A Machine Learning Approach Towards Prediction of Pancreatic Cancer Using Gene Expression and DNA Methylation

Paige Keller[1], Samuel Morehead[1], Zachary Caterer[1], Nijhum Paul[2], Rahul Gomes[1], Rick J. Jansen[3]
[1] University of Wisconsin-Eau Claire, Computer Science, [2] North Dakota State University, Public Health
[3] Biostatistician, Center for Biobehavioral Research, Sanford Health, Fargo and Masonic Cancer Center, University of Minnesota, Minneapolis

Introduction

DNA methylation can affect gene accessibility and therefore gene expression. Including those that suppress or promote tumor growth and progression. In this research, we examined the potential of developing a scalable feature selection and deep learning framework capable of processing high dimensional genomic datasets to identify methylation and gene expression sites in the human genome contributing to pancreatic ductal adenocarcinoma (PDAC).

Process

Dataset

The entire process is listed in Figure 1. Methylation and RNA-seq files were obtained from TCGA-PAAD project. From a total of 178 donors, 195 methylation files were obtained that comprised of 11 normal and 184 tumor samples. The RNA-seq dataset had 183 files comprising of 4 normal and 179 tumor samples.

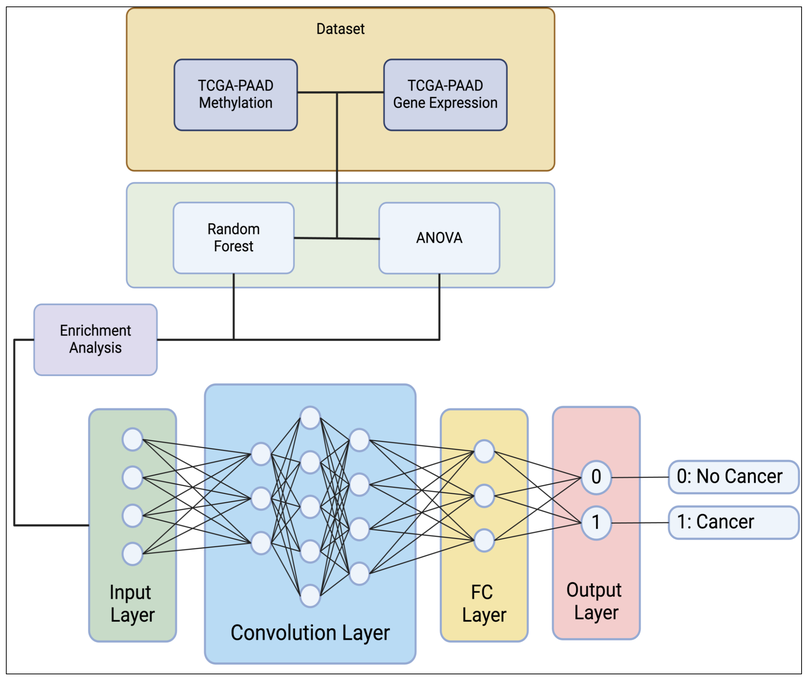

Figure 1: Diagram of methodology

Addressing Data Imbalance

- Due to data imbalance with the minority class being normal samples, direct application of feature selection was not possible due to model biased towards features that express tumor.
- Hence, normal samples were combined with a group of tumor samples using random undersampling where instances from the majority class (tumor) are subsampled until a more balanced class distribution is achieved.
- For the methylation dataset, this process was repeated 8 times with 23 tumor in each group.
- For the RNA-seq dataset, this process was repeated 7 times with 25 tumor in each group.
- The process is depicted in Figure 2.

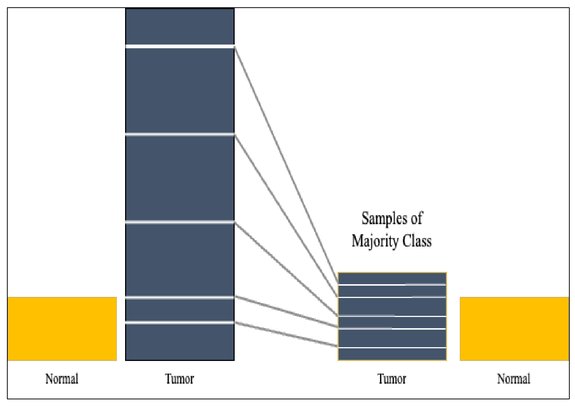

Figure 2: Representation of undersampling used in one cycle of feature selection

Feature Selection

- Feature selection performed separately for every sub-group in random undersampling process.
- ANOVA and Random Forest implemented for dimension reduction.
- For ANOVA, all features that reported to have a p-value < 0.05 were selected.
- Random Forest classifier utilized 500 decision trees to return the selected list.
- The features selected were combined to yield a final list of important features.
- CpG markers from Methylation and EnsembleIDs from RNA-seq were matched to gene names.
- Results from this gene overlap seen be in Figure 3.

Deep Learning

- In this stage, the CPG markers and ensemble gene ids that were identified as being important by both Random Forest and ANOVA were used to create a deep learning classification model for cancer prediction.
- The deep learning model consisted of 4 dense layers with 100, 200, 300, 200 neurons respectively.
- A default activation of Relu was used in all these layers.
- The learning rate was set to 0.001 with a patience of 20 epochs.
- The dropout rate was set to 0.1 with a loss function of sparse categorical cross entropy.
- For methylation data, the neural network was trained in eight incremental stages of 100 epochs each with a total of 800 epochs. 11 normal samples combined with 23 tumor samples at each stage.
- For RNA-seq data, the neural network was trained in seven incremental stages of 100 epochs each with a total of 700 epochs. Here 4 normals combined with 25 tumor samples were used.
- In this incremental approach, the deep learning model readjusted its weights based on a different combination of tumor and normal samples. The combined final training accuracy after both models on the entire dataset can be seen in Table 1.

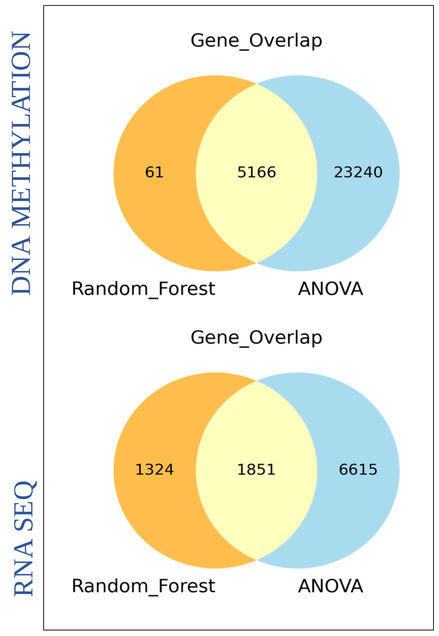

Figure 3: Results from feature selection

	Accuracy	Precision	Recall	F1	Cohen Kappa	ROC AUC
Methylation	0.995	1	0.994	0.997	0.954	0.997
RNA-seq	0.995	1	0.994	0.997	0.886	0.997

Table 1: Deep learning metrics for both methylation and RNA-seq data

Enrichment Analysis

- The 4039 genes obtained as an overlap from feature selection using Random Forest and ANOVA were used to perform an enrichment analysis. These genes included that of both methylation and RNA-seq feature selection
- When compared with a list of 221 tumor-related genes that were identified from literature, our results returned 98 common tumor genes.
- Enrichr database was used to identify common pathways across the overlap genes.
- Figure 4 shows the pathway analysis of the 98 cancer genes while Figure 5 shows the pathways analysis of all feature selected genes.
- These also included genes such as “KRAS", "TP53", and "SMAD4” revealing that the stratified feature selection technique is indeed useful at identifying important features.

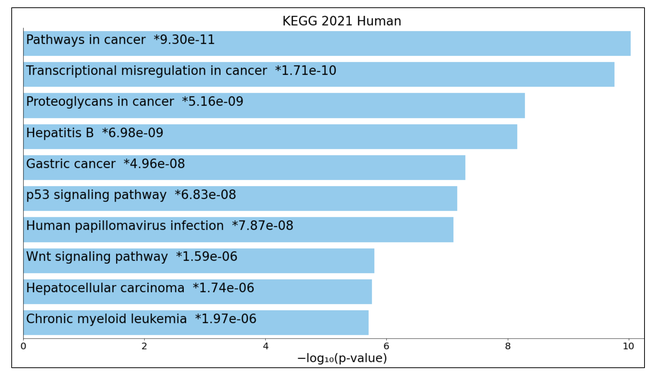

Figure 4: Pathway analysis of 98 cancer genes indeitified from literature and present in feature selection

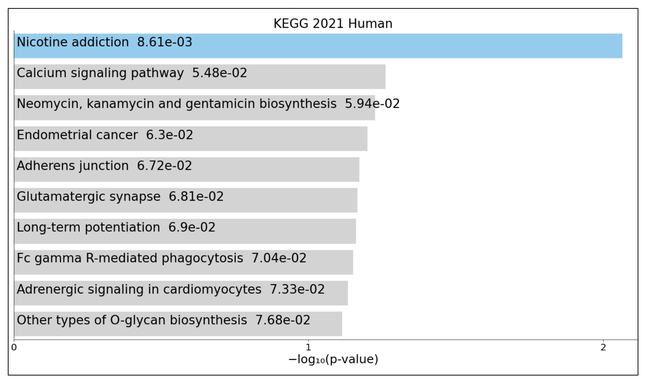

Figure 5: Pathway analysis of all feature selected genes identified from feature selection

Acknowledgements

We would like to acknowledge funding support from the National Institute of General Medical Sciences of the National Institutes of Health, under NDSU COBRE Award Number 1P20GM109024, the Office of Research and Sponsored Programs (ORSP) at UW-Eau Claire, and NIH grant P30 CA77598 utilizing the Biostatistics Core shared resource of the Masonic Cancer Center, University of Minnesota and by the National Center for Advancing Translational Sciences of the National Institutes of Health Award Number UL1TR002494. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The computational resources were provided by the Blugold Center for High Performance Computing funded by NSF CNS-1920220.