SCHOLARLY REVIEW

Machine Learning-Based Prediction on Diagnosis and Severity of Lung Cancer

By Solon Chan

Author Biography

Solon Chan is currently a junior attending Folsom High School. He has been very interested in STEM studies since a young age, especially in mathematics, cyber security and computer science. He self-learned Python and Java programming and enjoys utilizing computer techniques and machine learning to resolve real world problems. He founded his school's Cyber Security & Computer Science Club, in which he serves as the president and competition captain, leading multiple NASA challenges and cyber security competitions. With chemistry also being his favorite subject, he considers biochemistry as future career. Outside of school, Solon is an avid musician as he was trained classically with piano (9 years) and violin (8 years), receiving awards from many local and international competitions and currently a member of Sacramento Youth Symphony. In addition, he enjoys volunteer tutoring violin, piano, music theory and STEM subjects to promote these educations.

Abstract

Machine learning, combined with big data, is a powerful tool that can help researchers see the underlying and hidden patterns and correlations, to which we have not found the links and causations. This research seeks to make lung cancer diagnosis and severity predictions utilizing data science and machine learning tools with Python programming. Neural network classification models are built for both predictions and are trained with collected lung cancer datasets. Various data processing, modeling, and regression techniques are experimented and compared for analysis and the best accuracy. It is discovered that, while predicting lung cancer diagnosis remains inaccurate with only around 50-60% accuracy, predicting the severity of lung cancer has almost perfect outcomes. This proves that even though we can see positive correlations between lung cancer and its risk factors, predicting a positive diagnosis is complicated, and it cannot replace the actual medical diagnosis yet. However, the severity prediction after a positive diagnosis is highly accurate, and it can provide guidance and validation for further lung cancer triage.

Keywords: machine learning, neural network, artificial intelligence, data science, cancer, lung cancer, diagnosis, prediction, python, tensorflow, logistic regression

Introduction

Lung cancer is the second most commonly occurring cancer (World Health Organization [WHO], 2022) yet it is responsible for more deaths than any other cancer in the world. In the United States, it is the third most commonly occurring cancer for men and women (National Cancer Institute [NCI], n.d.). One in five of all cancer deaths is caused by lung cancer (American Cancer Society [ACS], n.d.). It is a silent killer, evading detection of the immune system before doctors can finally diagnose it. One of the reasons that it is so deadly is that when the symptoms have started to show, the lung cancer has already become advanced (Mayo Clinic, 2024). Though most people will not get checked for lung cancer until it is too late, blindly screening everyone yearly is costly and impractical for the healthcare system. This is the reason why it is imperative to determine the main risk factors and symptoms and how they put one at risk of lung cancer. There can be tens, if not hundreds of variables for researchers to sort out the relationships. With the new machine learning and data science methodologies, we have just the right tool to find the risk and also to diagnose the severity of lung cancer with the trained neural network model.

Literature review

Risks of lung cancer include smoking, air pollution (including air pollution from occupational hazards), genetic risk/family history, and age (ACS, n.d.). Symptoms include coughing, frequent chest infections, fatigue, weight loss, shortness of breath, wheezing, changes in the appearance of fingers (such as clubbing), and chest pain (National Health Service [NHS], n.d.). Life habits such as diet, stress, and lack of sleep can also play a role here. There can also be factors that are not as intuitive, such as peer pressure. It is well-established that smoking increases the risk of lung cancer. Thankfully, the number of people who suffer from lung cancer has decreased due to reduced smoking populations, which proves the high correlation between smoking and lung cancer (ACS, n.d.).

Air pollution is another risk factor. The pollution comes in the form of particles such as acids, chemicals, metals, soil, dust, asbestos, or radon (American Lung Association, 2016). In America, air pollution has been decreasing thanks to laws like the Clean Air Act.

Genetic and ethnicity can also be a risk factor, as the black population is 12% more likely to develop lung cancer than white population (ACS, n.d.).

Gender is considered a factor since males have higher rates of lung cancer than females, but the gap is closing (ACS, n.d.). This could be due to genetics or the difference in lifestyle.

The rate of lung cancer for men has been dropping for a few decades, but the rate of lung cancer for women has only been falling in the last decade. On average, men have a 1 in 16 chance of getting lung cancer and women have a 1 in 17 chance, but the chances are increased when they smoke (ACS, n.d.). There are many theories on this phenomenon. One of them states that women tend to start smoking at a later age, hence the benefit of reduction in smoking habit is delayed for the female population.

Methods

Two datasets are used in this project. One is concerned with lung cancer diagnosis and the other one is concerned with lung cancer severity. Both datasets are taken from Kaggle dataset collections. Based on the information written in the description, the dataset is compiled by students at the JIS College of Engineering in Kalyani, West Bengal, India (Nath & Biswas, 2024). The other dataset is collected from Kaggle (Kaggle, 2022), which is in turn collected from data.world (data.world, 2017), which is concerned with lung cancer severity.

Other than some common and intuitive factors included with both datasets, other factors could be indirect, such as yellow fingers, clubbed fingers, residence, and even family income, which could correlate to lung cancer.

One observation is how the variables are classified. Some variables are simple binary yes and no, and some variables like age, would be a spectrum of values. The column can also have a classification of limited and distinct categories.

To apply machine learning to establish a prediction model, Python programming language is utilized to pre-process the data and to call various data science/machine learning libraries. This is the most

common way to explore the data. Numpy, matplotlib, pandas, and seaborn libraries are imported to process and visualize the data. TensorFlow, Keras, and scikit-learn are included for the actual machine learning procedures such as data training. Instead of running the program locally, Google Colab can be used for free since the model in the research is not demanding.

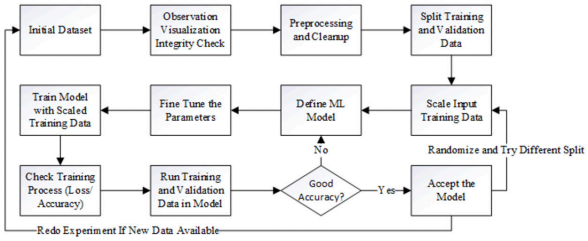
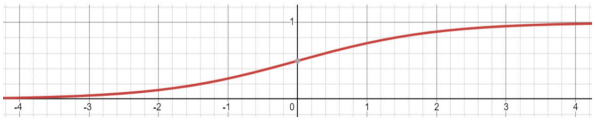


Figure 1. Machine learning flow employed in this research

After the dataset is obtained, first we need to observe the data. Each column and what it represents need to be well defined. Data needs to be reasonable in size (number rows) and free of blanks, especially for the columns with numerical values. We can also plot the data or do statistical analysis and see if the data is lopsided or skewed toward a particular outcome. Each factor (column) needs to have a balanced representation of values.

The next step is to preprocess the data which includes eliminating the unnecessary columns and removing/patching the blank data. Seriously incomplete rows also need to be dropped. Once we have a healthy dataset, we can go ahead to split the data into training and validation sets. A randomized 75-25 split is what we choose to implement. The training data is then scaled based on its average and standard deviation. This step helps the model's convergence and accuracy. The training dataset is used to train the model, while the validation set is used to verify the model.

Once the data is ready, we can now define the model. Logistic regression, which is different from linear regression, is a statistical method useful for binary classification. It is based on a sigmoid function (logistic function), which would saturate to 0 or 1 quickly and reach a binary decision.



$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}} = 1-\sigma(-x)$$

For the binary classification model, we start with 1 neuron with sigmoid activation. Based on the performance of the model, we can decide to add more ReLU (Rectified Linear Unit) activation layers with more neurons for improvement. This will make it a neural network model. There are other parameters to play with in the iteration, for example, the lambda parameter needs to be tweaked if the model is overfitting. We can observe that we have much higher accuracy on the training data than the validation data. The epoch parameter can also result in overfit or underfit and it has a big impact on the run time.

Finally, the scaled training data is fed to the model. We also monitor the loss and accuracy of the process. Once we have a model, the unscaled validating data is run through the model and we can check if the accuracy is satisfactory. Unscaled training data should also be tested in the model and compared with validation results. This way we can determine if we have an overfitting issue so that the process can be fine-tuned for improvements.

Lung Cancer Diagnosis Dataset

This dataset contained factors for lung cancer risk, habits that increase the risk of lung cancer, symptoms that typically define lung cancer, and finally, the classification of whether a person was diagnosed with lung cancer. The dataset contained 3000 rows of data usable data. It is mostly balanced, as shown in the tables on the Kaggle dataset (Nath & Biswas, 2024). The data contained only people between the ages of 30 and 80. The 1s and 2s under all columns except for "Lung Cancer" and "Gender" represent "Yes" and "No," respectively. Listed below are the first two rows of the data, and one row of factors and diagnoses.

Table 1. Sample data from the diagnosis dataset. 2 rows included. 3000 rows in total.[7]

Gender	Age	Smoking	Yellow Fingers	Anxiety	Peer Pressure
M	65	1	1	1	2
F	55	1	2	2	1

Chronic Disease	Fatigue	Aller	Wheezing	Alcohol Consuming	Coughing
2	1	2	2	2	2
1	2	2	2	1	1

Shortness of Breath	Swallowing Difficulty	Chest Pain	Lung Cancer
2	2	1	NO
1	2	2	NO

During data preprocessing, M and F are replaced with 0 and 1 respectively, and 1 and 2 are replaced with 0 and 1 respectively for the sake of consistency. The NO and YES in the diagnosis column are replaced with 0 and 1 respectively.

Lung Cancer Severity Dataset

This dataset contained factors for lung cancer risk, habits that increase the risk of lung cancer, symptoms that typically define lung cancer, and finally, the data on the severity of the cancer after a positive diagnosis. The dataset contained 1000 rows of data usable data. However, some of the descriptions about the data are vague, so assumptions have to be made. We assume that the 1s and 2s in the Gender column represent "Male" and "Female" based on the Kaggle dataset tables. Since there are more 1s than 2s, and males are more likely to get lung cancer than females, 1s is assumed to be male. We also assume that the numbers under each category except for "Gender" or "Level" to be a scaling system, where a higher number meant a more severe symptom or a more frequent habit. Listed below are the first two rows of the data, and one row of factors and diagnosis.

Table 2. Sample data from the severity dataset. 2 rows are included. There are 1000 rows in total.[5]

Patient ID	Age	Gender	Air Pollution	Alcohol use
P1	33	1	2	4
P10	17	1	3	1

Dust Allergy	Occupational Hazards	Genetic Risk	Chronic Lung Disease	Balanced Diet
5	4	3	2	2
5	3	4	2	2

Obesity	Smoking	Passive Smoker	Chest Pain	Coughing of Blood
4	3	2	2	4
2	2	4	2	3

Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty
3	4	2	2	3
1	3	7	8	6

Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
1	2	3	4	Low
2	1	7	2	Medium

During processing, the Patient ID column is removed since it is not relevant. The "Level" column is split into 3, by 1-hot encoding: a column each for "Low", "Medium", and "High". If a patient's severity level is low, the "Low" would have a 1 for that classification and 0 for the rest ("Medium" and "High") as shown below. The amount of "Low", "Mediums", and "High" are similar, indicating a balanced dataset.

Table 3. Sample data from the severity dataset. 2 rows included. Multiple classification columns are transformed into multiple classification columns.

Low	Medium	High
1	0	0
0	1	0

Models

For this research, two models are created in Google Colab and trained to diagnose lung cancer. Both models use RMSprop optimizers. Both datasets are split into a randomized 75% training set and a 25% validation set.

Lung Cancer Diagnosis Model

We utilize a neural network with multiple ReLU activator layers with multiple neurons and a final layer with a sigmoid activator. The final layer has a single neuron that represents a single binary classification of the diagnosis. The neural network has regularization to remove overfitting, with a small lambda value of 0.01. We have difficulty obtaining a model with good accuracy. After playing around with the number of layers, different activators, number of neurons, lambda, and epoch values, and even different data split and randomization, we are saturated at around 60% accuracy. We found that adding more than 4 layers does not result in any more improvement. Decreasing the lambda seems to heighten the overfitting issue and make the training data accuracy over 80%, while validation data accuracy lingers below 60%. Increasing the epoch from 100 does not seem to have much effect other than just increasing the run time. Eventually, we settle with 1 ReLU layer with 64 neurons. Since a satisfactory prediction cannot be obtained from the validation dataset with all the experiments, and there is no obvious issue of over-

representation with the dataset itself, in the end, we have to attribute the inaccuracy to the nature of the data. Here is the code segment related to the model specification:

```
lam = 0.01 #Lambda value to remove overfitting
model = Sequential() #Create sequential layers
model.add(Dense(64, activation='relu', kernel_regularizer=regularizers.l2(lam), input_dim =
X_train_scaled.shape[1])) #Extra ReLU layer with 64 neurons
model.add(Dense(1, activation='sigmoid', kernel_regularizer=regularizers.l2(lam))) #Logistic regression
with 1 neurons and sigmoid activator
model.compile(optimizer='RMSprop', loss='binary_crossentropy', metrics = ['accuracy']) #Compile model
with RMSprop optimizer
model.fit(X_train_scaled, y_train, epochs=100, verbose=0, validation_data = (X_val_scaled, y_val))
#Train the model with scaled training data, epoch = 100
```

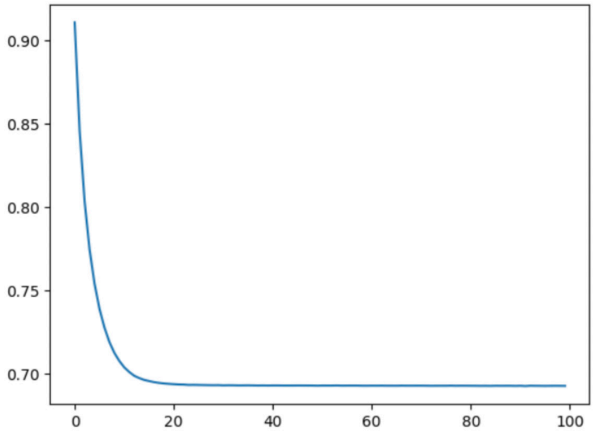


Figure 3. The plot of the loss value for each epoch on the lung cancer diagnosis model, showing the convergence. Epochs = 100

Lung Cancer Severity Model

On the contrary, for the severity dataset and model, we can only employ a simple logistic regression and obtain an accurate prediction. It is constructed with a final softmax activator layer with 3 neurons, to represent the 3-class diagnosis. Since the result is very good, there is no need to add extra layers and make it a neural network implementation. Here is the code segment related to the model specification:

```
lam = 0 #Lambda value to remove overfitting
model = Sequential() #Create sequential layers
model.add(Dense(3, activation='softmax', kernel_regularizer=regularizers.l2(lam), input_dim =
X_train_scaled.shape[1])) #Logistic regression with 3 neurons and softmax activator
model.compile(optimizer='RMSprop',loss='categorical_crossentropy', metrics = ['accuracy']) #Compile
model with RMSprop optimizer
model.fit(X_train_scaled,y_train,epochs=150,verbose=0, validation_data = (X_val_scaled,y_val)) #Train
the model with scaled training data, epoch = 150
```

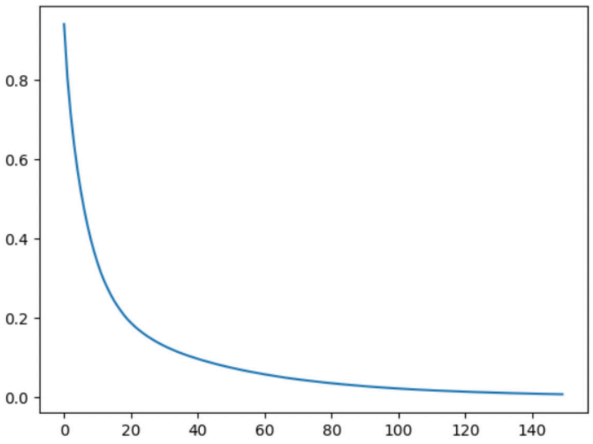


Figure 4. The Plot of the loss value for each epoch on lung cancer severity model, showing the convergence. Epoch = 150

Results

The resulting models are highly dependent on the datasets, as the diagnosis results are marginally over 50% and the severity results are almost perfect.

Diagnosis Model

The logistic regression model only has a 53% accuracy on the training set and a 52% accuracy on the validation set. The neural network with 1 extra ReLU layer on 64 neurons performed better, with 57% and 55% accuracy for the training and validation sets, respectively. Any extra optimization results only in marginal improvement but a much longer run time. We can raise the training data accuracy to over 80%, but it is proven to be inaccurate since it is a result of overfitting. The code is scrutinized and is void of any errors, so the assumption is that it is natural with this dataset. To verify this hypothesis, we single out the smoking data (smoking is a well-known cause of lung cancer) and count the number of people who smoked and had lung cancer, the number of people who did not smoke and had lung cancer, and the opposite scenarios. All four sets of numbers are in a similar range, which reveals that the dataset is not biased toward cancer diagnosis for smoking populations. This unintuitive behavior likely confuses the model for a good prediction. Some other factors also show the same behavior and the resulting model cannot classify a diagnosis with confidence.

```
[ ]
y_train_hat_cat = 1*(model.predict(X_train_scaled) > 0.5)
#y_pred = (y_pred_prob > 0.5).astype(int).flatten()
print(classification_report(y_train,y_train_hat_cat, zero_division=0))

71/71 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
              precision    recall  f1-score   support

           0       0.58      0.45      0.51      1113
           1       0.56      0.69      0.62      1137

    accuracy                           0.57      2250
   macro avg       0.57      0.57      0.56      2250
weighted avg       0.57      0.57      0.56      2250
```

Figure 5. Accuracy of the training data for the lung cancer diagnosis model = 0.57, note that a prediction of 1 is more accurate than 0, meaning a false negative is more likely than a false positive.

```
[ ]  y_val_hat_cat = 1*(model.predict(X_val_scaled) > 0.5)
     print(classification_report(y_val,y_val_hat_cat, zero_division=0))

24/24 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step
              precision    recall  f1-score   support

           0       0.55      0.44      0.49       369
           1       0.54      0.65      0.59       381

    accuracy                           0.55       750
   macro avg       0.55      0.54      0.54       750
weighted avg       0.55      0.55      0.54       750
```

Figure 6. Accuracy of the validation data for the lung cancer diagnosis model = 0.55, note that a prediction of 1 is more accurate than 0, meaning a false negative is more likely than a false positive.

Severity Model

On the contrary, the logistic regression model for the severity dataset returns a 100% accuracy for both the training and validation sets. We remain skeptical, and we assume that the codes have errors, or the dataset is problematic. However, after reviews, we determine that the codes are clean. The dataset has proven very effective for the model. This might be attributed to the fact that the factor columns have a spectrum of values instead of just binary values. Randomizing the training and validation partition results in close to 100% accuracy in the worst case. In that case, adding extra ReLU layers to the model is unnecessary. We decide to live with the simple logistic regression model.

```
[24]
y_train_hat_cat = 1*(model.predict(X_train_scaled) > 0.5)
print(classification_report(y_train,y_train_hat_cat, zero_division=0))

24/24 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       232
           1       1.00      1.00      1.00       252
           2       1.00      1.00      1.00       266

   micro avg       1.00      1.00      1.00       750
   macro avg       1.00      1.00      1.00       750
weighted avg       1.00      1.00      1.00       750
 samples avg       1.00      1.00      1.00       750
```

Figure 7. Accuracy of training data for lung cancer severity model = 100%

```
[25] y_val_hat_cat = 1*(model.predict(X_val_scaled) > 0.5)
     print(classification_report(y_val,y_val_hat_cat, zero_division=0))

8/8 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        71
           1       1.00      1.00      1.00        80
           2       1.00      1.00      1.00        99

   micro avg       1.00      1.00      1.00       250
   macro avg       1.00      1.00      1.00       250
weighted avg       1.00      1.00      1.00       250
 samples avg       1.00      1.00      1.00       250
```

Figure 8. Accuracy of validation data for lung cancer severity model = 100%

Discussion

As the results show, even with 3000 rows, 15 columns of data, and multiple layers of neural network, we still see only about 50 ~ 60% accuracy. This is somewhat in line with the general perception that lung cancer is elusive and hard to predict accurately. One reason could be that the lung is an organ that is exposed to external environments, so there are way more factors to be considered. For just one variable such as smoking, we would need to know not only if a patient is a smoker or exposed to secondhand smoke, but also consider when one starts/stops the habit, and at what age one does that. We would have to build the life history of many patients for the training data, let alone possible genetic information. We have merely 15 columns of factors based on obvious habits and symptoms for this dataset. Understandably, even machine learning with abundant data rows cannot generate satisfactory models. With 50 ~ 60% accuracy, it is only a little better than flipping a coin and guessing. It is possible to use statistics tools to link smoking habits with lung cancer, as well as other factors, however, we still have a long way to go for an accurate prediction.

On the contrary, once lung cancer does occur, predicting the seriousness of lung cancer becomes much easier. Our dataset has 20+ non-binary-value columns which might also contribute to the high accuracy (almost always 100%) of the prediction. However, practically, this prediction is more of a proof of concept because if a patient already has lung cancer, doctors can diagnose for the severity instead of using the machine learning model prediction. In practice, the model can only serve as a precaution when it mismatches the diagnosed result, to urge the doctor to take a second look.

Conclusion

Cancer is at the forefront of modern medical research. It is very deadly, and while we have made great strides, the solution remains elusive. Compared to well-established cancer types such as breast cancer, lung cancer is less predictable. The lung's constant exposure to external factors makes machine learning and data science applications difficult due to the vast amount of data that needs to be collected. We can observe the same problem in our research. With different approaches to building the model, and a good size of data, the diagnosis model is working a little better than a 50-50 prediction, hardly useful in real life.

On the severity model, we have a highly accurate model. This is likely because the factor columns include real diagnosed symptoms that can tell a lot about the seriousness of the cancer. This proves that machine learning methodology can work very well with relevant datasets. We say "You are what you eat", and similarly, the performance of the machine learning model is highly dependent on the data it consumes or is trained on. It is a great tool, but the data is what makes it useful. The collection of a good dataset is the first key to unlocking the power of machine learning when we want to apply it for medical purposes, such as cancer, Alzheimer's, or pharmaceutical research, as well as making real contributions.

Limitations

We observe that the quality of the dataset directly impacts the accuracy and usefulness of the model. The difficulty of gathering the data, either due to privacy, or lack of sources, would be our main limitation. The complexity of figuring out what kind of data is relevant also adds to the problem. Although some factors are not intuitive, including all the nitty-gritty details in the data collection would have no merit and only increase the resources needed to process them. We would need not only medical but also experts from other areas like social scientists, to find a more complete list of meaningful data.

Future Research

The next step of the research would be to find a better dataset. We plan to write a request to NIH for access to their huge anonymous cancer database. It might have enough factors for a more accurate lung cancer prediction model. We can also try different methodologies, such as collecting all different datasets and merging them. However, this process needs to be under heavy scrutiny to ensure data integrity because there will be blank columns for mixed datasets. We can expand the research to include breast and colon cancers by accessing related datasets and comparing the performance of their prediction model against lung cancer models. Since breast cancer receives more attention and research, the data might be more accessible and the results should be relatively easy to validate against real-life cases.

On the methodology side, one thing we would like to explore is the discrete nature of the result column. The datasets we use have either binary values or 3-category values. The sigmoid and softmax activation would pinpoint to each category as a prediction. However, we are wondering if it is possible to obtain a real "likelihood" to the yes/no answer, instead of "snapping" the result to yes/no as in the sigmoid function. This would require the model to generate a wider "linear" region before saturating to 0s or 1s. Changing the yes/no threshold away from 0.5 might also help weeding out non-deterministic results. For example, only $0.7 \sim 1.0$ and $0.0 \sim 0.3$ are counted as "confident" predictions, while $0.3 \sim 0.7$ would be inconclusive. If we look at the result this way, the model might be more accurate than we previously expected.

In addition to predicting the diagnosis and the advancement of the cancer stages, we can also apply the same methodology to predict the prognosis and survival rate of certain treatment protocols.

Machine learning is a powerful tool that might help us figure out a better treatment plan for cancers and the effectiveness of particular drugs. With good data preparation, we might find some factors and correlations that we previously missed or deemed non-intuitive.

Acknowledgements

I would like to thank my advisor from the School of Mathematics of Georgia Tech and his teacher's assistant, for guiding and supporting me through this research. I would also like to thank Kaggle, UCI, and data.world for the great collections of useful datasets for machine learning.

References

[1] American Cancer Society. (n.d.). Lung cancer risk factors: Smoking & lung cancer. Smoking & Lung Cancer | American Cancer Society. https://www.cancer.org/cancer/types/lung-cancer/causes-risks-prevention/risk-factors.html

[2] American Cancer Society. (n.d.). Lung cancer statistics: How common is lung cancer?. Lung Cancer Statistics | How Common is Lung Cancer? | American Cancer Society. https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html

[3] American Lung Association. (2016, June 20). The connection between lung cancer and Outdoor Air Pollution. American Lung Association. https://www.lung.org/blog/lung-cancer-and-pollution

[4] Lung cancer data - dataset by cancerdatahp. data.world. (2017, September 19). https://data.world/cancerdatahp/lung-cancer-data

[5] Lung cancer prediction. Kaggle. (2022, November 14). https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data

[6] Mayo Foundation for Medical Education and Research. (2024, April 30). Lung cancer. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620

[7] Nath, A., & Biswas, A. (2024, June 26). Lung cancer dataset. Kaggle. https://www.kaggle.com/datasets/akashnath29/lung-cancer-dataset/data

[8] National Cancer Institute. (n.d.). Common cancer types. NCI. https://www.cancer.gov/types/common-cancers

[9] NHS. (n.d.). Symptoms. NHS. https://www.nhs.uk/conditions/lung-cancer/symptoms/

[10] World Health Organization. (2022, February 3). Cancer. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer