RESEARCH

# Paper 4 in applied machine learning

Aakash Nepal and Cuong Gia Pham

Full list of author information is available at the end of the article
*Equal contributor

**Abstract**

**Goal of the project:** Develop a Random Forest(RF) classifier to distinguish between healthy and cancer samples. Compare performance with Support Vector Machine model(SVM) from the papers by Best[1] and Zhang[2]. Compare the top 18 features to the Zhang paper.

**Main results of the project:** By incorporating the sample cloning step, which results in a balanced distribution within the data groups, we can achieve a reliable result. Besides that, feature selection is also important to tackle the overfitting phenomenon from the training process.

**Personal key learning :**

1 Pham Gia Cuong: generate ROC with a confidence interval, t-SNE, GridSearchCV, and feature selection.

2 Aakash Nepal: was introduced to GridSearchCV, t-SNE ,and learned more about transcripomics and random forest.

**Estimation working hours:**

1 Pham Gia Cuong: 7 hours per week

2 Aakash Nepal: 7 hours per week.

**Project evaluation:** 1

**Number of words :** Approx. 1500(without abstract, captions and appendix)

## 1 Goal of the project

The goal of this project is to develop and test a Random Forest(RF) classifier for differentiating between healthy and cancer samples. The generated classifier is evaluated using a confusion matrix, accuracy, and ROC curve with AUC and compared with SVM models from the Best[1] and Zhang papers[2]. The top 18 features are identified using feature importance analysis and compared to the Zhang paper[2]. By establishing an effective machine learning classifier for identifying healthy and cancer samples, we can help improve the identification of cancer and perhaps aid in the development of individualized therapies.

## 2 Data and Preprocessing

### 2.1 Data

We used the public count matrix data GSE68086 from the NCBI database for this project. The data is originally RNA-sequencing data collected from 283 blood platelet samples, including 228 tumor-educated blood platelets (TEP) samples with six different malignant tumors (non-small cell lung cancer, colorectal cancer, pancreatic cancer, glioblastoma, breast cancer, and hepatobiliary carcinomas) and 55 healthy individuals. This dataset highlights the ability of TEP RNA-based 'liquid biopsies in patients with several types of cancer, including the ability for pan-cancer, multiclass cancer, and companion diagnostics. The downloaded count matrix is a single table containing the counts for all samples, with the genes in rows and the samples in columns, which means that each cell in the table indicates a number of reads that are mapped on a specific gene (row) of a specific sample (column). In this project, the count matrix data contains 285 samples (columns) and 57736 ensemble gene ids (rows).

### 2.2 Preprocessing

The above-mentioned count matrix data from NCBI database includes too many ensemble genes and some of them are even meaningless to the disease. Therefore, we have used gene dispersion analysis to purify the data. Firstly, we removed all the genes, which are mapped with a number of reads smaller than 5 in all samples. Those genes are supposed to be junk and a small number of reads are mapped on them by accident. The filtered count matrix is then transformed to be satisfied for the gene dispersion analysis function of DESeq2, a popular bioinformatics tool used for differential gene analysis. Based on the Benjamin Hochberg adjusted p-values from the negative binomial Wald Test of DESeq2, we keep all genes, that have the adjusted p-values smaller than 0,05 and log2 fold change is greater than 1 or smaller than -1. In the end, we have in total 86 significant genes.

Besides a large number of genes from the data, there is also a problem with the imbalance between cancer and non-cancer groups. The number of cancer samples is enormous around 4 times higher than the number of non-cancer samples. Hence, firstly all the non-cancer samples have been gathered. Afterward, those samples have been randomly picked and cloned. Adding the same samples to the data doesn't impact the efficiency of the model's result. Hence, some values in a range from -20 to 20 have been randomly generated and added those values in all the features of each cloned sample. In the end, there are 231 samples in each class.

## 3 Methods

The RF algorithm is a powerful machine-learning method used for tasks like regression and classification. It is based on an ensemble approach, where multiple decision trees, known as estimators, are combined to make predictions. This aggregation of predictions helps to improve the accuracy of the model.

To optimize the RF model, we employed the GridSearchCV function from the scikitlearn library. GridSearchCV systematically explores various combinations of hyperparameters and identifies the best parameter set based on a specified evaluation criterion. Once the optimal parameters are determined, the model is retrained using those parameters.

To assess the performance of the model, we utilized the confusion matrix. The confusion_matrix function from scikitlearn was employed to generate the matrix, and the ConfusionMatrixDisplay function was used to visually represent it in a more intuitive format. The confusion matrix provides insights into the accuracy of the model by showing the number of true positives, true negatives, false positives, and false negatives.

In addition to the confusion matrix, we also employed the ROC curve to evaluate the model's performance and assess overfitting. The ROC curve is a graphical representation of the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds. It allows us to visualize the tradeoff between sensitivity and specificity and determine the model's discriminatory power.
To have more reliability on the result, we used also the cross-validation method to plot out the confidence interval of AUC.
The feature importance analysis was performed using the build-in "feature_importances_" attribute of the RF. The extracted important features were then retrieved from the preprocessed dataset to obtain further information.
By employing these evaluation techniques, we gain a comprehensive understanding of the RF model's performance and its ability to generalize to unseen data.

## 4 Results and Discussion

### 4.1 Model evaluation

After filtering the features with the help of gene dispersion analysis and "cloning" non-cancer samples, the data includes 231 samples in each class and each sample has 86 features. The data is split into a train set and a test set with a ratio of 2:1. The above-mentioned hyperparameters tuning step gives us the RF that has a maximum depth of 10, criterion gini, and max features auto. The result in Figure 1 illustrates the ROC curves of the model on the train (left) and test (right) sets. Comparing the area under the ROC curve (AUC) of the RF model to that of the Best paper[1] reveals that the RF model outperforms with a mean AUC of 0.98 on the train set and 0.96 on the test set. Both ROC curves fall within the AUC confidence interval, indicating a reasonably reliable result. However, examining the t-SNE plot (see Appendix A) , we observe that the original data is already well separated, suggesting that our model may not readily generalize to unseen data. Additionally, there is a distinct cluster on the left side of the plot that is completely isolated from the rest. This indicates that the model might be more effective when classifying data with more than just two categories.

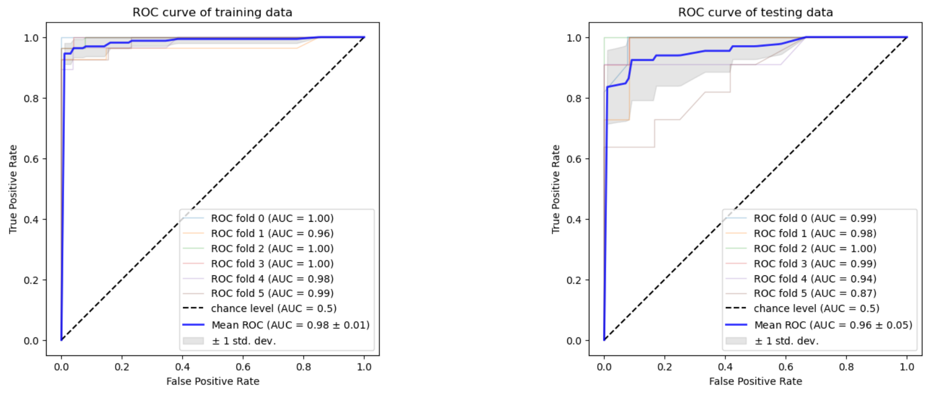

**Figure 1** Random forest's ROC curve for the train set (left) and for test set (right)

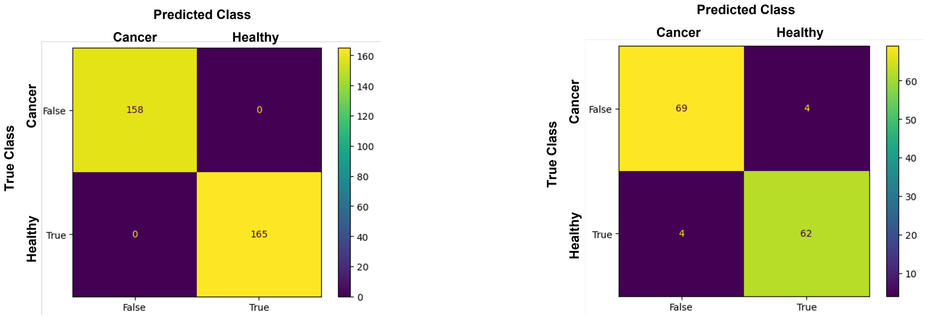

**Figure 2** Random forest's confusion matrix for the train set (left) and for the test set (right)

Besides that, the confusion matrix in Figure 2 shows us also an unusual result in the train set, in which the model achieved a perfect prediction on the train set. And hence, the confusion matrix has all the samples correctly classified, resulting in zero false positives (FP) and false negatives (FN). The diagonal elements of the confusion matrix (TP and TN) contain the counts of correctly predicted samples. In the test set, the confusion matrix indicates that there are 69 samples that belong to the cancer group and are correctly classified as positive by the model, 62 samples that belong to the non-cancer group and are correctly classified as negative by the model, 4 samples that belong to the non-cancer group but are incorrectly classified as cancer by the model and 4 samples that belong to the cancer group but are incorrectly classified as non-cancer by the model. From that, we can observe that the model performs well in terms of both true positives and true negatives, as indicated by the high counts in the corresponding cells. This suggests that the model effectively identifies cancer and non-cancer samples.

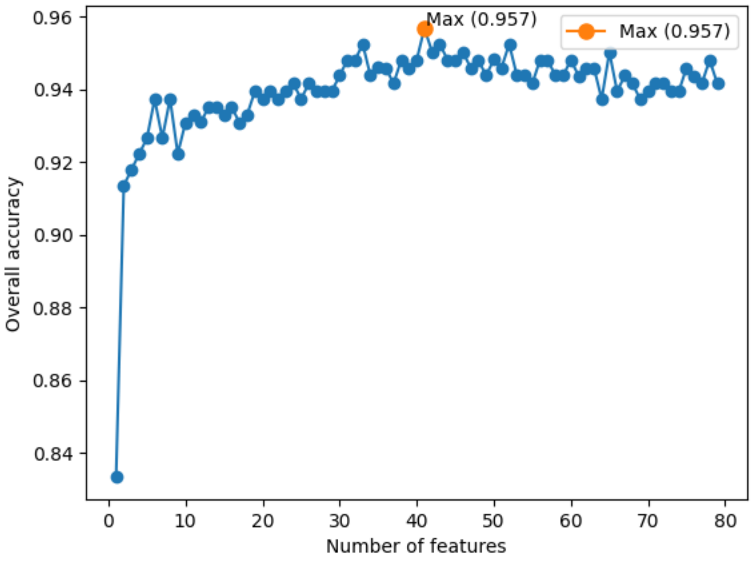

**Figure 3** Model's accuracy line from changing the number of features. The orange point depicts the maximum accuracy achieved.

In comparison with the Zhang[2] based on the accuracy line chart from Figure 3, we can see that our model outperforms the SVM model from Zhang's paper. The highest accuracy in Zhang's paper is 0.74 at the number of features equal to around 1600. Our highest accuracy is 0.96 at around 40 features. Besides that, our accuracy is also stable from 0.94 to 0.96 with the number of features from 40 to 80.

### 4.2 Feature importance

As shown in the Table 1, we can see that there were six similarities like genes RPSA(ENSG00000168028),RPL9(ENSG00000163682),RPS20(ENSG00000008988), RPL6(ENSG00000089009),RPLP2(ENSG00000177600), FAU(ENSG00000149806) in the 18 most important genes from RF model and 18 top genes from the Zhangs paper[2] that were essential for the classification of cancer and healthy sample(see Appendix B for Zhang's results). Among these 6 genes, most of them are ribosome-associated genes. An example is "RPS20", which is a ribosome-associated gene that contributes to ribosome biogenesis and it has been identified in cancer samples, including colorectal cancer and glioblastoma[2]. Also, other than genes that are related with ribosomes, genes like "FAU" has been found which has been demonstrated to contribute to the early stages of breast cancer, indicating that it may be a functional biomarker for the identification and differential diagnosis of breast cancer. On the other hand, many of the important confirmed tumor-associated genes like "TTN" and other significant genes like some cell surface protein genes as discussed in Zhangs paper[2] were not found in feature importance analysis of RF model.

**Table 1** The 18 most important features identified by RF model

| Order | Feature name | Gene name | Description |
|---|---|---|---|
| 1 | ENSG00000168028 | RPSA | ribosomal protein SA |
| 2 | ENSG00000163682 | RPL9 | ribosomal protein L9 |
| 3 | ENSG00000122406 | RPL5 | ribosomal protein L5 |
| 4 | ENSG00000100316 | RPL3 | ribosomal protein L3 |
| 5 | ENSG00000142541 | RPL13A | ribosomal protein L13a |
| 6 | ENSG00000123349 | PFDN5 | prefoldin subunit 5 |
| 7 | ENSG00000145425 | RPS3A | ribosomal protein S3A |
| 8 | ENSG00000008988 | RPS20 | ribosomal protein S20 |
| 9 | ENSG00000137154 | RPS6 | ribosomal protein S6 |
| 10 | ENSG00000089009 | RPL6 | ribosomal protein L6 |
| 11 | ENSG00000005961 | ITGA2B | integrin subunit alpha 2b |
| 12 | ENSG00000177600 | RPLP2 | ribosomal protein lateral stalk subunit P2 |
| 13 | ENSG00000166501 | PRKCB | protein kinase C beta |
| 14 | ENSG00000167526 | RPL13 | ribosomal protein L13 |
| 15 | ENSG00000149806 | FAU | FAU ubiquitin like and ribosomal protein S30 |
| 16 | ENSG00000035403 | VCL | vinculin |
| 17 | ENSG00000197956 | S100A6 | S100 calcium binding protein A6 |
| 18 | ENSG00000147403 | RPL10 | ribosomal protein L10 |

In conclusion, comparison of feature importance between our RF model and Zhang's paper indicated some significant similarities, notably in the discovery of ribosome-associated genes, which play critical roles in cancer classification. Furthermore, the finding of such ribosome-associated genes implies a strong association between ribosomal function and cancer. It suggests that modifications in ribosome-related activities, including ribosome assembly, translation, and protein synthesis, might have important consequences for the genesis and progression of cancer. Additional research and validation studies on these genes might lead to better diagnoses of cancer, differential diagnosis, and potentially specialized treatment approaches.

## 5 Contributions

1 Pham Gia Cuong: Overall 50% report and did the preprocessing with Deseq2, ran Random forest.

2 Nepal Aakash: Overall 50% report and ran RF plus extracted the feature importance.

## References

[1] Myron G Best et al. “RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics”. In: *Cancer Cell* 28.5 (2015), pp. 666–676. DOI: 10.1016/j.ccell.2015.09.018.

[2] Yuhong Zhang et al. “Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets”. In: *Oncotarget* 8.50 (2017), pp. 87494–87511. DOI: 10.18632/oncotarget.20903.

## Appendix A: t-SNE plot

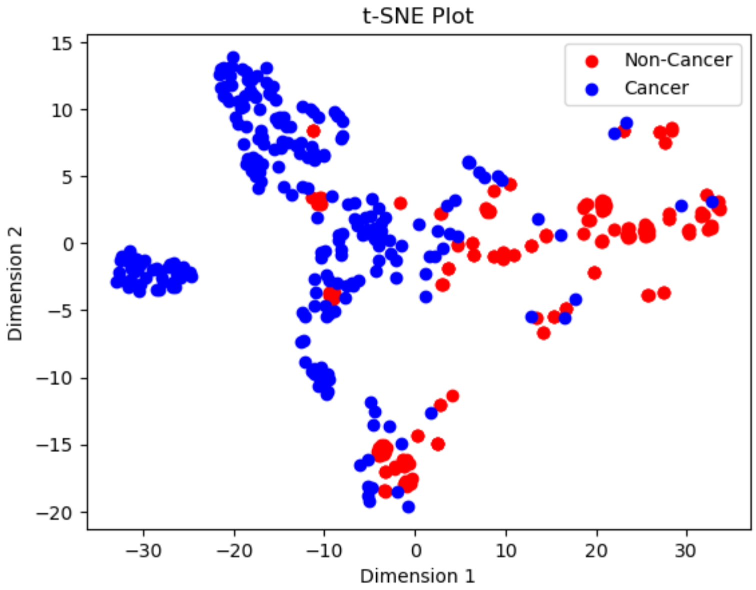

**Figure 4** t-SNE plot for the raw data set

## Appendix B: The top 18 features

**Table 2** The top 18 features in the MaxRel feature list as provided in Zhang paper[2]

| Order | Feature name | Gene name | Description |
|---|---|---|---|
| 1 | ENSG00000155657 | TTN | Titin |
| 2 | ENSG00000008988 | RPS20 | Ribosomal Protein S20 |
| 3 | ENSG00000177600 | RPLP2 | Ribosomal Protein Lateral Stalk Subunit P2 |
| 4 | ENSG00000211772 | TRBC2 | T Cell Receptor Beta Constant 2 |
| 5 | ENSG00000168028 | RPSA | Ribosomal Protein SA |
| 6 | ENSG00000142534 | RPS11 | Ribosomal Protein S11 |
| 7 | ENSG00000142676 | RPL11 | Ribosomal Protein L11 |
| 8 | ENSG00000105193 | RPS16 | Ribosomal Protein S16 |
| 9 | ENSG00000160654 | CD3G | CD3g Molecule |
| 10 | ENSG00000168421 | RHOH | Ras Homolog Family Member H |
| 11 | ENSG00000139193 | CD27 | CD27 Molecule |
| 12 | ENSG00000131469 | RPL27 | Ribosomal Protein L27 |
| 13 | ENSG00000163682 | RPL9 | Ribosomal Protein L9 |
| 14 | ENSG00000071082 | RPL31 | Ribosomal Protein L31 |
| 15 | ENSG00000149311 | ATM | ATM Serine/Threonine Kinase |
| 16 | ENSG00000149806 | FAU | FAU, Ubiquitin Like And Ribosomal Protein S30 Fusion |
| 17 | ENSG00000109475 | RPL34 | Ribosomal Protein L34 |
| 18 | ENSG00000089009 | RPL6 | Ribosomal Protein L6 |