analysis for some of the cancers, including lung cancer. We downloaded the blood mRNA differential gene analysis data for the lung cancer from the BBcancer[50] and compared it with the genes that we were getting as top markers. A total of 30 genes were found to be differentially expressed with an

adjusted p-value of greater than 0.05. 10 genes were found to be upregulated with a fold change of greater than 0.5.

Now let's look at the top marker and their current research status.**SCGB1A1** along with other seven genes are found to be related to the TRIM58 methylation and may be used as prognostic biomarkers in the lung squamous cell carcinoma[84].**LAMC2** is found to promote the metastasis potential of lung adenocarcinoma [85].**GRM8** is shown to act as a therapeutic target for the squamous cell lung cancer determined using genomic sequencing[86].**DCAF8L1** is found to be downregulated in the non-small cell lung cancer[87].**ApoB** high levels are correlated with increase lung cancer incidents[88].**HIGD1B** with other 4 genes are significant differentially expressing genes confirmed by qPCR done on LUAD tissues and cell lines[89].**DMP1** is found to be highly expressed in the human lung cancer[90].**RPPH1** is known to promote the nonsmall cell lung cancer development by miR-326 and wnt2B pathways[91].

Table 3.5: Differential gene comparison of simulated top Lung cancer genes with the BBcancer blood lung cancer

Best Markers	FoldChange	baseMean	p-value	adj-pvalue
SCGB1A1	1.652102728	8.312447867	0.025957	0.999857
CST2	1.599025411	8.513732612	0.027216	0.999857
MS4A10	1.507620204	6.779999493	0.226871	0.999857
TRIM63	1.33789871	4.868952084	0.082131	0.999857
LAMC2	1.11061504	7.683411704	0.231636	0.999857
GRM8	0.972792476	6.633838104	0.445775	0.999857
COL3A1	0.945393356	5.19341434	0.491498	0.999857
SLC13A5	0.930597664	7.339927116	0.272569	0.999857
CELA3A	0.880187595	5.553025191	0.511322	0.999857
FAM216B	0.63662126	8.567189477	0.489402	0.999857
SLC7A9	0.627918829	6.172108188	0.591753	0.999857
PDYN	0.553074376	7.78508295	0.473527	0.999857
FAM19A3	0.478290198	6.836092345	0.41222	0.999857
DCAF8L1	0.266646596	3.994098883	0.759212	0.999857
NOXO1	0.178738575	4.845255885	0.814453	0.999857
KIF4B	0.009929211	6.094944788	0.993844	0.999857
GSX2	-0.015377292	4.0259193	0.984578	0.999857
AWAT2	-0.055321893	4.337028509	0.961398	0.999857
APOB	-0.07076288	8.438110613	0.881987	0.999857
HIGD1B	-0.200379249	3.992140969	0.702843	0.999857
CDSN	-0.285395474	6.904809094	0.591724	0.999857
FOXB1	-0.326095477	4.215115982	0.563142	0.999857
UBD	-0.36779515	7.988895343	0.550719	0.999857
ATP6V0A4	-0.419571812	4.512713489	0.405605	0.999857
REP15	-0.664698972	7.168057949	0.366167	0.999857
DMP1	-0.93852312	4.368505255	0.227302	0.999857
GPR17	-1.133464439	5.167944349	0.254641	0.999857
SPINK6	-1.164437764	5.673590954	0.278308	0.999857
PECR	-1.248254958	8.429043167	0.04848	0.999857
A2ML1	-2.495479842	5.863675466	0.028976	0.999857
FAM71A	-3.637358939	5.251812396	0.011944	0.999857
ELK2AP	NA	NA	NA	NA
AKR7A2P1	NA	NA	NA	NA
RPPH1	NA	NA	NA	NA
PDZK1P1	NA	NA	NA	NA
KRT16P3	NA	NA	NA	NA

3.1.6 Simulated Prostate cancer inferences from results

We have calculated the top marker genes by using the simulated samples from TCGA and GTEx data. Now we see the differential expressed gene we used the BBcancer database, which is the only database that had blood data analysis for some of the cancers.We downloaded the prostate differential mRNA gene analysis data from the BBcancer[50] and compared it with the genes that we were getting as top markers for the simulated data. A total of 22 genes were found to be differentially expressed with an adjusted p-value of greater than 0.05. 11 genes were found to be upregulated, and 11 genes were downregulated.

Now let's look at the top marker and their current research status.**BMPR1B** is classified well-differentiated genes, and their expression correlates with tumor grade in prostate cancer[92].**KLK2** is known to proliferate the prostate cancer growth by ARA70 receptor binding transactivation[93].**THRB** and along with 4 other genes are classified as a discriminatory marker set to distinguish between prostate cancer and BPH in the blood plasma[94].**MMP10** is found to correlate with the different stages and progression of human prostate cancer.MMP10 is also upregulated in non metastatic prostate cancers[95].**CTGF** promotes the prostate cancer metastatic process in the bone by deregulating the osteoblast cell differentiation process[96].**FOXF2** is found to be a differential expressing gene and play role in FOXF2 in prostate homeostasis[97].**PCA3** is one of the most relevant marker which is used for early stages diagnosis of prostate cancer[98].

Table 3.6: Differential gene comparison of simulated top prostate cancer genes with the BBcancer data prostate cancer

Best Markers	FoldChange	baseMean	p-value	adj-pvalue
BMPR1B	1.132044758	2098.762981	0.19917	5.683808
KLK2	1.061537866	77686.4785	0.214654	4.945354
SLC9A2	1.041366069	646.468596	0.223788	4.653351
C1orf64	1.952626528	83.10634917	0.61966	3.151124
THRB	0.650743955	1173.42365	0.238586	2.727501
MSI1	0.444063087	424.8931199	0.173679	2.55681
MMP10	0.92019683	66.12185143	0.511005	1.800759
LAGE3	0.28799462	55.26291796	0.327246	0.880056
TMEM196	0.637136489	57.29572459	0.783179	0.813526
NPHS1	0.444296639	136.371438	0.737645	0.602317
C1QTNF2	0.245741833	78.86520021	0.472012	0.520626
RUNDC3B	0.212326645	123.1127114	0.52276	0.406165
ARGFX	0.433905465	14.05554598	1.17021	0.370793
REP15	-0.221259261	7.551922929	1.657444	-0.13349
SCGB2A1	-0.273793391	19.05472106	0.875958	-0.31256
CTGF	-0.152237841	5095.765309	0.345981	-0.44002
RAG2	-0.717031368	28.99097385	0.900617	-0.79616
LIG1	-0.196186401	474.1724976	0.210447	-0.93224
GPD1	-0.512130752	982.565965	0.21833	-2.34567
FOXF2	-0.848526704	108.8049437	0.358546	-2.36658
SCN7A	-0.571334648	897.7458806	0.191294	-2.98668
OSR2	-0.604539117	672.5476628	0.200315	-3.01794
RPPH1	NA	NA	NA	NA
AKR7A2P1	NA	NA	NA	NA
PCA3	NA	NA	NA	NA
OR1F2P	NA	NA	NA	NA
MIR4466	NA	NA	NA	NA
ELK2AP	NA	NA	NA	NA

3.1.7 Simulated Ovarian cancer inferences from results

We have calculated the top marker genes by using the simulated samples from TCGA and GTEx data. There was no differential gene expression data on the BBcancer database for the RNAs in ovarian cancer, so we searched the literature for our top markers to find their relevance.

Now we will see current research status for our top genes in the literature shown below.**SOX17** is highly expressed in the ovarian cancer and it also promotes angiogenesis in ovarian cancer by interacting with the transcription factor PAX8[99].**SOX3** which is found as a top gene is shown to be act as a oncogene in the epithelial ovarian cancer which act by targeting Src Kinase[100].**NTS** in ovarian cancer is an essential mediator for the progression and metastasis of high grade serous carcinoma in ovarian cancer[101].**LYPD1** is highly expressed in primary as well as metastatic ovarian cancer and can be used for therapeutic targeting[102].**MC4R** is mutation leads to development of ovarian teratomas in mice[103].**COL22A1** is one of the most differentially expressing gene in ovarian cancer reported present on chromosome 8 and might have role in ovarian cancer growth[104]. Other genes which are shown in the table need further validation for more inferences.

Chapter 4

Conclusion

Cancer detection is conventionally done using invasive solid tumor biopsies that can cause side effects on the body and are not so pocket friendly. So there is a need to look for new methods to solve the cancer prognosis and detection at early stages. Also, at the same time, make the process non-invasive and affordable. This problem can be solved by analyzing the liquid biopsies from the patients and looking for those particular markers that have significantly changed their expression in healthy vs. cancer patients. Then use sets of those particular markers for specific cancer and make a marker panel that can diagnose cancer.

Although liquid biopsies are an excellent alternative to replace conventional diagnostic, but the expression data of the liquid biopsies for cancer and normal persons is not available in the public domain. As this is a comparatively new field, researchers want to do their analysis first and then make data available. We have tried to bridge this gap by simulating liquid biopsies data with existing data in the public domain. We have used the TCGA pan-cancer data and GTEX whole blood data and created simulated liquid biopsies data at different concentrations, and then used machine learning on them to try to find the best possible marker for six types of cancers.

The findings from the simulated data are very promising. When we did the literature search on the best markers that we have detected in simulated cancer, we saw a lot of markers that we have detected are reported as biomarkers for those cancers. Also, we compared our best marker genes in simulated cancer with the available liquid biopsies samples for three cancer types and have found that the genes which we have identified are differentially expressed in those liquid biopsies with an adjusted p-value greater than 0.05.

For those genes that are not found in the current literature search but are detected in simulated data, those genes need further validation by doing lab experiments using qPCR or some other techniques. We believe that our research can help design the ideal set of markers that can be used in the diagnosis and tracking of different cancers.

Bibliography

[1] R. Palmirotta, D. Lovero, P. Cafforio, C. Felici, F. Mannavola, E. Pellè, D. Quaresmini, M. Tucci, and F. Silvestris, "Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology," *Therapeutic advances in medical oncology*, vol. 10, p. 1758835918794630, 2018.

[2] K. E. Aaltonen, V. Novosadová, P.-O. Bendahl, C. Graffman, A.-M. Larsson, and L. Rydén, "Molecular characterization of circulating tumor cells from patients with metastatic breast cancer reflects evolutionary changes in gene expression under the pressure of systemic therapy," *Oncotarget*, vol. 8, no. 28, p. 45544, 2017.

[3] F. Xie, P. Li, J. Gong, H. Tan, and J. Ma, "Urinary cell-free dna as a prognostic marker for kras-positive advanced-stage nsclc," *Clinical and Translational Oncology*, vol. 20, no. 5, pp. 591–598, 2018.

[4] S. Mastoraki, A. Strati, E. Tzanikou, M. Chimonidou, E. Politaki, A. Voutsina, A. Psyrri, V. Georgoulias, and E. Lianidou, "Esr1 methylation: A liquid biopsy–based epigenetic assay for the follow-up of patients with metastatic breast cancer receiving endocrine treatment," *Clinical Cancer Research*, vol. 24, no. 6, pp. 1500–1510, 2018.

[5] W. Gao, T. Huang, H. Yuan, J. Yang, Q. Jin, C. Jia, G. Mao, and J. Zhao, "Highly sensitive detection and mutational analysis of lung cancer circulating tumor cells using integrated combined immunomagnetic beads with a droplet digital pcr chip," *Talanta*, vol. 185, pp. 229–236, 2018.

[6] J. von Felden, T. Garcia-Lezana, K. Schulze, B. Losic, and A. Villanueva, "Liquid biopsy in the clinical management of hepatocellular carcinoma," *Gut*, vol. 69, no. 11, pp. 2025–2034, 2020.

[7] D. Roy and M. Tiirikainen, "Diagnostic power of dna methylation classifiers for early detection of cancer," *Trends in cancer*, vol. 6, no. 2, pp. 78–81, 2020.

[8] B. Pardini, A. A. Sabo, G. Birolo, and G. A. Calin, "Noncoding rnas in extracellular fluids as cancer biomarkers: the new frontier of liquid biopsies," *Cancers*, vol. 11, no. 8, p. 1170, 2019.

[9] K. I. Lei, L. Y. Chan, W.-Y. Chan, P. J. Johnson, and Y. D. Lo, "Diagnostic and prognostic implications of circulating cell-free epstein-barr virus dna in natural killer/t-cell lymphoma," *Clinical Cancer Research*, vol. 8, no. 1, pp. 29–34, 2002.

[10] M. S. Kopreski, F. A. Benko, L. W. Kwak, and C. D. Gocke, "Detection of tumor messenger rna in the serum of patients with malignant melanoma," *Clinical cancer research*, vol. 5, no. 8, pp. 1961–1965, 1999.

[11] J. Krol, I. Loedige, and W. Filipowicz, "The widespread regulation of microrna biogenesis, function and decay," *Nature Reviews Genetics*, vol. 11, no. 9, pp. 597–610, 2010.

[12] P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, *et al.*, "Circulating micrornas as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences*, vol. 105, no. 30, pp. 10513–10518, 2008.

[13] J. M. Thomson, M. Newman, J. S. Parker, E. M. Morin-Kensicki, T. Wright, and S. M. Hammond, "Extensive post-transcriptional regulation of micrornas and its implications for cancer," *Genes & development*, vol. 20, no. 16, pp. 2202–2207, 2006.

[14] J. D. Arroyo, J. R. Chevillet, E. M. Kroh, I. K. Ruf, C. C. Pritchard, D. F. Gibson, P. S. Mitchell, C. F. Bennett, E. L. Pogosova-Agadjanyan, D. L. Stirewalt, *et al.*, "Argonaute2 complexes carry a population of circulating micrornas independent of vesicles in human plasma," *Proceedings of the National Academy of Sciences*, vol. 108, no. 12, pp. 5003–5008, 2011.

[15] A. Gallo, M. Tandon, I. Alevizos, and G. G. Illei, "The majority of microrna detectable in serum and saliva is concentrated in exosomes," *PloS one*, vol. 7, no. 3, p. e30679, 2012.

[16] B. Pardini and G. A. Calin, "Micrornas and long non-coding rnas and their hormone-like activities in cancer," *Cancers*, vol. 11, no. 3, p. 378, 2019.

[17] C. H. Lawrie, S. Gal, H. M. Dunlop, B. Pushkaran, A. P. Liggins, K. Pulford, A. H. Banham, F. Pezzella, J. Boultwood, J. S. Wainscoat, *et al.*, "Detection of elevated levels of tumour-associated micrornas in serum of patients with diffuse large b-cell lymphoma," *British journal of haematology*, vol. 141, no. 5, pp. 672–675, 2008.

[18] R. Hamam, D. Hamam, K. A. Alsaleh, M. Kassem, W. Zaher, M. Alfayez, A. Aldahmash, and N. M. Alajez, "Circulating micrornas in breast cancer: novel diagnostic and prognostic biomarkers," *Cell death & disease*, vol. 8, no. 9, pp. e3045–e3045, 2017.

[19] X. Liu, B. Pan, L. Sun, X. Chen, K. Zeng, X. Hu, T. Xu, M. Xu, and S. Wang, "Circulating exosomal mir-27a and mir-130a act as novel diagnostic and prognostic biomarkers of colorectal cancer," *Cancer Epidemiology and Prevention Biomarkers*, vol. 27, no. 7, pp. 746–754, 2018.

[20] T. Kiss, "Small nucleolar rnas: an abundant group of noncoding rnas with diverse cellular functions," *Cell*, vol. 109, no. 2, pp. 145–148, 2002.

[21] M. T. Bohnsack and K. E. Sloan, "Modifications in small nuclear rnas and their roles in spliceosome assembly and function," *Biological chemistry*, vol. 399, no. 11, pp. 1265–1276, 2018.

[22] A. Baraniskin, S. Nöpel-Dünnebacke, M. Ahrens, S. G. Jensen, H. Zöllner, A. Maghnouj, A. Wos, J. Mayerle, J. Munding, D. Kost, *et al.*, "Circulating u2 small nuclear rna fragments as a novel diagnostic biomarker for pancreatic and colorectal adenocarcinoma," *International journal of cancer*, vol. 132, no. 2, pp. E48–E57, 2013.

[23] J. Köhler, M. Schuler, T. C. Gauler, S. Nöpel-Dünnebacke, M. Ahrens, A.-C. Hoffmann, S. Kasper, F. Nensa, B. Gomez, M. Hahnemann,

et al., "Circulating u2 small nuclear rna fragments as a diagnostic and prognostic biomarker in lung cancer patients," *Journal of cancer research and clinical oncology*, vol. 142, no. 4, pp. 795–805, 2016.

[24] T. Kitagawa, K. Taniuchi, M. Tsuboi, M. Sakaguchi, T. Kohsaki, T. Okabayashi, and T. Saibara, "Circulating pancreatic cancer exosomal rna s for detection of pancreatic cancer," *Molecular oncology*, vol. 13, no. 2, pp. 212–227, 2019.

[25] J. Liao, L. Yu, Y. Mei, M. Guarnera, J. Shen, R. Li, Z. Liu, and F. Jiang, "Small nucleolar rna signatures as biomarkers for non-small-cell lung cancer," *Molecular cancer*, vol. 9, no. 1, pp. 1–10, 2010.

[26] C. Xie, J. Yuan, H. Li, M. Li, G. Zhao, D. Bu, W. Zhu, W. Wu, R. Chen, and Y. Zhao, "Noncodev4: exploring the world of long non-coding rna genes," *Nucleic acids research*, vol. 42, no. D1, pp. D98–D103, 2014.

[27] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding rnas: insights into functions," *Nature reviews genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[28] S. Ren, F. Wang, J. Shen, Y. Sun, W. Xu, J. Lu, M. Wei, C. Xu, C. Wu, Z. Zhang, *et al.*, "Long non-coding rna metastasis associated in lung adenocarcinoma transcript 1 derived minirna as a novel plasma-based biomarker for diagnosing prostate cancer," *European journal of cancer*, vol. 49, no. 13, pp. 2949–2959, 2013.

[29] Y.-S. Tong, X.-W. Wang, X.-L. Zhou, Z.-H. Liu, T.-X. Yang, W.-H. Shi, H.-W. Xie, J. Lv, Q.-Q. Wu, and X.-F. Cao, "Identification of the long non-coding rna pou3f3 in plasma as a novel biomarker for diagnosis of esophageal squamous cell carcinoma," *Molecular cancer*, vol. 14, no. 1, pp. 1–13, 2015.

[30] W. Li, D. Notani, and M. G. Rosenfeld, "Enhancers as non-coding rna transcription units: recent insights and future perspectives," *Nature Reviews Genetics*, vol. 17, no. 4, pp. 207–223, 2016.

[31] W. Li, Y. Hu, S. Oh, Q. Ma, D. Merkurjev, X. Song, X. Zhou, Z. Liu, B. Tanasa, X. He, *et al.*, "Condensin i and ii complexes license full

estrogen receptor α-dependent enhancer activation," *Molecular cell*, vol. 59, no. 2, pp. 188–202, 2015.

[32] C. A. Melo, J. Drost, P. J. Wijchers, H. van de Werken, E. de Wit, J. A. O. Vrielink, R. Elkon, S. A. Melo, N. Léveillé, R. Kalluri, *et al.*, "ernas are required for p53-dependent enhancer activity and gene transcription," *Molecular cell*, vol. 49, no. 3, pp. 524–535, 2013.

[33] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[34] J. Tolles and W. J. Meurer, "Logistic regression: relating patient characteristics to outcomes," *Jama*, vol. 316, no. 5, pp. 533–534, 2016.

[35] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on mathematical software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[36] A. Maharjan, *Machine Learning Approach for Predicting Cancer Using Gene Expression*. PhD thesis, University of Nevada, Las Vegas, 2020.

[37] D. Petkovic, R. Altman, M. Wong, and A. Vigil, "Improving the explainability of random forest classifier–user centered approach," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pp. 204–215, World Scientific, 2018.

[38] AndreaBild'sLab, "Patient tumor samples." https://www.ncbi.nlm.nih.gov/biosample/SAMN03164102, nov 2011.

[39] GTExV8, "Patient rna seq data." https://gtexportal.org/home/datasets#filesetFilesDiv131, june 2017.

[40] Z. Zhang, J.-H. Lee, H. Ruan, Y. Ye, J. Krakowiak, Q. Hu, Y. Xiang, J. Gong, B. Zhou, L. Wang, *et al.*, "Transcriptional landscape and clinical utility of enhancer rnas for erna-targeted therapy in cancer," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[41] Z. Zhang, W. Hong, H. Ruan, Y. Jing, S. Li, Y. Liu, J. Wang, W. Li, L. Diao, and L. Han, "Hera: an atlas of enhancer rnas across human tissues," *Nucleic acids research*, vol. 49, no. D1, pp. D932–D938, 2021.

[42] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[43] DavisVaughan, "Dplyr." https://github.com/tidyverse/dplyr, jan 2014.

[44] LengHan, "eric (enhancer rna in cancers)." https://hanlab.uth.edu/eRic/download, jan 2008.

[45] Hanlab, "eric (hera (human enhancer rna atlas))." https://hanlab.uth.edu/HeRA/download, jan 2008.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[47] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *the Journal of machine Learning research*, vol. 9, pp. 1871–1874, 2008.

[48] S. Han, C. Qubo, and H. Meng, "Parameter selection in svm with rbf kernel function," in *World Automation Congress 2012*, pp. 1–4, IEEE, 2012.

[49] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.

[50] Z. Zuo, H. Hu, Q. Xu, X. Luo, D. Peng, K. Zhu, Q. Zhao, Y. Xie, and J. Ren, "Bbcancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers," *Nucleic Acids Research*, vol. 48, no. D1, pp. D789–D796, 2020.

[51] L. Mercatali, V. Valenti, D. Calistri, S. Calpona, G. Rosti, S. Folli, M. Gaudio, G. Frassineti, D. Amadori, and E. Flamini, "Rt-pcr determination of maspin and mammaglobin b in peripheral blood of healthy donors and breast cancer patients," *Annals of oncology*, vol. 17, no. 3, pp. 424–428, 2006.

[52] S. Li, D. Wu, H. Jia, and Z. Zhang, "Long non-coding rna lrrc75a-as1 facilitates triple negative breast cancer cell proliferation and invasion

via functioning as a cerna to modulate baalc," *Cell death & disease*, vol. 11, no. 8, pp. 1–12, 2020.

[53] T. Z. Parris, A. Kovács, L. Aziz, S. Hajizadeh, S. Nemes, M. Semaan, E. Forssell-Aronsson, P. Karlsson, and K. Helou, "Additive effect of the azgp1, pip, s100a8 and ube2c molecular biomarkers improves outcome prediction in breast carcinoma," *International journal of cancer*, vol. 134, no. 7, pp. 1617–1629, 2014.

[54] T. Smirnova, L. Bonapace, G. MacDonald, S. Kondo, J. Wyckoff, H. Ebersbach, B. Fayard, A. Doelemeyer, M.-M. Coissieux, M. R. Heideman, *et al.*, "Serpin e2 promotes breast cancer metastasis by remodeling the tumor matrix and polarizing tumor associated macrophages," *Oncotarget*, vol. 7, no. 50, p. 82289, 2016.

[55] J. Lin, L. Goldstein, A. Nesbit, and M. Y. Chen, "Influence of hormone receptor status on spinal metastatic lesions in patients with breast cancer," *World neurosurgery*, vol. 85, pp. 42–48, 2016.

[56] K. Ratajczak-Wielgomas, J. Grzegrzolka, A. Piotrowska, R. Matkowski, A. Wojnar, J. Rys, M. Ugorski, and P. Dziegiel, "Expression of periostin in breast cancer cells," *International Journal of Oncology*, vol. 51, no. 4, pp. 1300–1310, 2017.

[57] D. E. Ramirez-Ardila, K. Ruigrok-Ritstier, J. C. Helmijr, M. P. Look, S. van Laere, L. Dirix, E. M. Berns, and M. P. Jansen, "Lrg1 mrna expression in breast cancer associates with pik3ca genotype and with aromatase inhibitor therapy outcome," *Molecular oncology*, vol. 10, no. 8, pp. 1363–1373, 2016.

[58] G. Xiong, L. Deng, J. Zhu, P. G. Rychahou, and R. Xu, "Prolyl-4-hydroxylase α subunit 2 promotes breast cancer progression and metastasis by regulating collagen deposition," *BMC cancer*, vol. 14, no. 1, pp. 1–12, 2014.

[59] K. A. Snyder, M. R. Hughes, B. Hedberg, J. Brandon, D. C. Hernaez, P. Bergqvist, F. Cruz, K. Po, M. L. Graves, M. E. Turvey, *et al.*, "Podocalyxin enhances breast tumor growth and metastasis and is a target for monoclonal antibody therapy," *Breast cancer research*, vol. 17, no. 1, pp. 1–14, 2015.

[60] Shana, "Mttpisdifferentiallyexpressedinbrainmetastatichumanbreast-cancer." https://osf.io/ajns7, feb 2021.

[61] T. Leung, R. Rajendran, S. Singh, R. Garva, M. Krstic-Demonacos, and C. Demonacos, "Cytochrome p450 2e1 (cyp2e1) regulates the response to oxidative stress and migration of breast cancer cells," *Breast Cancer Research*, vol. 15, no. 6, pp. 1–12, 2013.

[62] D.-n. Dai, Y. Li, B. Chen, Y. Du, S.-b. Li, S.-x. Lu, Z.-p. Zhao, A.-j. Zhou, N. Xue, T.-l. Xia, *et al.*, "Elevated expression of cst1 promotes breast cancer progression and predicts a poor prognosis," *Journal of Molecular Medicine*, vol. 95, no. 8, pp. 873–886, 2017.

[63] H. Zhang, X. Sun, Y. Lu, J. Wu, and J. Feng, "Dna-methylated gene markers for colorectal cancer in tcga database," *Experimental and therapeutic medicine*, vol. 19, no. 4, pp. 3042–3050, 2020.

[64] D. Huang, C. Du, D. Ji, J. Xi, and J. Gu, "Overexpression of lamc2 predicts poor prognosis in colorectal cancer patients and promotes cancer cell proliferation, migration, and invasion," *Tumor Biology*, vol. 39, no. 6, p. 1010428317705849, 2017.

[65] K. Yoneda, H. Iida, H. Endo, K. Hosono, T. Akiyama, H. Takahashi, M. Inamori, Y. Abe, M. Yoneda, K. Fujita, *et al.*, "Identification of cystatin sn as a novel tumor marker for colorectal cancer," *International journal of oncology*, vol. 35, no. 1, pp. 33–40, 2009.

[66] S. Li, J. Huang, M. Qin, J. Zhang, and C. Liao, "High expression of cdca7 predicts tumor progression and poor prognosis in human colorectal cancer," *Molecular Medicine Reports*, vol. 22, no. 1, pp. 57–66, 2020.

[67] N. Jagadish, D. Parashar, N. Gupta, S. Agarwal, V. Suri, R. Kumar, V. Suri, T. C. Sadasukhi, A. Gupta, A. S. Ansari, *et al.*, "Heat shock protein 70–2 (hsp70-2) is a novel therapeutic target for colorectal cancer and is associated with tumor growth," *BMC cancer*, vol. 16, no. 1, pp. 1–13, 2016.

[68] F. Li, Q. Li, and X. Wu, "Construction and analysis for differentially expressed long non-coding rnas and microrna mediated competing

endogenous rna network in colon cancer," *PloS one*, vol. 13, no. 2, p. e0192494, 2018.

[69] Z.-x. Liang, H.-s. Liu, F.-w. Wang, L. Xiong, C. Zhou, T. Hu, X.-w. He, X.-j. Wu, D. Xie, X.-r. Wu, *et al.*, "Lncrna rpph1 promotes colorectal cancer metastasis by interacting with tubb3 and by promoting exosomes-mediated macrophage m2 polarization," *Cell death & disease*, vol. 10, no. 11, pp. 1–17, 2019.

[70] K. Munakata, M. Uemura, I. Takemasa, M. Ozaki, M. Konno, J. Nishimura, T. Hata, T. Mizushima, N. Haraguchi, S. Noura, *et al.*, "Scgb2a1 is a novel prognostic marker for colorectal cancer associated with chemoresistance and radioresistance," *International journal of oncology*, vol. 44, no. 5, pp. 1521–1528, 2014.

[71] Y. Ren, Y. Li, W. Zhang, K. Yang, J. Li, Y. Hu, Z. Zuo, C. Xu, Y. Pan, and X. Zhang, "Mir-4746 inhibits the proliferation of colorectal cancer cells in vitro and in vivo by targeting ccnd1," *Biochemical and Biophysical Research Communications*, 2022.

[72] R. Mazzolini, H. Dopeso, S. Mateo-Lozano, W. Chang, P. Rodrigues, S. Bazzocco, H. Alazzouzi, S. Landolfi, J. Hernández-Losa, E. Andretta, *et al.*, "Brush border myosin ia has tumor suppressor activity in the intestine," *Proceedings of the National Academy of Sciences*, vol. 109, no. 5, pp. 1530–1535, 2012.

[73] J. Liu, C. Lu, M. Xiao, F. Jiang, L. Qu, and R. Ni, "Long non-coding rna snhg20 predicts a poor prognosis for hcc and promotes cell invasion by regulating the epithelial-to-mesenchymal transition," *Biomedicine & Pharmacotherapy*, vol. 89, pp. 857–863, 2017.

[74] R. Liu, P. Deng, Y. Zhang, Y. Wang, and C. Peng, "Circ_0082182 promotes oncogenesis and metastasis of colorectal cancer in vitro and in vivo by sponging mir-411 and mir-1205 to activate the wnt/β-catenin pathway," *World Journal of Surgical Oncology*, vol. 19, no. 1, pp. 1–14, 2021.

[75] Y. Huang, L.-Z. Li, C. Z.-Y. Zhang, C. Yi, L.-L. Liu, X. Zhou, G.-B. Xie, M.-Y. Cai, Y. Li, and J.-P. Yun, "Decreased expression of zinc-alpha2-glycoprotein in hepatocellular carcinoma associates with poor

prognosis," *Journal of translational medicine*, vol. 10, no. 1, pp. 1–10, 2012.

[76] Y.-H. Wang, C.-L. Liu, W.-C. Chiu, Y.-C. Twu, and Y.-J. Liao, "Hmgcs2 mediates ketone production and regulates the proliferation and metastasis of hepatocellular carcinoma," *Cancers*, vol. 11, no. 12, p. 1876, 2019.

[77] P. Sancho and I. Fabregat, "Nadph oxidase nox1 controls autocrine growth of liver tumor cells through up-regulation of the epidermal growth factor receptor pathway," *Journal of Biological Chemistry*, vol. 285, no. 32, pp. 24815–24824, 2010.

[78] H. S. Seol, S. E. Lee, J. S. Song, J.-K. Rhee, S. R. Singh, S. Chang, and S. J. Jang, "Complement proteins c7 and cfh control the stemness of liver cancer cells via lsf-1," *Cancer letters*, vol. 372, no. 1, pp. 24–35, 2016.

[79] J. Oliva, F. Bardag-Gorce, A. Lin, B. A. French, and S. W. French, "The role of cytokines in ubd promoter regulation and mallory-denk body-like aggresomes," *Experimental and molecular pathology*, vol. 89, no. 1, pp. 1–8, 2010.

[80] X. Yan, M. Yao, X. Wen, Y. Zhu, E. Zhao, X. Qian, X. Chen, W. Lu, Q. Lv, L. Zhang, *et al.*, "Elevated apolipoprotein b predicts poor post-surgery prognosis in patients with hepatocellular carcinoma," *OncoTargets and therapy*, vol. 12, p. 1957, 2019.

[81] X. Wang, Y. Gong, T. Deng, L. Zhang, X. Liao, C. Han, C. Yang, J. Huang, Q. Wang, X. Song, *et al.*, "Diagnostic and prognostic significance of mrna expressions of apolipoprotein a and c family genes in hepatitis b virus-related hepatocellular carcinoma," *Journal of Cellular Biochemistry*, vol. 120, no. 10, pp. 18246–18265, 2019.

[82] S. Lu, W. Zhou, H. Wei, L. He, and L. Li, "Mtbp promotes the invasion and metastasis of hepatocellular carcinoma by enhancing the mdm2-mediated degradation of e-cadherin," *Digestive diseases and sciences*, vol. 60, no. 12, pp. 3681–3690, 2015.

[83] G. Chen, X. Fan, Y. Li, L. He, S. Wang, Y. Dai, C. Bin, D. Zhou, and H. Lin, "Promoter aberrant methylation status of adra1a is associated

with hepatocellular carcinoma," *Epigenetics*, vol. 15, no. 6-7, pp. 684–701, 2020.

[84] W. Zhang, Q. Cui, W. Qu, X. Ding, D. Jiang, and H. Liu, "Trim58/cg26157385 methylation is associated with eight prognostic genes in lung squamous cell carcinoma," *Oncology Reports*, vol. 40, no. 1, pp. 206–216, 2018.

[85] Y. Moon, G. Rao, J. Kim, H. Shim, K. Park, S. An, B. Kim, P. Steeg, S. Sarfaraz, L. Changwoo Lee, *et al.*, "Lamc2 enhances the metastatic potential of lung adenocarcinoma," *Cell Death & Differentiation*, vol. 22, no. 8, pp. 1341–1352, 2015.

[86] P. Zhang, B. Kang, G. Xie, S. Li, Y. Gu, Y. Shen, X. Zhao, Y. Ma, F. Li, J. Si, *et al.*, "Genomic sequencing and editing revealed the grm8 signaling pathway as potential therapeutic targets of squamous cell lung cancer," *Cancer letters*, vol. 442, pp. 53–67, 2019.

[87] R. Sun, X. Meng, W. Wang, B. Liu, X. Lv, J. Yuan, L. Zeng, Y. Chen, B. Yuan, and S. Yang, "Five genes may predict metastasis in non-small cell lung cancer using bioinformatics analysis," *Oncology Letters*, vol. 18, no. 2, pp. 1723–1732, 2019.

[88] S. Borgquist, T. Butt, P. Almgren, D. Shiffman, T. Stocks, M. Orho-Melander, J. Manjer, and O. Melander, "Apolipoproteins, lipids and risk of cancer," *International journal of cancer*, vol. 138, no. 11, pp. 2648–2656, 2016.

[89] Y. Zhou, B. Xu, Y. Zhou, J. Liu, X. Zheng, Y. Liu, H. Deng, M. Liu, X. Ren, J. Xia, *et al.*, "Identification of key genes with differential correlations in lung adenocarcinoma," *Frontiers in Cell and Developmental Biology*, vol. 9, 2021.

[90] M. Chaplet, L. De Leval, D. Waltregny, C. Detry, G. Fornaciari, G. Bevilacqua, L. Fisher, V. Castronovo, and A. Bellahcène, "Dentin matrix protein 1 is expressed in human lung cancer," *Journal of Bone and Mineral Research*, vol. 18, no. 8, pp. 1506–1512, 2003.

[91] Y. Wu, K. Cheng, W. Liang, and X. Wang, "lncrna rpph1 promotes non-small cell lung cancer progression through the mir-326/wnt2b axis," *Oncology Letters*, vol. 20, no. 4, pp. 1–1, 2020.

[92] I. Y. Kim, D.-H. Lee, H.-J. Ahn, H. Tokunaga, W. Song, L. M. Devereaux, D. Jin, T. K. Sampath, and R. A. Morton, "Expression of bone morphogenetic protein receptors type-ia,-ib, and-ii correlates with tumor grade in human prostate cancer tissues," *Cancer research*, vol. 60, no. 11, pp. 2840–2844, 2000.

[93] Z. Shang, Y. Niu, Q. Cai, J. Chen, J. Tian, S. Yeh, K.-P. Lai, and C. Chang, "Human kallikrein 2 (klk2) promotes prostate cancer cell growth via function as a modulator to promote the ara70-enhanced androgen receptor transactivation," *Tumor Biology*, vol. 35, no. 3, pp. 1881–1890, 2014.

[94] H. Schwarzenbach, F. K.-H. Chun, I. Müller, C. Seidel, K. Urban, A. Erbersdobler, H. Huland, K. Pantel, and M. G. Friedrich, "Microsatellite analysis of allelic imbalance in tumour and blood from patients with prostate cancer," *BJU international*, vol. 102, no. 2, pp. 253–258, 2008.

[95] S. Maruta, Y. Miyata, Y. Sagara, S. Kanda, T. Iwata, S.-i. Watanabe, H. Sakai, T. Hayashi, and H. Kanetake, "Expression of matrix metalloproteinase-10 in non-metastatic prostate cancer: Correlation with an imbalance in cell proliferation and apoptosis," *Oncology Letters*, vol. 1, no. 3, pp. 417–421, 2010.

[96] S. Zhang, B. Li, W. Tang, L. Ni, H. Ma, M. Lu, and Q. Meng, "Effects of connective tissue growth factor on prostate cancer bone metastasis and osteoblast differentiation," *Oncology Letters*, vol. 16, no. 2, pp. 2305–2311, 2018.

[97] L. van der Heul-Nieuwenhuijsen, N. Dits, W. Van Ijcken, D. de Lange, and G. Jenster, "The foxf2 pathway in the human prostate stroma," *The Prostate*, vol. 69, no. 14, pp. 1538–1547, 2009.

[98] H. van Poppel, A. Haese, M. Graefen, A. de la Taille, J. Irani, T. de Reijke, M. Remzi, and M. Marberger, "The relationship between prostate cancer gene 3 (pca3) and prostate cancer significance," *BJU international*, vol. 109, no. 3, pp. 360–366, 2012.

[99] D. Chaves-Moreira, M. A. Mitchell, C. Arruza, P. Rawat, S. Sidoli, R. Nameki, J. Reddy, R. I. Corona, L. K. Afeyan, I. A. Klein, *et al.*,

"The transcription factor pax8 promotes angiogenesis in ovarian cancer through interaction with sox17," *Science signaling*, vol. 15, no. 728, p. eabm2496, 2022.

[100] Q. Yan, F. Wang, Y. Miao, X. Wu, M. Bai, X. Xi, and Y. Feng, "Sex-determining region y-box3 (sox3) functions as an oncogene in promoting epithelial ovarian cancer by targeting src kinase," *Tumor Biology*, vol. 37, no. 9, pp. 12263–12271, 2016.

[101] E. J. Norris, Q. Zhang, W. D. Jones, D. DeStephanis, A. P. Sutker, C. A. Livasy, R. N. Ganapathi, D. L. Tait, and M. K. Ganapathi, "Increased expression of neurotensin in high grade serous ovarian carcinoma with evidence of serous tubal intraepithelial carcinoma," *The Journal of pathology*, vol. 248, no. 3, pp. 352–362, 2019.

[102] A. A. Lo, J. Johnston, J. Li, D. Mandikian, M. Hristopoulos, R. Clark, D. Nickles, W.-C. Liang, K. Hötzel, D. Dunlap, *et al.*, "Anti-lypd1/cd3 t-cell-dependent bispecific antibody for the treatment of ovarian cancer," *Molecular Cancer Therapeutics*, vol. 20, no. 4, pp. 716–725, 2021.

[103] A. A. Naser, T. Miyazaki, J. Wang, S. Takabayashi, T. Pachoensuk, and T. Tokumoto, "Mc4r mutant mice develop ovarian teratomas," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[104] M. Ramakrishna, L. H. Williams, S. E. Boyle, J. L. Bearfoot, A. Sridhar, T. P. Speed, K. L. Gorringe, and I. G. Campbell, "Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis," *PloS one*, vol. 5, no. 4, p. e9983, 2010.