TEP RNA-based 'liquid biopsies' for diagnosing multiple types of cancer

Shubhangi Kaushik

Abstract

The goal of the project: Train a machine learning model to accurately classify healthy and cancer samples, from RNA-sequencing data of blood platelet samples.

Main results of the project: Confusion matrix and important features were obtained for the study and compared with the previous studies.

Estimated working hours: 20 hours(6 hour:- reading the research paper, 8 hours:- understanding the data, cleaning it and building the model, 6 hours:-writing down the report)

Challenges Faced: Tried to run DESeq on Python first but the kernel would stop working every time we tried to import 'ro' module. Therefore, due to unfamiliarity with using DESeq2 on Python, R-script was used instead for the preprocessing of the data.

Goal

The project aimed to develop a machine learning-based classification model for cancer classification using RNA-sequencing data from blood platelet samples, and compare it to the works from Best et al[1] and, Zhang et al.[5] for the same data set. Confusion matrix, measures like accuracy, and graphs like ROC curves were used to evaluate the model. Important features were also obtained and compared with Zhang's study. This study is important because it highlights the role of platelets in tumor biology and metastasis. Understanding the hidden patterns using Data Science can help find valuable insights leading to the development of new methods related to precision oncology and targeted treatments.

Data and Preprocessing

Data

The data analyzed is RNA-sequencing data from 283 blood platelet samples, including 228 tumor-educated platelets (TEP) samples from patients with six different malignant tumors. The rest of the 55 samples were from healthy individuals' platelets. Blood platelets were isolated from blood samples using centrifugation. sing the Truseq Nano DNA Sample Preparation Kit, total RNA was isolated from the platelet pellet. It was then subjected to cDNA synthesis and SMARTer amplification, followed by fragmentation by Covaris shearing. The processed data was sequenced using the Illumina Hiseq 2500 platform.

There were 57000s rows and 286 columns, the columns described the cancer and healthy samples, whereas the rows described the level expression of different genes for the samples. The transcriptomics data was downloaded in text format from GEO database with accession ID: GSE68086.

Preprocessing

Here, the methods involving DESeq2[3] were chosen to be applied to the transcriptomics data. In the first step, the samples which were labeled as 'Type.Unknown' were removed from the data set. Then the labels which were healthy or control, i.e. 'HD' or 'Control' were marked as 'Non Cancer' and the rest of the labels were marked as 'Cancer'. All the gene ids, sample names along with the columns representing whether it is 'Non Cancer' or 'Cancer', and the data without the gene ids were stored in separate data frames.

The rows in the data without gene ids were observed to have many zero values, and values as low as 5 and below. It meant that many genes had no or very low expression for many samples hinting at the presence of noise in the data and were excluded. Ultimately, what was left out of 57000s original rows, were the 668 rows consisting of gene expression values for all samples above 5.

In order to run the analysis on DESeq2, the data frame consisting of the sample names was considered to be the metadata. The design consisted of the column names in the metadata, whereas, the count data consisted of the data without gene ids in a matrix form.

After running the analysis with DESeq2, the gene ids with p values above 0.05 and log2foldchange above 1 were filtered. The filtered data with 286 columns and 90 rows were transposed and labeled as "1" for Cancer and "0" for 'Non Cancer' state. The classes were found to be imbalanced. "1" had 4 times more occurrence than value "0", hence Synthetic Minority Oversampling Technique[2] was used to increase the sample size of "0". The new data was synthetically generated based on k-nearest neighbors of the minority class. Later, the data was split into 'X' and 'Y' set, where 'Y' consisted of the 'labels', and 'X' or the 'predictor variables' consisted of 90 feature columns. The 'X' set was scaled because many columns were not on the same scale.

After all the preprocessing, the X and Y data sets were split into test and train set with 154 and 308 rows respectively.

Methods

The analysis started with finding the optimized hyper-parameters for the chosen machine learning algorithm, Random Forest(RF). For the hyperparameter tuning, an array of integers were tested by fitting the model, where each value of the array denoted the number of trees. The model worked the best with 150 trees having AUC score of 0.972. The number of trees was used as 150 for training the final model.

0.1 Random Forest

Random Forest[4] is a bagging method that uses Decision Trees. Decision Trees are created for the combination of random subsets of observations and random subsets

of features. The final prediction is made from the majority vote after each tree casts a vote.

The choice of predictor variables used to split nodes in decision trees is determined by specific optimization criteria. In classification, one commonly used criterion is entropy(E),

$$E = -\sum_{i=1}^{c} p_i \log(p_i) \tag{1}$$

where c represents the number of unique classes and p_i is the prior probability of each class. The goal is to maximize the entropy value, aiming to gain the most information with each split in the decision tree.

0.2 Evaluation

The machine learning model was evaluated using the prediction probabilities calculated using K-Fold cross-validation technique where the number of folds was set to 10. Based on the prediction probabilities, the class was predicted for both train and test datasets. These predictions were evaluated using the following:

0.2.1 Confusion Matrix

A confusion matrix is a table that summarizes the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions. It is used to calculate evaluation measures like Accuracy, True Positive Rate, etc.

Predicted/Actual	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Table 1: Confusion Matrix

0.2.2 ROC Curve

(Receiver Operating Characteristic) Curve[**?**] for each class was also generated using the probabilities. It plots the true positive rate against the false positive rate at different classification thresholds.

0.2.3 Number of features against Accuracy graph

The feature size was modified from 1 to 90. The feature vectors were sampled for each feature size, and the y values were predicted based on a 10-fold cross val predict approach. The process repeated 10 times for each feature size. Average accuracy was then calculated for each feature size, and a graph was plotted to depict the relationship between feature size and average accuracy.

Results and Discussion

We evaluated our models using Confusion Matrix and ROC Curve for both Test and Training data sets. The results were compared with the previous studies done by Best

et al.[1] and Zhang et al.[5]. We also compared the most important features that we got from our model with the ones that were obtained in Zhang et al.'s research paper. The performance of the random forest in this study was assessed using the Confusion Matrix for the training set(Figure 1) and test set(Figure 4). They were compared with the results from Best et al.'s paper(Figure 2, Figure 5).

Predicted/Actual	Positive	Negative
Positive	149	12
Negative	5	142

Figure 1: Confusion Matrix with an accuracy of 94.4% for Train set from Random Forest model

Predicted/Actual	Positive	Negative
Positive	131	3
Negative	5	36

Figure 2: Confusion Matrix with an accuracy of 95% for Train set from Best et al.'s SVM model.

Figure 3: Comparison of Confusion matrix of Train set for this project and Best et al. study

Predicted/Actual	Positive	Negative
Positive	72	8
Negative	5	69

Figure 4: Confusion Matrix with an accuracy of 91.5% for Test set from Random Forest model

Predicted/Actual	Positive	Negative
Positive	89	1
Negative	3	15

Figure 5: Confusion Matrix with accuracy of 96% for Test set from Best et al.'s SVM model.

Figure 6: Comparison of Confusion matrix of Test set for this project and Best et all study

ROC curves were also compared for both models. ROC curve was plotted for both train and test sets for the Random Forest model(Figure 7).

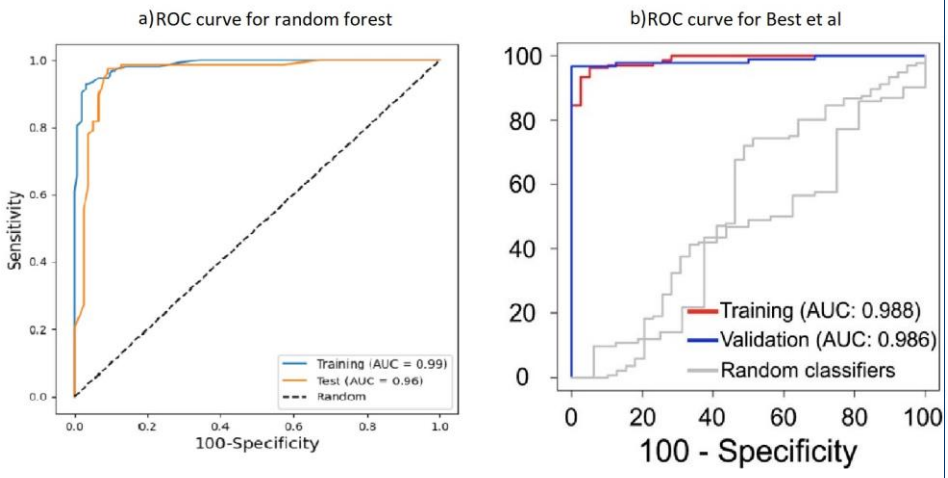

Figure 7: **a)** the ROC curve obtained from the random forest method. The blue line represents the training set, and the orange line, represents the test set. The auc score for the training set was 99% where as for the test set it was 96%. **b)** The ROC curve is for the SVM LOOCV method used in the Best et al study, the red line is for the train set and the blue line for the validation set. the auc score for the training and validation set is 98.8% and 98.6%. respectively.

For comparison with the results from Zhang et al.'s paper(Figure 9), an average accuracy line graph was plotted over 90 preprocessed features for Random Forest(Figure 8). Based on the comparison, Random Forest seemed to be working better with 95% maximum accuracy for 65 features.

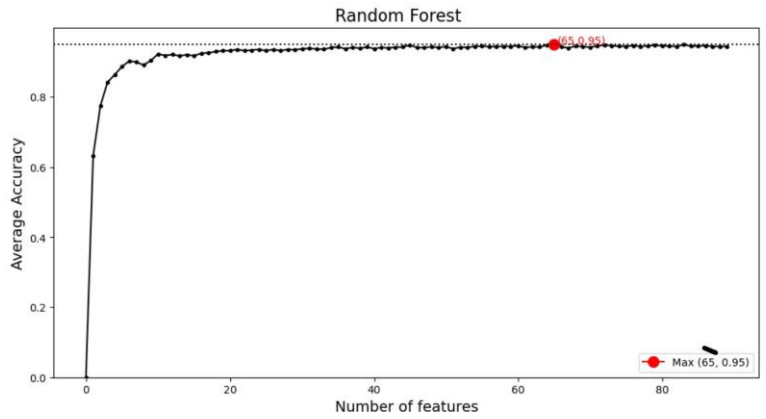

Figure 8: Relationship between the number of features and the accuracy for the Random Forest model. Each data point on the graph represents the average accuracy of the model for a specific number of features, "n" where n ranges from 1 to 90. To obtain these values, the model was created 10 times for each "n". In each iteration, "n" were randomly sampled. The average accuracy for each "n" was calculated considering all 10 results/iteration.

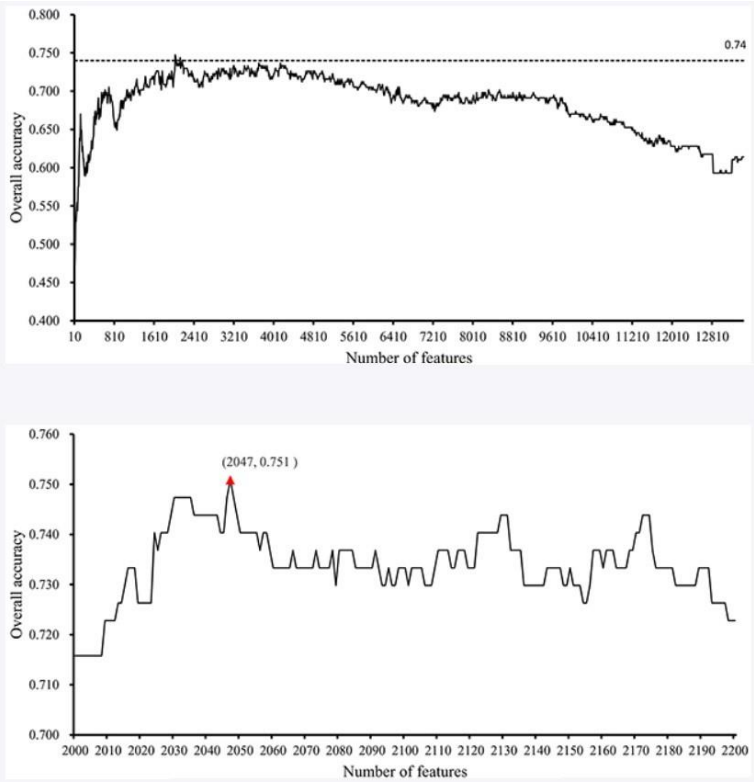

Figure 9: Overall Accuracy plot based on IFS method from Zhang et al.'s paper.
The Y-axis represents the overall accuracy, and the X-axis represents the number of features used for classification. Feature subsets with a multiple of ten features were analyzed for SVM models for all the features, and overall accuracies were calculated for each. 75% was the best accuracy for their model overall.
The top figure shows the accuracies for the number of features in the range [1, 13,445]. 13,445 was the count of all of their features.
The bottom figure shows a zoomed-in view of the top figure for the range [2,000 to 2,200] allowing for a more detailed analysis within the range of interest from the top figure.

Based on the comparison of the ROC curves in Figure 7, the SVM LOOCV used in Best et al, turned out to be doing better than the Random forest used in this project. However, this could be improved by changing the random seed value and number of trees during the cross-validation. When compared with Zhang's study the random forest used in this project turned out to be better. And when the most important features were identified for our model and compared to the ones obtained from Zhang,

6 of the important features were observed to be the same. This was an indication that the preprocessing worked fine with the data.

Random Forest	Importance	Zhang Study 3	Importance
ENSG00000089009	0.086294	ENSG00000155657	0.416
ENSG00000163682	0.057992	ENSG00000008988	0.407
ENSG00000005961	0.052483	ENSG00000177600	0.405
ENSG00000035403	0.052274	ENSG00000211772	0.396
ENSG00000177600	0.042513	ENSG00000168028	0.393
ENSG00000108846	0.038821	ENSG00000142534	0.384
ENSG00000005249	0.030065	ENSG00000142676	0.381
ENSG00000168028	0.026965	ENSG00000105193	0.380
ENSG00000162909	0.025972	ENSG00000160654	0.379
ENSG00000138326	0.022634	ENSG00000168421	0.373
ENSG00000142676	0.022292	ENSG00000139193	0.369
ENSG00000162368	0.021546	ENSG00000131469	0.368
ENSG00000235162	0.021043	ENSG00000162368	0.368
ENSG00000171858	0.020720	ENSG00000071082	0.367
ENSG00000071082	0.020382	ENSG00000149311	0.367
ENSG00000114391	0.018795	ENSG00000149806	0.366
ENSG00000196924	0.017209	ENSG00000109475	0.366
ENSG00000198858	0.016071	ENSG00000089009	0.366

Table 2: The top 18 feature importance from this project and Zhang study

Common Features	Position in this Study	Position in Zhang study
ENSG00000177600	5 Metabolite	3 Metabolite
ENSG00000168028	8 Metabolite	5 Metabolite
ENSG00000142676	11 Metabolite	7 Metabolite
ENSG00000163682	2 Metabolite	13 Metabolite
ENSG00000071082	15 Metabolite	14 Metabolite
ENSG00000089009	1 Metabolite	18 Metabolite

Table 3: From the top 18 feature importance obtained from this project and Zhang study as mentioned in the table2, these were the common metabolites found, along with their position in the table2

Overall, preprocessing was the most important part of the metabolomics data analysis. Already available bioinformatics tools like DESeq2 are capable of assisting in providing insightful results. In the end, Random Forest worked relatively well with the data that was obtained after preprocessing using DESeq2.

References

1. Myron G Best, Nik Sol, Irsan Kooi, Jihane Tannous, Bart A Westerman, Fran¸cois Rustenburg, Pepijn Schellen, Heleen Verschueren, Edward Post, Jan Koster, et al. Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer cell*, 28(5):666–676, 2015.
2. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
3. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.
4. Matthias Schonlau and Rosie Yuyan Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.
5. Yu-Hang Zhang, Tao Huang, Lei Chen, YaoChen Xu, Yu Hu, Lan-Dian Hu, Yudong Cai, and Xiangyin Kong. Identifying and analyzing different cancer subtypes using rna-seq data of blood platelets. *Oncotarget*, 8(50):87494–87511, 2017.