Reinvestigating the Gene Expression in the Diagnostic Evaluation of Smokers with Suspect Lung Cancer

XI CHEN, LAP PUI CHUNG, SICHENG YANG

1. ABSTRACT

Lung cancer is the second most common cancer in the United States. According to the American Cancer Society (ACS), around 130,180 people will die from lung cancer in 2022[4]. Early detection of cancer can significantly reduce the chance of dying. In 2007, a paper was published that investigated the gene expression in cells from people that were suspected of having lung cancer. In the end, the research team was able to identify 80 genes that were associated with lung cancer. Using those 80 genes to classify cancer (n=180), the research team was able to achieve an accuracy of 83%. We used the same dataset from the research team with the hope of finding a model that can achieve higher accuracy. At the same time, we hoped to identify genes that the research team did not identify. With various tuned models that we developed, we were able to achieve an accuracy of 72 ± 2.13%. In addition, using a decision tree model, we identified three new genes that the research team missed. These findings not only helped us learn more about the performance of machine learning models in the field of genetics but also provided more insights into genes that are associated with cancers.

2. INTRODUCTION

2.1 Background information

Cancers are the leading cause of death besides heart diseases[5]. Among all types of cancer, lung cancer is one of the most lethal. Within one year of diagnosis, more than half of the people died[6]. However, when diagnosed at an early stage using imaging techniques such as CT scans, the death rate can be reduced by 14 to 20%[9][10]. Imaging techniques have helped detect cancers at a macroscopic level. On the other hand, microscopic concepts such as epigenetics also have the potential to be implemented as a technique to detect early stage cancers.

Epigenetics is a subfield of genetics. It is the study of how one's behavior and environment can change the way genes express in the body[11]. Genes are sequences of DNA that code for proteins. Some genes are essential for daily function, such as body moving, so they are

expressed all the time. However, in most cases, genes are only expressed under certain conditions. In the case of cancer, cancerous cells have different gene expressions than regular cells, and these variations can play an important role in cancer classification.

Cancer epigenetics specifically targets the varying gene expressions between cancerous and normal cells. Between the years 1980 and 2018, there were approximately 147,000 citations from the 100 most-cited papers on cancer epigenetics[12]. It is an unsaturated field with a lot of potential. In our area of interest, the exact epigenetics of lung cancers is not well understood. Various attempts were trying to identify the genetic factors behind lung cancers. One of which was a team from 2007 that studied the gene expression in cancerous and normal lung cells in people that were suspected of having lung cancer.

The research team collected samples from a total of 187 people. Next, the samples were analyzed using DNA microarrays. The microarrays would return numeric results that represented the relative gene expression levels. Using the data collected from the microarray, the team identified 80 associated genes[3]. Moreover, those genes were used for an ensemble model, and the model was able to produce an 83% accuracy[3]. Not limited to just lung cancer, more and more genes have been labeled as associated with cancer. In an ongoing effort to identify more cancer genes, the Catalogue of Somatic Mutations in Cancer (COSMIC) initialized a program called the Cancer Gene Census (CGC). As of today, the CGC contains a total of 729 genes that are associated with cancers[13].

2.2 Statement of the task

The 2007 study had two major goals. One was to identify genes that are associated with lung cancer. Another goal was to create a combinational way, which involved macroscopic and microscopic analysis, to detect lung cancer at an early stage. With the dataset remaining available to the public, our study aimed to reinvestigate the dataset and possibly provide a better model using random forest, XGBoost, and neural networks. At the same time, we hoped to identify genes that the research team missed back in 2007.

3. METHODS

3.1 Data processing

3.1.1 label transformation

The gathered data was already cleaned with RMA normalization[3]. First, there were samples that were considered as neither having nor not having cancer, so those samples were removed. For binary classifications, the remaining labels were converted from strings to integers. A value of 0 refers to people not diagnosed with cancer, and a value of 1 refers to the opposite.

3.1.2 axis flipping and features isolation

The dataset contains 22,216 genes for each sample. Due to the limited capacity of CSV files, the dataset was originally stored with the axis flipped. To make the original data available for machine learning, we flipped the horizontal and vertical axis. Figure 1 below is our flipped-axis data.

	Diagnostic	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	1316_at	1320_at	1405_i_at	...	90610_at	91580_at	91617_at	91682_at	9
0	0.0	10.696879	4.23617	5.345251	7.919187	3.478706	7.142456	4.814028	4.112987	6.948436	...	7.373103	3.232875	5.466897	5.952681	
1	1.0	10.560653	4.173343	5.788414	8.180133	3.745022	7.667475	5.254227	4.303188	5.243215	...	7.714989	3.491614	5.597842	6.394273	
2	1.0	9.939516	4.559698	5.924607	8.207125	3.630689	7.807055	5.364942	4.498574	5.108143	...	8.707332	3.655863	5.934463	6.547563	
3	1.0	10.299866	4.359738	5.583276	8.213886	3.653105	7.5098	5.103016	4.137036	5.450892	...	7.253566	3.333656	5.259695	6.299795	
4	1.0	10.467122	4.245888	5.785865	8.184135	3.718874	7.588953	5.233377	4.029887	4.745138	...	7.898104	3.511188	5.409353	6.27696	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
182	0.0	9.867062	4.347193	6.077813	8.386588	4.069206	7.144349	5.638766	4.329403	4.826731	...	7.786588	3.599751	5.267918	6.524704	
183	0.0	9.655055	4.556167	5.856299	8.251081	4.003426	7.345543	5.531227	4.647367	6.8435	...	8.474055	3.606413	5.693594	6.505405	
184	0.0	10.072952	4.356708	5.657181	8.118552	3.887196	7.696534	5.335557	4.48411	6.477846	...	8.247335	3.769988	6.173629	6.379607	
185	0.0	10.407795	4.37766	5.667295	8.110224	3.713854	7.621459	5.252291	4.241374	6.598051	...	8.028244	3.546674	5.857019	6.226623	
186	0.0	9.622821	4.359618	6.150318	8.277467	3.829989	7.255572	5.418568	4.457643	6.83864	...	7.917359	3.518806	5.2645	6.441155	

187 rows × 22216 columns

Figure 1: Part of Dataset

Due to computational constraints, instead of doing classification with all genes available in the dataset, we stuck with the 80 genes identified from the 2007 research team. As a result, there were 80 features for our training models.

3.1.3 train set and test set

The dataset was split into a train set and a test set; 70% of data was used as a train set, and 30% as the test set. Train set aimed at training and tuning the models. Because GridSearchCV and XGBoost both perform cross-validation, the validation set was eliminated. The test set was used to test the performance of each model and check overfit.

To make sure the train_test_split() function offers us the same train and test result, the random state was set to a constant value of 888.

3.2 Apply data to models and achieve better scores using GridSearchCV

After the models were trained, they would predict cancer based on the test set; and confusion matrices were used to calculate and visualize accuracy for each model. Accuracy on the test set would be used to evaluate models' performance.

3.3 Use decision tree to find new genes

Different from above, a decision tree was trained and tuned with the full train set (22,216 genes for each individual). Because the decision tree usually overfits the data, we tuned it with various cost complexity pruning (ccp_alpha) values. According to figure 2 below, ccp_alpha = 0.085 minimized the overfit.

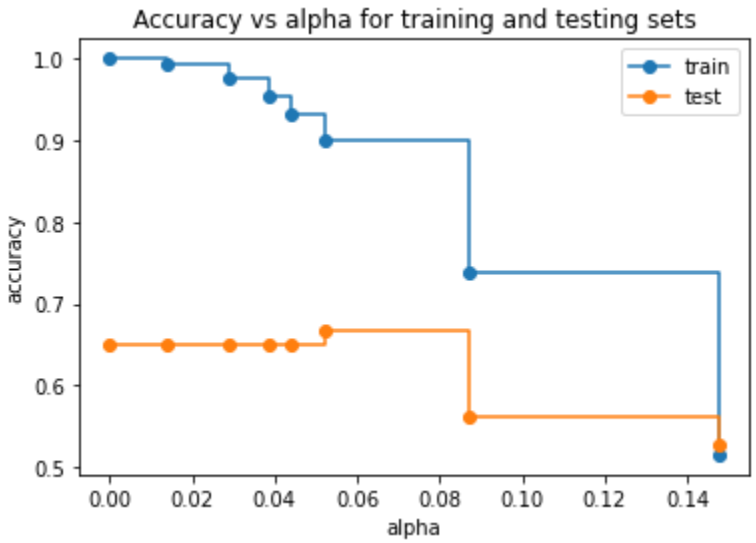

Figure 2: Accuracy vs Alpha for Training and Testing Sets

After that, we exported the inner structure of the tree. By looking at key gene parameters, we were able to know more genes that are associated with lung cancer.

4. ANALYSIS AND RESULTS

To make results stable for each run, our team chose 888 to be the numpy seed and models' random state.

4.1 Performance for each model

Figure 3 below plots the accuracy for each model; and table 1 in the appendix contains the best parameters, accuracy, and confusion matrix of each model.

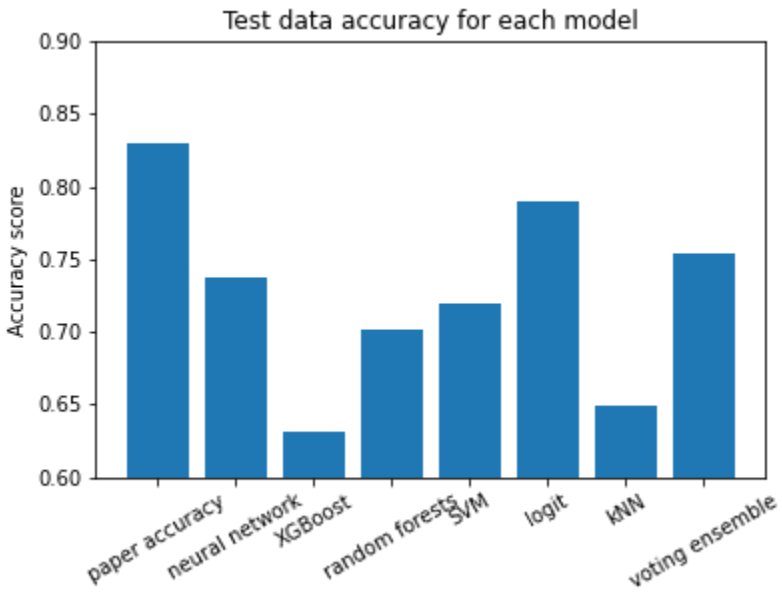

Figure 3: Test Data Accuracy for Each Model

4.2 Analysis on performance

Our best model was logistic regression with an accuracy of 78.9% on the test set. It was unfortunate that none of our models outperform the original research's model (83% accuracy)[3]. However, according to figure 3, as all models have greater than 60% accuracy, and two of them are beyond 75%, our models are acceptable.

Because gene expression is not the sole factor of lung cancer, it is reasonable to have 60% - 80% accuracy. It would be surprising if a model could achieve 90% accuracy. According

to the Center for Disease Control and Prevention (CDC)[2], smoking, radon, radiation therapy, etc could also cause lung cancer; without information on these factors outside of genetics, it was hard to reach high accuracy. Therefore, 60% - 80% accuracy was acceptable and reasonable. In addition, our models mainly made mistakes on false-negative samples.

False-negative samples were cancer patients wrongly classified as healthy people. Because our trained data did not include all possible genes that cause cancer, our models might make mistakes because of those unidentified genes.

4.3 Decision tree models to find more genes

Because it is possible to look through how decision trees classify samples, we are able to obtain genes that are associated with lung cancer. According to table 1 in the appendix, the accuracy of decision trees after pruning was 66%. This means the genes used in the decision tree might be important factors of lung cancer. Figure 4 below is the inner structure of our decision tree trained on the full train set.

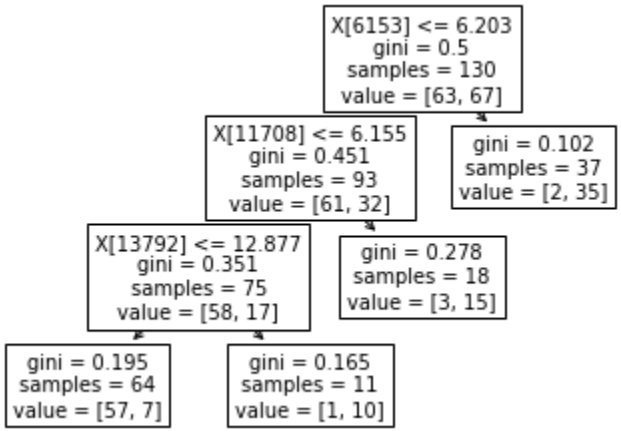

Figure 4: Decision Tree Structure

According to figure 4, there are three new genes X[6153], X[11708], and X[13792] that the decision tree uses to classify patients; according to the dataset, they are 206628_s_at, 212324_s_at, and 214414_x_at. When translated to actual gene names, they are SLC5A1, VPS13D, and HBA1 respectively. Research[1] shows that SLC5A is linked to skin cancer. A gene detection patent[7] marks VPS13D as one of the breast cancer genes. Lastly, HBA1 is a factor in

leukemia[8]. Because the abnormal expression of these 3 genes may result in other types of cancer, it is also possible that these abnormal expressions may lead to lung cancer.

5. SUMMARY AND CONCLUSION

All in all, with various tuned models that we developed, we were able to achieve an accuracy of 72 ± 2.13%. The best model was logistic regression which achieves 78.9% accuracy.

We also found 3 genes—SLC5A1, VPS13D, and HBA1—that may be factors of lung cancer. These genes were also confirmed by other research and government agencies. We hope our results could make a contribution to future cancer research.

5.1 Future research

There are a few improvements that we could make. First, our future research on model training could be adding features outside of genetics; according to the CDC[2], they can be risk factors such as patients' smoking frequency, radiotherapy history, etc. We believe all models could have better accuracy with these newly added data if they are available.

Moreover, we could also do more research on the three newly found genes. Research shows they may result in cancer in other organs, but we could not find any research showing the relationship between lung cancer and these genes. As our decision tree could achieve 66% accuracy based on these genes, we believe these genes are factors of lung cancer.

6. ACKNOWLEDGEMENT

We appreciate Prof. Steinwand from Augustana University for offering advice and assistance throughout our research. We also appreciate the National Center for Biotechnology Information (NCBI) for keeping records of the dataset.

APPENDIX:

[1] Program for Objective 1: COSC380_final_project_obj1.ipynb in attachment

[2] Program for Objective 2: COSC380_final_project_obj2.ipynb in attachment

[3] Table 1:

Model Name	Best Parameters	Confusion Matrix	Accuracy on Test Set
Neural Network	'alpha': 0.1, 'hidden_layer_sizes': 7, 'max_iter': 1500, 'random_state': 888, 'solver': 'lbfgs'	[20 7 8 22]	0.737
Logistic Regression	'C': 0.1, 'penalty': 'l2'	[22 5 7 23]	0.789
Random Forest	'criterion': 'gini', 'n_estimators': 100, 'random_state': 888	[22 5 12 18]	0.702
XGBoost	'subsample': 0.6, 'n_estimators': 500, 'max_depth': 10, 'learning_rate': 0.3, 'colsample_bytree': 0.9, 'colsample_bylevel': 0.9	[20 7 14 16]	0.632
SVM	'C': 0.1	[20 7 9 21]	0.719
K-Nearest Neighbor	'leaf_size': 20, 'n_neighbors': 4, 'p': 2	[19 8 12 18]	0.649
Voting Ensemble	KNN, SVM, and Logistic	[21 6 8 22]	0.754
Decision Tree	criterion = 'gini', ccp_alpha = 0.085	[19 8 11 19]	0.667

REFERENCES

[1] Josefa P. Alameda et al. 2016. IKKA regulates the stratification and differentiation of the epidermis: Implications for Skin Cancer Development. (November 2016). Retrieved January 25, 2022 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5363549/

[2] CDC. 2021. What are the risk factors for lung cancer? (October 2021). Retrieved January 25, 2022 from https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm

[3] Avrum Spira et al. 2007. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. (March 2007). Retrieved January 18, 2022 from https://www.nature.com/articles/nm1556

[4] American Cancer Society. 2022. Lung cancer statistics: How common is lung cancer? (January 2022). Retrieved January 26, 2022 from https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

[5] CDC. 2022.(January 2022). Retrieved January 26, 2022 from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

[6] U.S. National Institute Of Health, National Cancer Institute. SEER Cancer Statistics Review, 1975–2015.

[7] Van Ryan. 2013. GENE MARKER SETS AND METHODS FOR CLASSIFICATION OF CANCER PATIENTS. (August 2013).

[8] Gene Visible. Team at Nebion AG. 2016.(2016). Retrieved January 26, 2022 from https://genevisible.com/cancers/HS/UniProt/P69905

[9] de Koning HJ, Meza R, Plevritis SK, ten Haaf K, Munshi VN, Jeon J. Benefits and Harms of Computed Tomography Lung Cancer Screening Strategies: A Comparative Modeling Study for the U.S. Preventive Services Task Force. Annals of Internal Medicine. 2014; 160(5):311-20. doi:10.7326/M13-2316.

[10] Aberle DR, Adams AM, Berg CD, et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. New *England Journal of Medicine*. June2011; 365(5):395-409. doi:10.1056/NEJMoa1102873.

[11] CDC. 2020.(August 2020). Retrieved January 26, 2022 from https://www.cdc.gov/genomics/disease/epigenetics.htm#:~:text=Epigenetics%20is%20the%20study%20of,body%20reads%20a%20DNA%20sequence.

[12] Ignacio Jusue-Torres, Joshua E. Mendoza, Malcolm V. Brock, and Alicia Hulbert. 2020. The

100 most cited papers about cancer epigenetics. (April 2020). Retrieved January 26, 2022 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7213660/

[13] COSMIC. 2021. Cancer gene census. (November 2021). Retrieved January 26, 2022 from https://cancer.sanger.ac.uk/census