**Abstract:**

The goal of this project was to develop a cloud-based lung cancer classification machine learning model. To this end, 3 different lung cancer datasets were concatenated and combined along common genes. The features from this data set were analyzed using a Random Forest classifier to determine feature importance. The feature importances were evaluated using a standard z-score statistic test, using 2 as the cutoff value, the top ≈2% of features were selected (129 of 5233). Those features were fed into numerous machine learning models, with the best performing model being SVM with a linear kernel, with >90% accuracy.

**Introduction:**

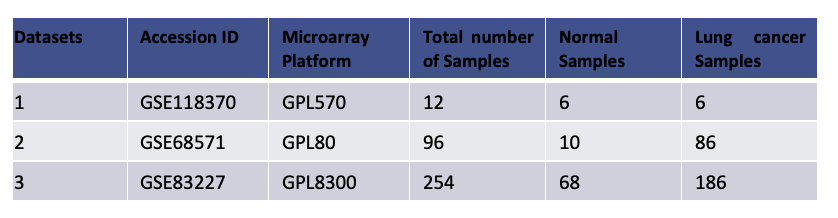
Lung cancer is one of the most widely diagnosed cancer type in developed countries such as the United States, and is divided into two major types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) [1,2]. Adenocarcinoma, squamous cell carcinoma, and large cell carcinoma are the three subtypes of NSCLC, accounting for 85% of lung cancer incidence. SCLC accounts for 10 – 15% of the total lung cancer [3] incidence but it is one of the most aggressive forms of cancer due to rapid development and progression. When choosing a target for any drug, one needs to identify either genes which are mutated in cancer cells or genes whose expression is only essential to cancer cells. With a significant increase in the number of cancer causative factors, both intrinsic and extrinsic, there exists a gap in knowledge in terms of the genetic predisposition for each tissue or type of cancer. This gives room for exploring, defining complex networks of regulation and gene function, and dissecting it to design better drugs, tailor-made to each type of cancer.

Problem statement

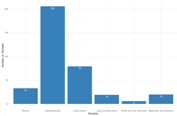
Every year, each case of lung cancer diagnosis leads to an extensive amount of both imaging and gene expression data being generated. There emerges a need to develop platforms or tools that can effectively process such data allowing doctors and researchers alike to make decisions at quicker pace. In recent years, some research groups are working on employing cloud-based machine-learning algorithms to differentiate between tissue types for being cancer or normal. For instance among others, two groups from India and Saudi Arabia have implemented an SVM model on image data from Computed Tomography (CT) scans to distinguish between normal and cancerous class of tissues [4,5]. While most studies are focused on a binary sort of classification and lung cancer in itself is quite complicated, much is required to be done. Cancer genes are inherent to every cell and are harmless until their activation by mutagens/carcinogens. As the number carcinogens is increasing over the years, analysis of imaging data alone might not suffice. Designing an algorithm that can use expression data to effectively tell if a tissue is cancerous and if so also define its cancer type, can be extremely helpful. Using the machine learning component of spark, we propose to build a multi-class classification model that can, based on expression levels of genes, differentiate between cancerous tissue and types and provide a consensus that can help frame formulations of more specific drugs and targets. In a clinical setting, this could be broadly applied to help pair specific cancer types with targeted therapies, in contortion with traditional approaches: chemotherapy, radiotherapy or more generalized approaches.

Data Description

For this study, transcriptomic data captured using microarray platforms were downloaded from the Gene Expression Omnibus (GEO). The datasets were chosen from experiments performed on Lung cancer types and also normal samples. Below is the table that summarizes the number of samples in each dataset. The accession ID can be used to download the data from the NCBI’s GEO dataset.



Dataset 1 (GSE118370) was obtained from a study conducted on patients with lung adenocarcinoma to predict the survival rate using the gene expression profiles [6]. The second dataset (GSE68571) was from a study on lung adenocarcinoma patients paired with normal samples to find the new molecular regulators that can help improve the therapeutic targets of lung adenocarcinoma [7]. The third dataset (GSE83227) was obtained from a classification study which stated integration of the gene expression data from multiple samples with the clinical data could help in quicker diagnosis of lung cancer [8]. The dataset also had sub-classes of lung cancer i.e. Lung adenocarcinoma, Small-cell lung cancer, Squamous cell carcinoma, lung carcinoid tumor along with the normal samples. The number of samples in each sub-class is represented using a bar-plot shown below:

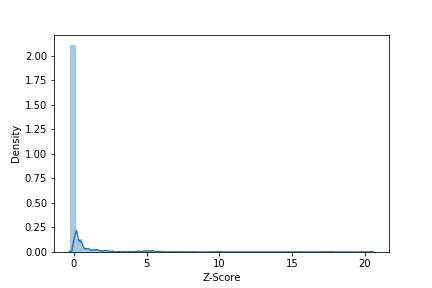


Data Pre-processing

The transcriptomic data from the GEO was downloaded in the .CEL format. The raw data (.CEL) file was directly downloaded into the R-Studio using the GEOQuery package [9]. The downloaded raw data was curated using packages such as affy, affyQCReport, oligo and converted into a comma separated file (.csv). The CSV file contains rows that lists the genes and the columns represent the samples. After conversion of the file format, the probes will be mapped to genes and list of genes will be used as features for the classification models. Normalization was also performed for each dataset using the R-package ‘rma’. In this Robust- Multi-Array Average method of normalization, the dataset will be background corrected for any local artifacts and noise, normalized and summarized to get the final expression matrix [10]. This method of is one of the effective normalization methods for microarray data. The pre-processed datasets were merged together to a single large expression matrix containing 362 samples and 5233 genes.

Feature Selection

Feature selection was performed using a Random Forest classifier with all available features. Each feature’s significance in the classification of the data was calculated. A z-score analysis was performed on the the feature significance. The z-score distribution is shown below. All features with z-scores greater than 2 were selected and composed the finalized set of features.



Model Creation and Cross-validation

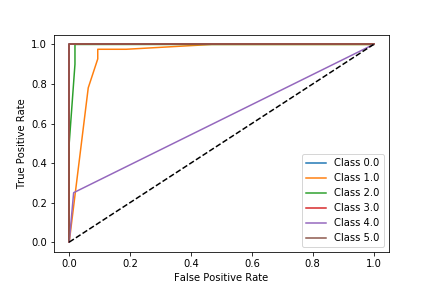
To classify the different types of lung cancer along with the normal samples, four different machine learning algorithms were implemented. To determine the distribution of datapoints in each clusters K-means clustering was performed and to check accuracy of classification of different lung cancer types SVM, Random forest classifier, Decision Tree Classifier was implemented. Cross-validation was also performed by splitting the datasets randomly into 80:20 ratio. The training set had 80% of the features while the test set had 20% of the data. Receiver Operating Characteristic (ROC) curve was plotted for the best performing model in terms of accuracy.

**Results:**

Due to the uncertain nature of Machine Learning, we decided to test a variety of different machine learning algorithms including Decision Tree, K-means, Random Forest Classifier, and Support Vector Machines Classifier (SVM). These models were selected because they are supported by pyspark.ml library, support multiple label classification, and have relatively high sensitivity to limited number of samples for a specific label. Models were evaluated by comparing accuracy values from a k = 5 fold cross-validation. Additional scrutiny was applied to the best performing model, through analyzing the Receiver Operating Characteristic (ROC) curve, to determine strengths and weaknesses of the model.

In 5-fold cross-validation, the data is shuffled and subset into 5 equally sized folds. 4 of those of folds were used for training data, and the remaining fold was used for testing the model’s predictions. The results for the different models performed relatively equally, between 85 – 90 percent, with only the K-mean clustering being a significant outlier with 53.0 percent accuracy. The Support Vector Machine Classifier performed marginally better at 91.7 percent accuracy relative to Decision Tree with 84.9 percent and Random Forest with 89.0 percent. Therefore, SVM was chosen as the best performing model for additional analysis.

An ROC curve was generated for the SVM model using a pseudo random fold of the data, that contained at least one entry for each class of the model. The ROC curve is shown in Figure XXX. The ROC was analyzed by calculating the area under the curve(AUC) for each class, This is a measure of the accuracy of the model at predicting subjects of a particular class, with the closer the number is to 1, the more accurate the predictions are likely to be. The result of the AUC showed some evidence of overfitting for numerous classes; 0, 2, 3, and 5, with numbers exceeding 0.99. Additionally, the AUC showed very poor accuracy for class 4 with 0.618. This lack of accuracy for class 4, is likely as a result of the lack of class 4 samples present in the data.



**Conclusion:**

This project serves as a proof of concept for a cloud-based high precision lung cancer type classifier. With additional data entries, the precision and accuracy of this tool could likely be expanded; able to determine different stages of developing lung cancer and identify specific strains and subtypes of lung cancer. By utilizing the virtualization of cloud platforms, this tool could be easily utilized by medical staff to quickly diagnose patients and begin to determine treatment.

References:

1. Lungcancer.org:Diagnosing Lung Cancer. https://www.lungcancer.org/find\_information/publications/163-lung\_cancer\_101/267-diagnosing\_lung\_cancer
2. Lungcancer.org: Types and Staging of Lung Cancer. lungcancer.org/find\_information/publications/163-lung\_cancer\_101/268-types\_and\_staging
3. Cancer Research UK: Lung cancer types. https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types
4. Khan S.A, Nazir M, Khan M, Saba T, Javed K, Rehman A, Akram T, Awais M. Lungs nodule detection framework from computed tomography images using support vector machine. MRT 2019 April;82(8):1256-1266.
5. Valluru D, Jeya IJS. IoT with cloud based lung cancer diagnosis model using optimal support vector machine. Health Care Manag Sci. 2019 Jul;doi.org/10.1007/s10729-019-09489-x
6. Xu L, Lu C, Huang Y, Zhou J, Wang X, Liu C, Chen J, Le H. SPINK1 promotes cell growth and metastasis of lung adenocarcinoma and acts as a novel prognostic biomarker. BMB Rep. 2018 Dec;51(12):648-653.
7. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med. 2002 Aug;8(8):816-24.3.
8. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C,Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W,Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001 Nov 20;98(24):13790-5. Epub 2001 Nov 13.4.
9. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007 Jul 15;23(14):1846-7. Epub 2007 May 12. 5.
10. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U,Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr;4(2):249-64.

Appendix: Group Member Contributions:

Alexander Krohannon: Xander found the 3 datasets, performed the feature selection anaylsis, created the Random Forest model, and did the ROC curve analysis.

Gayathri Panangipalli:

Sharmila Selvaraj: