**Molecular Characterization of Pancreatic Cancer Survival using Machine Learning**

***LINES FOR AUTHORS***

**Abstract ID: XXX**

**Abstract**

Despite therapeutic advances, pancreatic ductal adenocarcinoma (PDAC) remainsa very deadly cancer with an average 5-year survival rate around 10% and anincreasing incidence rate. PDAC is difficult to detect due to lack of symptoms during early stages inhibiting early diagnosis, hence lowering the chances to treat them surgically. CA19-9 is currently the most effective biomarker for pancreatic cancer recurrence but not as a screening tool. Yet, the analysis of cellular abnormalities and molecular expression profiles has flourished with the advances in biotechnology and could aid the identification of possible biomarkers. Therefore, this work aims to extract a genomic signature panel to assess survival risk of pancreatic cancer patients using machine learning approaches and two large-scale and molecularly comprehensive datasets, The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC). Since these are composed of high dimensional and heterogeneous data structures, feature selection methods (i.e., Recursive Feature Selection (RFE)) were implemented before constructing the predictive models (i.e., Support Vector, Random Forest). RFE revealed 100 and 10 genes for the TCGA and ICGC cohorts, respectively, both showcasing high performance (TCGA: ~74.5% accuracy, ~74.6% AUC; ICGC: ~82.2% accuracy, ~78.5% AUC) for predicting survival. Even with the discrepancies in size, these sets shared genes in the UCK, UNC, and WNT families. These reveal interesting biological insights on the possible link to survival with active WNT signaling from outer to inner cells and UCK as a target in anticancer therapies. Meanwhile, UNC genes are identified by experiments with worms to identify genetic defects.

**Keywords**

Pancreatic Cancer, Survival Rate, mRNA Gene Expression, Feature Selection, Biomarkers

1. **Introduction**

The human pancreas is in charge of regulating blood sugar; it uses exocrine cells to produce digestive juices and endocrine cells to produce hormones. These cells can sometimes grow out of control, which forms malignant tumors which develop into cancer. Pancreatic cancer is a deadly disease in which these cancerous cells form in the tissues of the pancreas. According to the ACS, PDAC is the most common subtype of cancer, associated with exocrine cells which make up most of the pancreas. What troubles pancreatic cancer patients’ survival is its lack of diagnostics in early stages (given three percent of incidence) and limited treatments, only ten percent are expected to live after 5 years. This body tissue is located in an uncomfortable area to identify tumors, and most patients do not show symptoms until after the third stage. During these advanced stages, the cancer is often spread throughout the rest of the body in problematic areas near blood vessels [2] which is why surgery is only recommended to less than twenty percent of patients. Even so, surgeries to remove lethal large-scale tumors do not necessarily help with survival (ninety percent of cases are fatal). The statistics for pancreatic cancer position it as the fourth leading cause of cancer-related deaths, accounting for seven percent of all cancer-related deaths [1].

Research to improve cancer diagnostics involve studies on gene expression and their advantages to identify biomarkers of pancreatic cancer. Gene expression is the process by which cells build proteins based on DNA instructions carried through RNA, it controls how and when proteins are made [3]. It helps analyze genetic mutations and study cellular abnormality associated with tumor cells, characterization of gene expression has been used in other types of cancer research. As for pancreatic cancer, where the only approved biomarker is CA 19-9 [4], the search for other biomarkers is important to assess patients’ survival rates. To find possible biomarkers across large cohorts of genes, statistical data analytics can build classification models with the optimal genes associated with pancreatic cancer survival rate. Hence, the objective of this work is to implement machine learning approaches to identify possible biomarkers to assess survival risk of pancreatic cancer patients. This will be achieved by selecting important features using machine learning methods and conducting a comparison across TCGA and ICGC, to evaluate if similar findings provide information about putative validated biomarkers.

1. **Methodology**

To extract the important genes to assess survival risk, this work implements a feature selection step using the recursive feature elimination (RFE) wrapper method. Once the features are selected these are evaluated across three classification methods (Logistic Regression (LR), Support Vector Classifier (SVC), and Random Forest Classifier (RFC)). The methodology is implemented in two large-cohort of pancreatic cancer patients in the aims to finds similarities among the results.

* 1. **Data Description**

To categorize a gene as a possible expression-based biomarker, information from two large cancer patient cohorts were incorporated. TCGA is a historical portal whose research has helped characterize around 33 types of cancer including PDAC. From this repository, survival status from 184 patient with their respective mRNA expression values for 20,531 gene probes were extracted and preprocessed. After pre-processing (e.g.,eliminate null values)a resulting dataset with 177 patient samples characterized across 20,025 gene probes were obtained. The ICGC is a voluntary scientific organization among the world's leading cancer and genomic researchers. ICGC provided samples of pancreatic cancer survival status and mRNA expression for 92 patients, and after preprocessing, resulted in a 90 patients by 35,601 gene probes dataset. A large number of predictors often lead to complex models with unreliable predictions, which is why TCGA and ICGC datasets needed to be reduced using computational methods. Their high dimensionality required feature selection to obtain the most relevant predictors to distinguish between living and deceased status from gene expression. Comparison between the features obtained for the TCGA and ICGC datasets would also be executed in both Python and R programming languages to evaluate the fit with classification models.

* 1. **Feature Selection: RFE**

Feature selection is a method in computer programming that takes a set of data and reduces the number of variables. Inside feature selection, the genes can be reduced based on a scoring method (also known as filters), training a subset of genes to a machine learning algorithm for evaluation (otherwise known as wrappers), and even a combination of both (called embedded methods). Each of these contains shared and unique qualities as they search the optimal features, where the wrapper methods proved to incorporate dependency between variables while adding genes to an initial subset and evaluating the fit. The wrapper technique used for this study was the RFE, which starts with all features in the training dataset and removes features by their rank of importance depending on the algorithms or statistical tests [5]. Although using this method can be advantageous to understand the biological interdependencies between genes, wrappers are often characterized by extensive computational times. In the case of RFE feature selection in Python language with ‘sklearn’ library, its effects on number of predictors for analysis limited its selection parameter for 3, 10 and 100 predictors for TCGA and ICGC datasets. On the other hand, R programming’s version of RFE evaluated up to 200 and 300 predictors with step 5 and chose the best subset for evaluation through ‘caret’ and ‘randomForest’ libraries.

An initial learner evaluation using cross-validation was applied in Python and R using 3 repeats and 5 folds to reduce overfitting, a recommended step when working with machine learning methods as well as the classifiers within the wrapper methods. Python computational run times were collected using a Mac 16GB RAM device with a 2.3 GHz processor while R codification was run in a Mac 8GB RAM device with a 1.1 GHz processor. This included the addition of data storage necessary to moderate prolonged run time for RFE and validate features between datasets while building classification models. To store the results from feature selection, the built-in functions would save the reduced matrix in a CSV file and information regarding run time and cross-validation scores for every computational run for further evaluation.

* 1. **Classification Methods**

Once the feature selection step was completed, the predictive performance of the reduced set of important genes obtained from TCGA and ICGC were evaluated using cross validation of 10 folds in three classification models: Logistic Regression (LR), Support Vector Classifier (SVC), and Random Forest Classifier (RFC). LR predicts the probability of a target variable where the dependent variable is a binary response [6]. SVC adapts to the data in a hyperplane that ideally categorizes it. Lastly, RFC is a model consisting of a large number of decision trees to obtain predictive performance [9]. Classification performance involved four metrics: accuracy represents the ratio of correct predictions for the test data, precision measures the number of predictors (true positives) the model correctly predicted divided by the total number of predictors, recall analyzes how many positive predictions were selected (true positive rate) and ROC AUC measures the probability of a model selecting a positive example over a negative [7]. Likewise, ROC AUC curves graphed the true positive rate versus the false positive rate, which assists in visualizing how much the models distinguish between a living and deceased patient [7]. In addition to the respective performance for each subset, predictors extracted from R were validated by evaluating TCGA’s optimal subset in ICGC’s complete dataset and vice versa to analyze the selected features’ relevance in another set of data using the same performance metrics.

* 1. **Biological Annotation and Enrichment Analysis**

Gene ontology and enrichment analysis was performed using the R packages, *clusterProfiler,* and *DOSE* with the KEGG pathway and Gene Ontology databases to evaluate the biological mechanisms possible involved with the significant gene sets to understand survival. To verify if a selected group of genes is represented more than the random expectation, statistical tests are performed (i.e., hypergeometric tests, contingency tests) using the previously established R libraries.

1. **Results**

Two different implementations of the RFE were executed from the sklearn and caret libraries from Python and R programming platforms.

* 1. **Python sklearn implementation**

Tables 1 and 2 shows a summary of the predictive performance metrics from Python’s RFE function across different sets of important genes and classifiers for TCGA, and ICGC, respectively. The optimal prediction results are highlighted in the tables to display which number of features best predict survival status along with its computational time.

**Table 1.** Classification Model Metrics in TCGA/ICGC in Python’s RFE using Random Forest Classifier

Table

Description automatically generated with low confidence

Even though the ICGC was a smaller dataset in terms of the number of samples, its computational times were slightly higher than those from TCGA, TCGA took an average of 22.4 hours while ICGC 24.9 hours. Also, increasing the number of features had little to no effect on the run time, as shown in **Table 1**, the SVC for TCGA offered higher accuracy, precision, and ROC AUC score when using 100 genes than 10 genes except for recall which was higher using 10 features. Similarly, RFC presented optimal metrics when using 10 features, adding more features did not increase results (see **Table 1)**. Fitting the features to a Logistic Regression model did not present optimal metrics for neither of the datasets.

The optimal features obtained by the RFE sklearn; 10 and 100 optimal features from ICGC and TCGA, respectively; were compared in terms of their intersection or commonality. There was no overlap when gene names were assessed. However, when studying the families of genes, UCK, UNC, and WNT appeared on both datasets as relevant variables to predict survival status. WNT proteins’ role is to send signals from outer cells to inner cells; variations of UCK have been found in cancerous cells and considered a target in anti-cancer therapies, and UNC families are used in experiments with worms to identify genetic defects. Furthermore, predictions achieved with Python’s RFE function held an apparent bias towards the features at the end of the list for both datasets, which were initially ordered alphabetically. Therefore, we proceeded to investigate if this behavior was similar or not using another programming language such as R.

* 1. **R caret implementation**

For functions executed in R’s RFE package, **Table 2** presents predictive performance metrics in each classifier after finding the optimal features in TCGA datasets and their common features when evaluated with ICGC samples as well as performance metrics from ICGC along with its corresponding validation on TCGA samples. Highlighted sections in both tables indicate which classifier model offers the best metrics in terms of the subset selected by RFE. Computational run time for RFE functions in R did not exceed 15 minutes for all executions, substantially lower than RFE in Python.

**Table 2.** Classification Model Metrics in TCGA/ICGC and Validation in R’s RFE using Random Forest Classifier



**Table 2** proves 136 features extracted from 200 predictor size and 271 from 300 limit size for TCGA dataset. SVC model obtained the highest precision and ROC AUC score when 271 features were selected, yet recall was higher in RFC. Top accuracy was achieved both in SVC and RFC with 271 features also, more predictors increased classification metrics. Their respective validations with ICGC dataset are displayed below their separate optimal predictors, where 115 out of 136 predictors were modeled in ICGC patient samples and 226 out of 271 predictors, respectively.

ICGC on RFE considered 166 features when RFE function size went up to 200 predictors and later 216 when limited to 400 predictors, as described on Table 4. Out of the optimal 166 predictors, 117 were found in TCGA patient samples while 157 were discovered when 216 predictors were chosen. Similar to TCGA metrics, adding predictors presented higher overall metrics. LG displayed highest accuracy, precision and ROC AUC score with 216 features, RFC model provided higher recall. **Figure 1** illustrates these genes in heatmaps to highlight their expressions on deceased and living patients. Several features displayed distinction between their expression in deceased and living patients in **Figure 1a** for TCGA subset, but lower half of the features did not show expression in deceased patients. The heatmap on **Figure1b** presents more features with higher expressions in living patients than deceased patients in visual terms, variability in both plots are reflected in the metrics (see **Table 2)**.

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

(a) (b)

**Figure 1**. Heatmaps for optimal predictors: (a) 271 TCGA features plot and (b) 216 ICGC features plot

Each intersection between TCGA and ICGC subsets decreased the number of optimal features one dataset considered important for prediction, the absence of many predictors lowered the majority of classification model metrics. Exceptions were found on TCGA validation, where 115 features increased LG precision; 226 features also presented higher recall on SVC and RFC. Validation metrics significantly affected accuracy metrics when evaluated on an independent dataset, optimal genes found in TCGA and ICGC held no bias when modeled in each other’s dataset. The largest effect was observed when only 157 out of 216 features of ICGC were evaluated with TCGA, LG previously scored highest accuracy (86.7%) but the difference of 59 predictors resulted in 40.9% decrease.

The selected subsets from TCGA and ICGC found two mRNA expressions in common, which were MIER1 and LIG3. Studies on them revealed that MIER1 is described as a transcriptional regulator and Gene Ontology (GO) annotations associate it to chromatin binding and signal transducer activity [10]. LIG3 is a protein coding gene with annotations related to DNA ligase activity and excision repair [11].

[Enrichment Results]

|  |  |
| --- | --- |
| **Cohort Study** | **Annotation & Enrichment Results** |
| TCGA: 271 gene set | Chart, scatter chart, bubble chart  Description automatically generated A picture containing diagram  Description automatically generated |
| ICGC: 216 gene set | Chart, bubble chart  Description automatically generated Chart, scatter chart  Description automatically generated |

**Figure 2**. Annotation and enrichment results for optimal gene sets selected for TCGA and ICGC patients’cohorts

1. **Conclusions**

The ability of wrapper methods to incorporate dependency between variables can be relevant in studying potential biomarkers given computational time limitations since some genetic expressions require the presence of others to generate certain proteins [8]. The computer device used to generate the selected features took around a day (22.4 and 24.9 hours for TCGA and ICGC, respectively); however, the number of predictors did not affected the run time. Adding more features to the model does not necessarily increase the performance, proven when ICGC reached higher metrics with 10 variables rather than 100. In terms of accuracy, ICGC dataset scored 82.22% while TCGA scored 74.57%. Although the comparison did not find identical genes in common, UNC, UNC, and WNT families were present in both models and are known to be relevant in cancer-related studies such as therapies, genetic defects, and protein signals.

R did not present computational time limitations on the given number of predictors as Python. Different sets of features were obtained by feature selection in R when assigning from 200 to 400 predictors, its effect on classification models increased along with the selected features. TCGA and ICGC accuracy respectively improved with 271 and 216 predictors, scoring 74.60% and 86.7%. These sets of data shared MIER1 and LIG3 genes in common, which are respectively important in regulating the expression of number in genes and part of the DNA ligase family.

This opens the door to generate codes that calculate the optimal number of features by taking into account estimated running times in each programming language. Applying other variations of the RFE method such as different learners aside from Random Forest Classifier is needed to possibly improve the selection of markers. Pathway enrichment analysis should be the next step for common genes in both datasets, which would incorporate statistical approaches in search for association between these findings and biological pathways. Though the results by Python were highly predictive and encouraging, it also showed some possible bias on the selection of important features since most of them seem to be at the end of the gene list if ordered alphabetically which could be a problem in terms of biological interpretation of these putative biomarkers. Before conducting correlation analysis on the mRNA expressions, investigations on how Python libraries generate their results are needed, as well as exploring other implementations where the R programming is a possibility since no bias on list order was observed.

**References**

[1] American Cancer Society. Key Statistics for Pancreatic Cancer. (2021, January 21)., from <https://www.cancer.org/cancer/pancreatic-cancer/about.html>

[2] Marks, Julie. Inoperable Pancreatic Cancer. (2020, October 18)., from <https://www.healthline.com/health/inoperable-pancreatic-cancer>

[3] [J Cancer.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6036716/) 2018; 9(13): 2249–2265. Published online 2018 Jun 5. doi: [10.7150/jca.24744](https://dx.doi.org/10.7150%2Fjca.24744)

[4] Hasan, Syed, et al. "Advances in pancreatic cancer biomarkers." Oncology reviews 13.1 (2019).

[5] Brownlee, Jason. Recursive Feature Elimination (RFE) for Feature Selection in Python. (2020, May 25). from <https://machinelearningmastery.com/rfe-feature-selection-in-python/>

[6] Machine Learning – Logistic Regression. Tutorials Point., from https://www.tutorialspoint.com/machine\_learning\_with\_python/machine\_learning\_with\_python\_classification\_algorithms\_logistic\_regression.htm

[7] Classification: ROC Curve and AUC., from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[8] Nature Education. Gene Expression. (2014)., from <https://www.nature.com/scitable/topicpage/gene-expression-14121669/>

[9] Yiu, Tony. Understanding Random Forest. (2019, June 12)., from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

[10] The Human Gene Database. MIER1 Gene. (2013, May)., from <https://www.genecards.org/cgi-bin/carddisp.pl?gene=MIER1>

[11] The Human Gene Database. LIG3 Gene. (2008, July)., from https://www.genecards.org/cgi-bin/carddisp.pl?gene=LIG3