The effects of gene expression on pulmonary adenocarcinoma prognosis

1 ABSTRACT

This project focuses on investigating the effects of gene expression on the prognosis, or the likely course, of pulmonary adenocarcinoma, which is one of the most common types of lung cancer. Due to the large variety in lung cancer types, as well as the wide range in symptoms and patient conditions, it is often very difficult to reach an accurate prognosis. By using computational methods such as artificial neural networks, decision trees and Bayesian networks, on gene expression data from the patient, it may be possible to reach more accurate and objective prognosis. The gene expression data for this project was obtained from The Cancer Genome Atlas (TCGA) Lung Adenocarcinoma Project (LUAD) and after pre-processing, run through classifiers on WEKA (Waikaito Environment for Knowledge Analysis). The dataset after preprocessing consisted of the gene expression data of 188 different genes on 176 patients diagnosed with pulmonary adenocarcinoma. An average accuracy of 74% accuracy was reached on the classifiers, and 17 genes were identified to have some prognostic value. Some of these genes include ERBB4, TAB2, and MAGI1, which regulate cascades that play important roles in cell proliferation, differentiation, and apoptosis, as well as NCKAP5, and EML4, which play a role in the production and organization of microtubules, which are an essential structure for mitotic cell division. Such results indicate the high potential of classifiers in the field of cancer research in not only deducing prognosis, but in also in discovering potential genes that may have a significant effect on the patient’s symptoms and survival.

2 INTRODUCTION

2.1 MOTIVATIONS

Lung cancer is the most common type of cancer worldwide, causing over 140,000 deaths in the USA in 2019 alone with an average 5-year survival rate of 25% for all SEER (Surveillance, Epidemiology, and End Results) stages (American Cancer Society). Pulmonary adenocarcinoma is a type of non-small cell lung cancer, and around 40% of all total lung cancer diagnoses are of pulmonary AD. This type of lung cancer most often occurs in smokers but is also the most common type of lung cancer diagnosed in non-smokers as well. The prognosis of cancers is extremely difficult due to their heterogeneity and the wide array of mutations in the genome which could result in the growth of cancerous cells. Targeted gene therapy, as well as gene-data dependent prognosis is an upcoming field in cancer research, as individual gene data of patients allow for doctors to make a more accurate prognosis and come up with treatments personalized towards individual patients. However, due to the vast amount of data which can be obtained from gene expression, some of which may not directly impact prognosis, it is important to develop efficient and accurate computational methods which can use this data in a productive way.

2.2 Vocabulary and Data Sources

FPKM-UQ count: The Fragments per Kilobase of transcript per Million mapped reads upper quartile is a normalized value of HTseq counts, which is the number of mapped reads to each gene. (National Cancer Institute) The normalization formula is as follows:

$FPKM = \frac{[RM_g * 10^9]}{[RM_{75} * L]}$ RM_g = number or reads mapped to the gene, RM_{75} = number of reads mapped to the 75th percentile gene in the alignment, L = length of the gene in base pairs.

Driver Gene: Genes whose mutations increase net cell growth under specific microenvironmental conditions.

Multilayer Perceptron: Classification algorithm in which an activation function is used and each perceptron outputs a Boolean value based on the weight and label of the edge and node. The WEKA multilayer perceptron model uses the sigmoid function on all nominal inputs. While this method is widely used in similar classification problems, their structure makes them very time-inefficient, and it is usually impossible to know the source of the hidden layers making them very difficult to analyze.

Bayesian Network: The Bayesian Network algorithm outputs a directed acyclic graph based on the probability distributions of input data. They can be used to deduce the dependencies of input variables and attributes.

J48 Decision Trees: Generates a rooted, directed tree in which the leaves correspond to outcomes and the nodes correspond to input variables. They are generally faster to construct than the two other data structures and can allow for reasoning based on the labeling of inner nodes.

These three classifiers were used, since they were the most commonly used methods in literature based on similar research of influences of gene expression on cancer prognosis. Another aim of this project was to investigate which of these three classifiers were most suited for the purpose of cancer prognosis based on their output accuracy rates.

2.3 Materials and Methods

The first step of the process after downloading data samples from the TCGA LUAD database was to construct a summarized database of patient data based on the extent of available clinical information, including survival times. Of the original 595 patients, specific survival times were only available for 179 patients, since many of the patients either did not have clinical data available on the database or had a null value survival rate. A null value survival rate indicates that either patient data was missing from the database or that the patients were currently alive; patients with null survival rates were eliminated from the analysis dataset in order to remove potential for bias. The survival rate of patients in the dataset used for analysis is as below in Figure 1. This data is not a very accurate representation of the actual prognosis of lung AD patients, however, as average 5 year survival rate is 23% in reality, while the 5-year survival rate for the patients in this dataset is 5%.

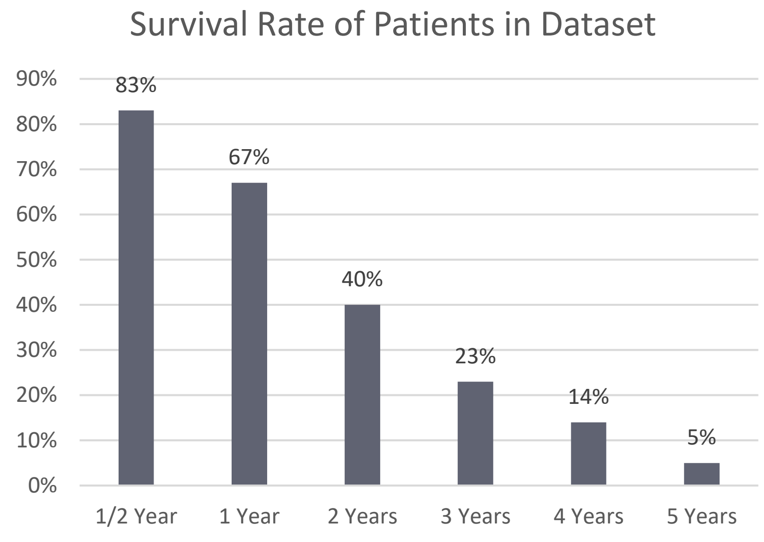

Figure 1: Survival rate of patients in dataset by year

For each patient, the expression of 60,483 genes were available in the original dataset, but because the number of attributes were too large and could not be processed in a reasonable time by the WEKA classifiers, the number of gene attributes were narrowed down based on potential driver genes found on previous research done on the topic. Some of the more significant of these driver genes are ERBB/EGFR, EGFR, MET, and LKB1. ERBB (Receptor tyrosine-protein kinase erbB), which is a parent type of EGFR is currently a key paradigm of molecular targeted therapy for lung cancer, and many reports suggest that ERBB mutations may carry significant prognostic value. It plays a role in the regulation of the ERK1 and ERK 2 cascade, affecting cellular response to stimulus, as well as negatively regulating apoptotic processes. Cell population proliferation, and cell fate commitment are also affected by mutations in this gene as well. KRAS is a central mediator of downstream growth factor receptor signaling and is critical for cell proliferation, survival, and differentiation having a relatively high rate of occurrence in adenocarcinoma patients. MET is also a type of receptor tyrosine kinase and facilitates receptor phosphorylation and activation. It is also thought to be a mediator of effects like tumor grown, survival, branching morphogenesis, migration, invasion, and metastasis; overexpression of MET often correlates with decreased survival of patients. LKB1 is a well-known tumor suppression gene which is commonly seen with inactivating mutations in non-small cell lung cancer tumors. Like ERBB, it plays an important role in cell metabolism, apoptosis, and DNA damage response. In addition to these well-known driver genes of lung adenocarcinoma, additional driver genes found on Candidate Cancer Genome Database.

The numeric FPKM-UQ values of the 188 potential driver genes were converted into nominal Boolean values such that FPKM-UQ values larger than 0 were interpreted as positive expression, and those equal to 0 were interpreted as negative expression. Then based on the days of survival of the patient, each subject was assigned 6 Boolean values; if the patient survived for 4 years, then the values for ½, 1, 2, 3, and 4 years would be true while the value or 5 years would be false. The gene expression data and the survival information were put into an ARFF attribute relation file format which could then be directly inputted into WEKA to be run on the classifiers using 10-fold cross validation. A separate model was created 6 separate times, such that the first model will provide a true or false answer to the question of whether a patient would survive 6 months, the second model to 1 year survival, and so on for up to 5 years. The results obtained from this method proved to be the most accurate and the most applicable.

In addition to the method of using Boolean nominal values above, three other ways of data representation were tested. Initially, before narrowing down the data based on potential driver genes, the J48 decision tree classifier was run on the entire gene attribute list of over 60,000 genes, all of which were converted into Boolean nominal values. Such a dataset could not be run on the remaining two classifiers, as the dataset was too large to be processed in a reasonable amount of time. However, most likely due to the problem of dimensionality from the unreasonably large number of attributes, the model was only able to correctly classify 52% of instances, which is basically equivalent to a random coin toss. Secondly, the multilayer perceptron classifier was run on Boolean dataset with 188 gene expressions represented as Booleans and trained to return a specific number of years survived from the given nominal values (6 months, 1 year, 2 years, 3 years, 4 years, 5 years). This model was only able to return the correct prognosis with 38% accuracy. Another type of data representation that was tested was using numeric FPKM-UQ gene expression values instead of converting them to Boolean values. This data was run on the multilayer perceptron model, as the Bayesian Network and J48 Decision tree only took nominal values inputs. The resulting model only had a correlation coefficient of -0.1204, indicating little to no correlation, indicating that providing numeric values to these models reduced the accuracy of the models.

3 Results

Table 1: Accuracy by classifiers and by year

Years	Bayesian Network	J48 Tree	Multilayer Perceptron	Average
1/2 Year	83%	82%	73%	80%
1 Year	65%	64%	56%	62%
2 Years	56%	57%	55%	56%
3 Years	77%	76%	63%	72%
4 Years	86%	86%	76%	83%
5 Years	95%	95%	91%	94%
Average	77%	77%	69%	74%

Across the three classifiers tested, an average accuracy of 74% was reached for all years of survival. Both the Bayesian Network and J48 Decision Trees had very similar accuracies overall, with the highest accuracy prognosis of survival for 5 years, 4 years, and 6 months. This was also true for the accuracy of the Multilayer Perceptron, though its accuracy was lower than that of the other two models overall. The disparity in accuracy both in relation to the years of survival and the different classifiers may be attributed to the method of output, as well as the nature of the initial dataset. As can be seen from Figure 1 above, the training dataset contained patients whose survival rates were distributed very unevenly, mirroring the overall survival rates of lung AD patients in general. However, this meant that even without the gene expression data, it would have been possible for the Bayesian Network and J48 classifiers to develop only one node that would return only true or only false values and achieve an extremely high accuracy during cross-validation. For instance, only 5 out of the 179 patients survived for 5 years, meaning that if the model consistently returned false for the question of whether a given patient had survived for 5 years regardless of the information given by the other attributes, it would be correct for 97% of the time. This was exactly what occurred in the trees produced for these models, in which there was only one leaf that would consistently return a single Boolean value. The same can be said for 4 year survival and 6 month survival as well, since most patients survived for 6 months but did not make it beyond 4 years. In comparison, survival rates of patients for the 1 year, 2 years, and 3 years after diagnosis was relatively evenly distributed, meaning that the models would have had to develop multiple connections and take gene expression data into account to reach a better prognosis. Nonetheless, the accuracy of these models decreased drastically, most likely due to the lack of training data, as well as small correlation between provided attributes and prognosis. This may also provide a reason for the lower accuracy of the Multilayer Perceptron model, as this model is forced to consider the gene expression data and cannot return simple answers like the other two models. While the accuracy for this model may be lower than that of the other two, it may be providing a more accurate picture of the correlation between gene expression and prognosis.

From the models obtained from the three different classifiers, the genes that were most frequently found to be on nodes or to have significant weights in the models were counted, as can be seen from the top 14 genes seen in Table 2 below, along with the ontology or function of these genes. Figure 2 through 4 are some sample trees created by the J48 Decision Tree. Many of the genes in the table can also be seen in these trees, indicating that they may have a significant effect on prognosis.

Table 2: Top 14 most frequent genes found in classifiers and their Gene and Ensembl ID

Gene ID	Ensembl ID	Frequency
CNTNAP5	ENSG00000155052	686
PET/FEV	ENSG00000163497	686
MAGI1	ENSG00000240175	684
TAB2	ENSG00000225924	684
LRP1B	ENSG00000168702	684
EGFR-AS1	ENSG00000224057	684
NCKAP5	ENSG00000233729	682
EHD4	ENSG00000259883	682
STARD13	ENSG00000230300	682
WDR7	ENSG00000267225	680
ANK3	ENSG00000254271	680
SLC38A3	ENSG00000188338	680
ERBB4	ENSG00000178568	680
EML4	ENSG00000234217	678

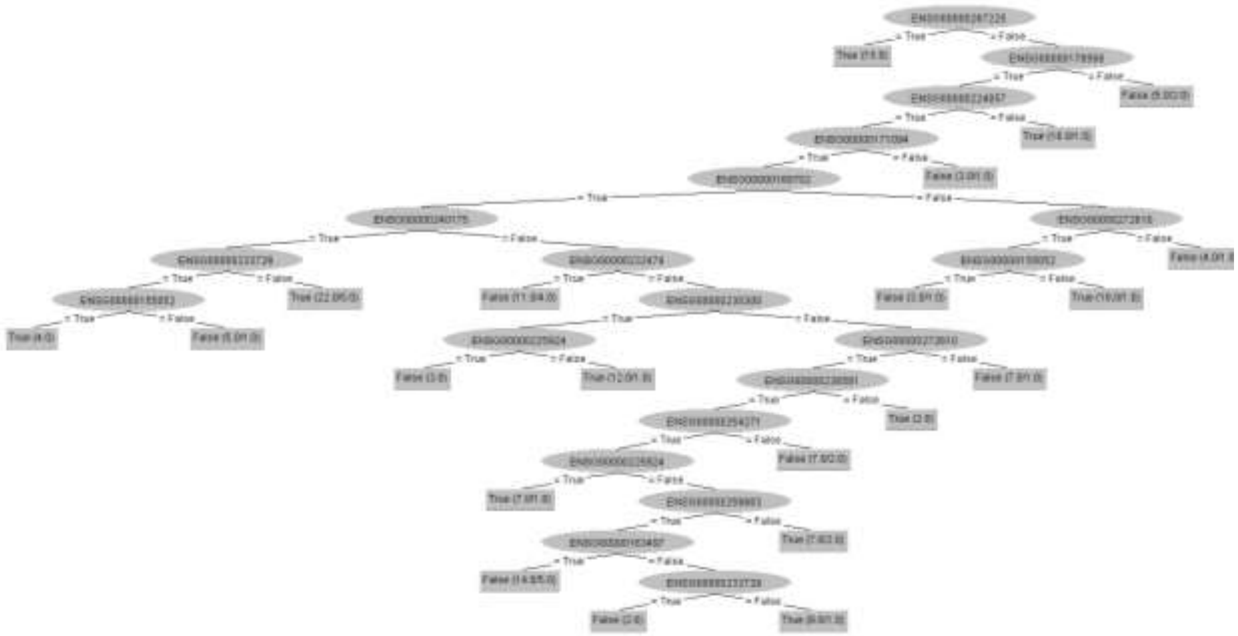

Figure 3: Tree produced by J48 Decision Tree for 1 year survival rates.

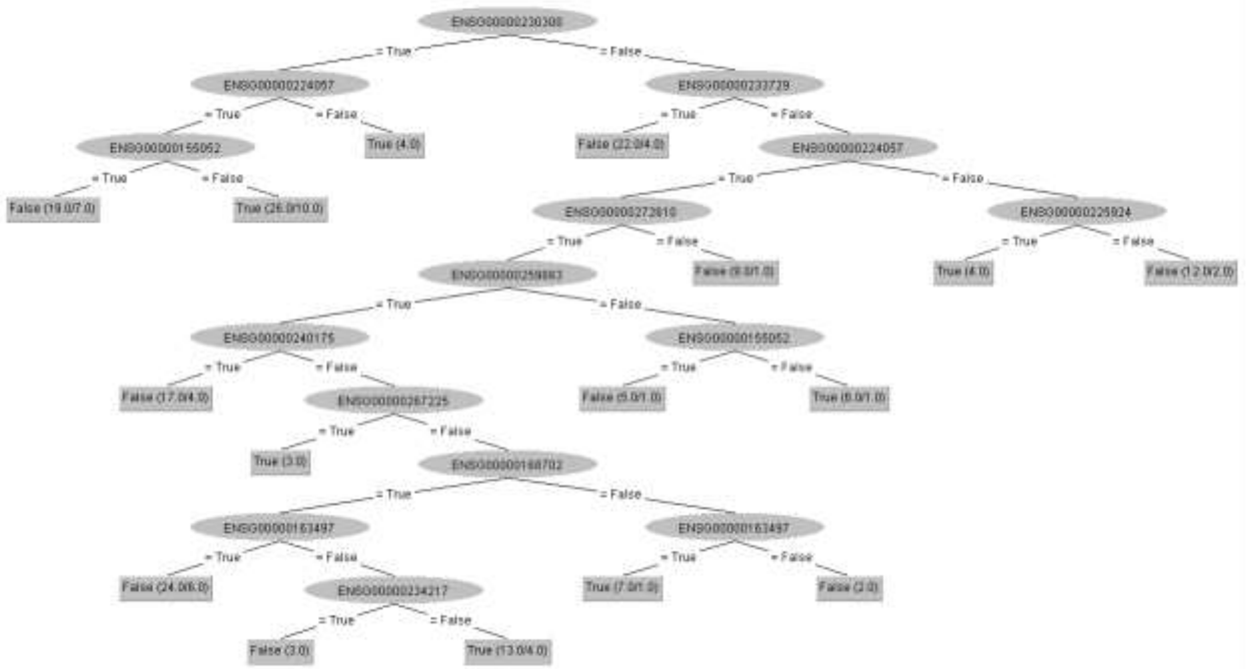

Figure 2: Tree produced by J48 Decision Tree for 2 year survival rates.

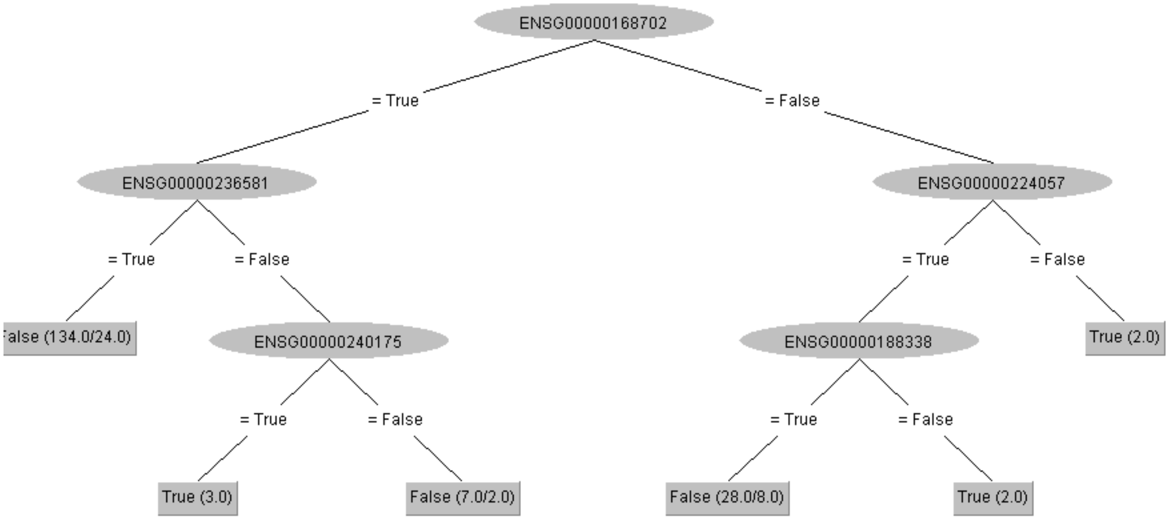

Figure 4: Tree produced by J48 Decision Tree for 3 year survival rates.

4 DISCUSSION

From the most frequently occurring genes in the models as can be seen in Table 2 and in the trees, it could be deduced that the genes that most prominently affected the prognosis of pulmonary AD patients could be divided into two separate categories.

First are the genes that directly have a role in the regulation of the ERK1 and ERK2 cascade as well as the MAPK cascade. The ERK cascades are part of the central signaling pathway and regulate many of the stimulated cellular processes such as cell proliferation, differentiation, and survival, as well as cell apoptosis and stress response. (Wortzel, Seger) It also plays an important role in cross talk between other cellular pathways, implying that it may have an effect on the tumor cell growth and suppression. The MAPK cascade is also a central signaling pathway which regulates stimulated cellular processes, much like the ERK cascades. (Plotnikov, et al.) Some of these genes include MAGI1, TAB2, and ERBB. MAGI1 is a connector enhancer kinase suppressor of RAS 3 and plays a role in the negative regulation of the ERK1 and ERK2 cascade. MAGI1 also has a role in stabilizing cell junctions and has been known cause migration and invasion of cells when silenced. (Zaric et al.) Overexpression has also been known to suppress primary tumor grown and spontaneous lung metastasis. It has also been found to have significant prognostic value in colorectal cancer prognosis, suggesting that it may also have an impact on lung cancer prognosis as well. (Protein Atlas) TAB2 is a TGF-beta-activated kinase which positively regulates I-kappaB kinase, which is transcription factor controlling inflammatory and immune responses. In addition to regulating the MAPK cascade, it also negatively regulates autophagy, which is a process in which cells digest parts of their own cytoplasm in order to remove potentially harmful dysfunctional parts of the cell machinery. MAGI1 and TAB2 are driver genes which do not come up often in the discussion of gene expression in cancer prognosis, and not much literature can be found on the subject. However, based on their ontology, as well as their high frequency within the models created, it may be inferred that they do play a role of some significance in lung cancer prognosis. In the case of TAB2 and its role in controlling inflammatory responses, many clinical studies suggest a strong association between inflammation and cancer. (Gomes, et al.) Inflammatory reactions often promote tumor progression and may also affect the proliferation of malignant cancer cells, as well as tumor response to chemotherapeutic drugs.

Secondly, several genes that regulate the production and organization of microtubule filaments have been found to occur often in the models created, including EML4 and NCKAP5. EML4 is the Echinoderm microtubule-associated protein-like 4 and plays a role in microtubule binding, as well as microtubule cytoskeleton organization in the mitotic cell cycle. In addition, when EML4 and ALK fusion occurs, it may affect the regulation of cellular survival and proliferation. Drugs specifically targeted towards EML4-ALK-positive non-small cell lung cancers have recently been approved by the FDA and have been successful as a targeted cancer drug. NCKAP5 is an NCK-associated protein and plays a role in microtubule bundle formation as well as microtubule depolymerization. Microtubules play a vital role in the mitotic cycle, especially in the accurate separation of chromosomes during mitosis, influencing the genetic integrity of product cells. Changes in microtubule expression and stability have been reported for a wide range of cancers, often being associated with poor prognosis, as well as chemotherapy resistance in solid cancers such as lung adenocarcinomas. (Parker, et al.) Though the mechanisms of the relation of microtubules to cancer prognosis are not very well known, it seems reasonable to deduce that mutations of genes controlling microtubule organization and production may have significant prognostic value.

5 CONCLUSION

Some potential ways to improve accuracy of classifiers may first be to obtain more data from more patients and to train the classifiers on a less skewed dataset in which the survival rates of patients are more evenly distributed. In addition, further specification and modification of the classifier models, such as editing of multilayer perceptron hidden layers or of the weighting of various datapoints may allow for increase accuracy in prognosis when using these classifiers. Furthermore, incorporation of clinical data into the lists of attributes passed into the classifiers along with the gene expression data may allow for more accurate prognosis. For instance, survival rate varies drastically in relation to the stage at which the cancer is initially diagnosed; patients diagnosed at the local stage typically have a 60% survival rate, while those diagnosed at distant stage on the SEER scale have a 6% chance of survival statistically. Factors such as age of diagnosis, as well as other conditions such as diabetes or blood pressure may affect prognosis as well. The incorporation of such clinical data to prognostics in addition to gene expression data may allow for more object and accurate prognosis for lung adenocarcinoma patients.

Overall, it may be concluded that while the three classifiers tested in this project did not produce the most accurate prognosis and may be skewed or overfitted by the characteristics of the training dataset, the information they provided on potential driver genes for lung adenocarcinomas and prognosis proved insightful. In particular, several genes not typically associated with lung cancer but with related processes such as MAGI1 and TAB2, as well as those associated with microtubule organization such as EML4 and NCKAP were found. While the results of this project do not provide any information about the prognostic value of these genes in real clinical situations, further research into the topic specifically based on the expression of these genes may result in the discovery of potential gene target therapies and other effective methods of cancer treatment.

6 Works Cited

American Cancer Society. “Key Statistics for Lung Cancer.” *American Cancer Society*, www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.

“Cancer Candidate Gene Database (CCGD).” *Cancer Candidate Gene Database (CCGD)*, ccgd-starrlab.oit.umn.edu/.

Gomes, Mónica, et al. “The Role of Inflammation in Lung Cancer.” *Advances in Experimental Medicine and Biology*, U.S. National Library of Medicine, 2014, www.ncbi.nlm.nih.gov/pubmed/24818717.

National Cancer Institute. “Encyclopedia.” *GDC Docs*, docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM-UQ/.

Parker, Amelia L, et al. “Microtubules and Their Role in Cellular Stress in Cancer.” *Frontiers in Oncology*, Frontiers Media S.A., 18 June 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4061531/.

Plotnikov, Alexander, et al. “The MAPK Cascades: Signaling Components, Nuclear Roles and Mechanisms of Nuclear Translocation.” *Biochimica Et Biophysica Acta (BBA) - Molecular Cell Research*, Elsevier, 16 Dec. 2010, www.sciencedirect.com/science/article/pii/S0167488910003228.

“The Cancer Genome Atlas.” *National Institutes of Health*, U.S. Department of Health and Human Services, tcga-data.nci.nih.gov/docs/publications/tcga/.

“The Human Protein Atlas.” *The Human Protein Atlas*, 15 Nov. 2018, www.proteinatlas.org/.

University of Wakaito. “Weka 3: Machine Learning Software in Java.” *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*, www.cs.waikato.ac.nz/ml/weka/.

Wortzel, Inbal, and Rony Seger. “The ERK Cascade: Distinct Functions within Various Subcellular Organelles.” *Genes & Cancer*, SAGE Publications, Mar. 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3128630/.

Zaric, J, et al. “Identification of MAGI1 as a Tumor-Suppressor Protein Induced by Cyclooxygenase-2 Inhibitors in Colorectal Cancer Cells.” *Oncogene*, U.S. National Library of Medicine, 5 Jan. 2012, www.ncbi.nlm.nih.gov/pubmed/21666716.