Revolutionizing Early Lung Cancer Detection with AI and ML Solutions

Om Gagrani

SUMMARY

Lung cancer is the leading cause of cancer-related deaths worldwide, and early detection is critical for improving patient survival. Advances in machine learning (ML) provide powerful tools for analyzing medical data, yet it is unclear which ML approaches are most effective for lung cancer prediction. This study tested the hypothesis that ensemble methods combining multiple ML models would achieve higher predictive accuracy than individual models, reaching over 90% accuracy for both patient demographic models and convolutional neural network (CNN) models. It was further predicted that increasing hyperparameters (such as tree depth, number of estimators, or epochs) would improve accuracy but reduce computational speed. A publicly available dataset from Kaggle containing clinical and diagnostic data was used. Four models were trained and evaluated: Decision Tree, Random Forest, Logistic Regression, and a Classifier algorithm. Their outputs were then combined using the Multiplicative Weight Update Method to create an ensemble prediction. Model performance was evaluated using accuracy, precision, and recall. Results showed that the Decision Tree achieved 92.2% accuracy, Random Forest 95.1%, Logistic Regression 65.1%, and the Classifier 91.2%. The ensemble model significantly improved prediction, reaching 96.62% accuracy. Hyperparameter tuning further improved accuracy but at the cost of slower performance. These findings support the hypothesis and highlight the trade-off between accuracy and computational efficiency in ML-based lung cancer prediction.

INTRODUCTION

Lung cancer is a disease that forms in a patient's lung and can go undetected until it is too late to treat it. It is the leading cause of cancerous deaths worldwide.

Your lungs are 2 sponge-like organs that allow you to inhale oxygen and distribute it to all your cells in your body. Each lung is divided into sections called lobes. Your right lung has 3 lobes and your left lung has 2. The left lung is smaller because the heart takes up more space on that side of the body. When you inhale (breathe in), the air enters your body and your lungs through the trachea, also known as the windpipe. The trachea divides into tubes called bronchi when it enters the lungs and divides even further into smaller bronchi. These divide into even smaller branches called bronchioles. At the end of all bronchioles are tiny air pockets, also known as alveoli, that contain the air you just inhaled. The alveoli absorb oxygen into your blood from the

inhaled air and remove carbon dioxide from the blood when you exhale (breathe out). Taking in oxygen and getting rid of carbon dioxide are the lung's main functions. This is how your lungs normally work. (1)

Lung cancer is one of the most common cancers in the United States and the world, with someone being diagnosed about every two and a half minutes. About 1 in 16 people have been diagnosed with lung cancer in their lifetime. In 2023, an estimated 238,340 people were diagnosed in only the U.S. In 2024, 234,580 people were diagnosed with this disease. Due to lung cancer, approximately 127,070 American lives are lost annually. Today, approximately 650,620 people in the U.S. have been diagnosed with lung cancer at some point in their lives. (2)

Lung cancer mainly occurs in older people. Most people diagnosed with lung cancer are 65 or older. Not many people are diagnosed with lung cancer under the age of 45. The average age of people when diagnosed with lung cancer is about 70. Lung cancer is by far the leading cause of cancer deaths in the US. Lung cancer itself takes about 1 in 5 of all cancer lives and more lives than the other three most common cancers combined. (3)

Lung cancer blocks the airways in which air flows through. This results in not enough air reaching the lungs, and therefore not enough oxygen supply for the rest of the body. (4) The main types of lung cancer are non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). About 80% - 85% of lung cancers are NSCLC. The main subtypes of NSCLC are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. All the subtypes of NSCLC are grouped because they all start from different types of cells in the lung. Their treatment and prognoses are often quite similar. About 10% - 15% of all lung cancers are SCLC. This type of lung cancer grows and spreads faster than NSCLC. In most people with SCLC, the cancer has already spread beyond the lungs at the time it is diagnosed. Since this cancer grows so quickly, it can be treated well with chemotherapy and radiation therapy. Unfortunately, for most people, the cancer will come back at some point. (1)

Smoking is the number one risk factor for lung cancer. In the United States, about 80% to 90% of all lung cancer deaths are a result of cigarette smoking. People who smoke cigarettes are 15 to 30 times more likely to be diagnosed with lung cancer or die from the disease than people who do not smoke. The more frequently a person smokes and the more time that a person smokes the more the risk increases. Radon is the second leading cause of lung cancer in the United States. Radon is a naturally occurring radioactive gas that forms in rocks, soil, and water. It cannot be seen, tasted, or smelled. When radon gets into homes or buildings through cracks or holes, it can get trapped and build up in the air inside. People who are exposed to high levels of radon over long periods can be diagnosed with lung cancer. (5) Figure 1 shows the difference between a healthy lung and a cancerous lung.

All in all, lung cancer is a very deadly disease, and it is also very common. It is most diagnosed when it is too late to do anything about it. This makes it extremely difficult to cure lung cancer. But what if we use a tool that can help us identify lung cancer in a person earlier? Using such tools can help detect lung cancer earlier, which leads to a higher chance of surviving the disease.

To solve this problem, our project mainly focuses on using the rapidly advancing technology called Artificial Intelligence (AI) to help diagnose people with lung cancer earlier. I hypothesize that AI can help diagnose lung cancer in its initial phase, which will make it much easier to cure patients with the disease. By the usage of different features, such as age, gender, if they smoke, if they have shortness of breath, etc., the model could use these features and many more to detect if a person is at risk of lung cancer during the initial stage itself. The model could also analyze histopathological images of patients to diagnose a patient. This dual-methodology could make it much easier for the disease to be cured, and it could save many lives.

RESULTS

In our project, there are three types of results - the training accuracy, the testing accuracy, and the weights. As the names suggest, the training accuracy is the accuracy the AI models get during the training phase. Similarly, the testing accuracy is the accuracy the models get during

the testing phase. The weights are all the coefficients multiplied by all the models. This helps the models achieve a higher accuracy. This is also part of the "Multiplicative Weight Update Method" that is used to combine multiple models to get a final prediction. Each of the four models in my project has its own training accuracy and testing accuracy. Then, using the Multiplicative Weight Update Method, the project has combined training accuracy, testing accuracy, and weights. Another method used was weight normalization, which made it so that all the weights added together would be equal to one. Figure 2 shows the training and testing accuracies of the models along with the weights. Figure 3 shows the relationship between the training and testing accuracies for all the models.

Another type of variable collected was the maximum depth and maximum iteration for the models. Some models require a parameter such as maximum depth and maximum iteration. These parameters determine some properties of a model and can affect the accuracy of the model. Therefore, it is crucial to pass numbers that are most effective for the model's accuracy. Using the hyperparameter tuning method, I was able to come up with values for the maximum depths for the classifier (CLF) model, the decision tree model, and the random forest model. The logistics regression model does not need the value. Figures 4, 5, and 6 show each model's training and testing accuracy for each maximum iteration or maximum depth. The highlighted rows show the highest testing and training accuracies.

From the data collected, we can see that the optimal maximum iteration for CLF is 80 (Figure 4), the optimal maximum depth for the decision tree is also 80 (Figure 5), and the optimal maximum depth for the random forest classifier is 90 (Figure 6). These numbers allow us to increase the accuracy of the models.

DISCUSSION

In conclusion, Artificial Intelligence, if used correctly, has the power to help save numerous lives by helping detect lung cancer in a patient in its early stage. This is showcased by our project which can identify lung cancer with 96.62% accuracy in its initial stage. This has the potential to

help save many lives. In the future, a question that could build on this idea could be whether artificial intelligence can use images and image classification to detect lung cancer. For example, can AI use medical scans such as MRI scans and clinical scans to detect lung cancer in its initial stage? Another question that could be answered is could patients and doctors use this tool to help detect lung cancer in its initial stage?

MATERIALS AND METHODS

When creating an AI model, we need to follow some steps. When creating an AI model using a CSV file, the first step is to load the data into Python in a format that can be easily manipulated. This is typically done using the "pandas" library, which allows for reading CSV files. By examining the data, you can identify missing values, inconsistencies, or non-numerical values that may need attention. Ensuring the dataset is clean and well-structured is critical to achieving accurate model performance. Our dataset was found on https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.

Next, it's important to convert all non-numerical values into a numerical format. Machine learning models require numerical data, so categorical or text-based features need to be changed into numerical equivalents. For example, the gender feature can be converted into 0 and 1, allowing the model to process the information effectively.

Once the data is properly formatted, the next step is feature selection. This involves identifying which features (columns) in the dataset are most relevant to the model's task. By removing irrelevant features, we can prevent the model from overfitting and improve the model's accuracy. Our project did not require this to be done, as all the features were relevant to the model.

After selecting the necessary features, the dataset needs to be split into training and testing sets. The training data is used to teach the model, while the testing data evaluates its performance on unseen information to the model. Our project uses the split ratio 66:33. Splitting the data into two

sets helps ensure the gets a new set of data each time to get an unbiased estimation of the performance of the model.

At this point, it's time to identify and select a machine-learning model suitable for the problem. Each model has its strengths depending on the data and the goals of the project, so it's important to experiment with multiple models. Our model uses four different types of models: classifier, decision tree, random forest, and logistic regression.

Once the model(s) is chosen, the next step is to train it using the training dataset. The model learns patterns and relationships in the data, adjusting its internal parameters through the training process. This is done by fitting the model using the "fit()" method in libraries like "scikit-learn".

The trained model is then evaluated using the test data, with metrics such as accuracy or precision used to assess its performance. A well-trained model should make accurate predictions on new, unseen data.

In our project, we used four different types of models: CLF, decision tree, random forest, and logistic regression.

Classification models, or CLF models, are a type of machine learning algorithm used to predict based on input features. They are specifically good at classifying classes for the input features. They work by learning from a labeled training dataset, where each example consists of input features and a corresponding target class label. The model identifies patterns in the data. Figure 7 shows an example diagram of how a CLF model works.

A decision tree model works by using a series of "if statements" to predict the target column. Each node in the tree represents a decision point. At each node that the input value arrives at, the decision is made about which node the input value will go next. The input value will then travel through the branches to the node that it had been assigned. At the new decision point that the input value arrives at, it will again repeat the same process. At the very end of the tree, the value will arrive at a leaf node, where the final prediction of the target feature will be made. Each leaf has a different prediction, and according to what the input value lands at, the leaf's prediction will be the prediction of the model. The tree-like structure that is formed in the model and making decisions at each node is the reason why the model is called the decision tree model. Decision trees are easy to interpret as the process is visually shown. Figure 8 shows an example diagram of how a decision tree model works.

A random forest is a learning method that improves the accuracy of decision trees. A random forest, as the name suggests, is a collection of multiple decision trees that collectively make the final prediction or decision. In a random forest, there are multiple decision trees built independently using different subsets of the data. Each tree will then make its prediction and the random forest will call the final decision by majority vote. In other words, the final prediction will be what most of the trees are voting for. Random forests are efficient because they can handle complex datasets with high accuracy. Figure 9 shows an example diagram of how a random forest classifier works.

Logistic regression is a method used for models where the outcome has two values. In our case, cancerous and non-cancerous. Unlike linear regression, where the model sees the input and output as consistent, logistic regression adds a little bit of logic and estimates the probability that a given input belongs to a particular category by using the logistics function. This classifies the input value in the given categories and the value is placed in the category with the higher estimation percentage. Figure 10 shows an example diagram of how a logistics regression model works.

Overfitting in AI models occurs when the training accuracy of a model is much higher than the testing accuracy. This means that the AI model is picking up on data that is irrelevant and not necessary for the classification. For example, in our model that classifies if a person has cancer, it will also be trained on the dates that the data was collected. This data is irrelevant for classification purposes and the model is being trained on extra data. An easy fix to this problem is removing all the irrelevant pieces of data that the model does not need. For example, the random forest classifier initially had the problem of overfitting. The training accuracy was 96% and the testing accuracy only came out to be 81%. This occurred when the random forest classifier's maximum depth was set to 5. Another example is when the decision tree model also was overfitting. The training accuracy was 96%, but the testing accuracy was only 84%. This also occurred when the decision tree model's maximum depth was set to 5.

The Multiplicative Weight Update is a technique used to combine multiple AI models to get a single output. This technique works by using weights. To start, all the models are assigned the same weight. Now, let's take an example. Let's say that in our data file, the first row in the target column says "cancerous" Let's also say that we have 4 types of models, Model A, B, C, and D, and currently, all the models have the same weight 10. For the first row, Model A predicts "not cancerous," Model B predicts "not cancerous," Model C predicts "cancerous," and Model D predicts "not cancerous." In this case, all the models have the same weightage, meaning that all of the votes are worth the same. In this cycle, we can see that Models A, B, and D did not predict the target column. This means that we trust these models a bit less than we used to before. So, we decrease the weight of the models. If we divide the weights by 2, Models A, B, and Ds' weights will be 5. Model C, however, predicted the target column correctly. Therefore, we trust this model a bit more and we multiply the weight of Model C by 2, resulting in it being 20. By following this cycle for the rest of the rows in the data file, we can find weights for all the models that will maximize the accuracy of the models.

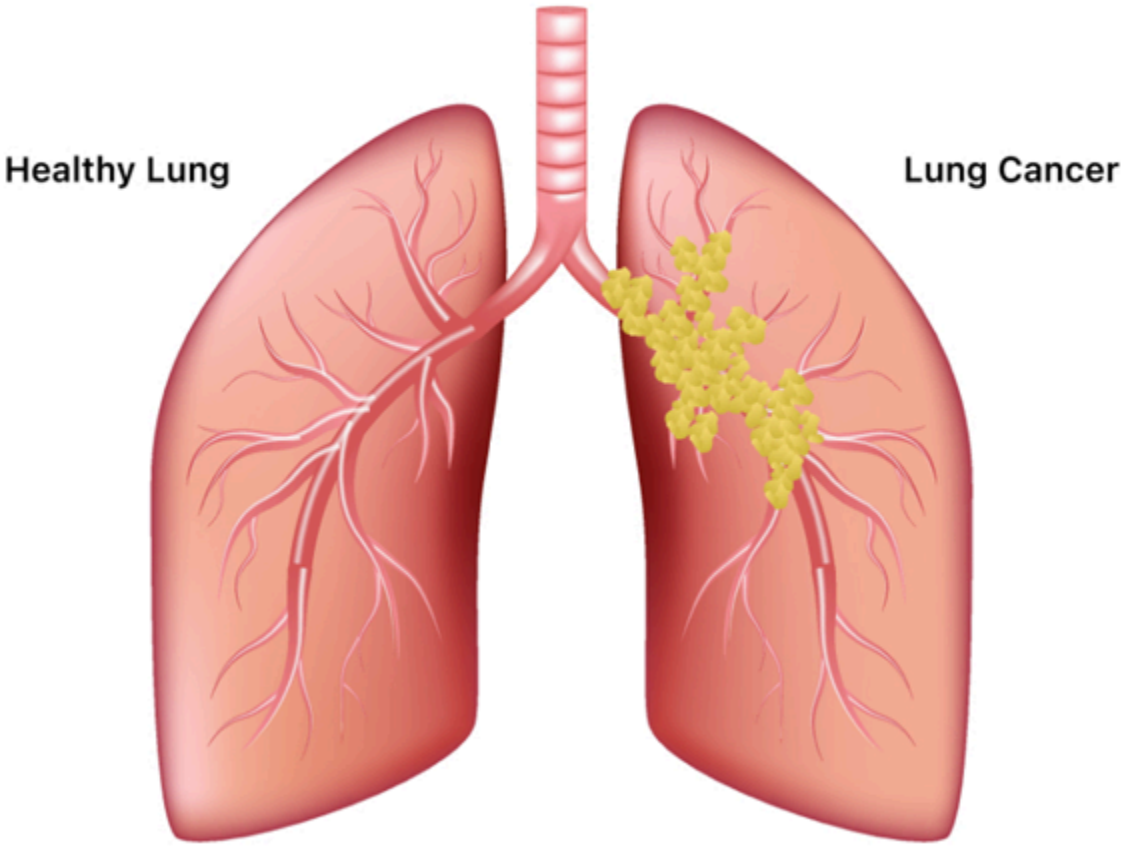
By the end of the training process, we will have weights suitable for all the models that are most suitable for them and the model. By using this method, it will help bring the accuracy of all the models up and give exactly one prediction.

Hyperparameter tuning is used to tune certain models' maximum depth and maximum iteration parameters. Some models, for example, a decision tree model, ask for a max depth input. This input means how many levels the decision tree will have. For a random forest classifier, it means the maximum depth of each decision tree in the forest. Tuning these numbers is crucial because it can greatly improve the model's accuracy. This method allows us to find a number that improves accuracy the most. By the end of the training process, we will have the maximum depth and iteration parameters that increase the accuracy of the models.

REFERENCES

1. "What Is Lung Cancer?: Types of Lung Cancer." *American Cancer Society*, 29 Jan. 2024, www.cancer.org/cancer/types/lung-cancer/about/what-is.html.
2. "Facts About Lung Cancer." *Lung Cancer Research Foundation*, www.lungcancerresearchfoundation.org/lung-cancer-facts/#:~:text=1%20IN%2016%20PEOPLE%20will,and%201%20in%2017%20women.&text=Approximately%20127%2C070%20AMERICAN%20LIVES%20are%20lost%20annually.&text=654%2C620%20PEOPLE%20IN%20THE%20U.S.,some%20point%20in%20their%20lives.
3. "Key Statistics for Lung Cancer." *American Cancer Society*, 29 Jan. 2024, www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html#:~:text=The%20American%20Cancer%20Society%27s%20estimates,men%20and%20118%2C270%20in%20women.
4. "Manage Shortness of Breath with Lung Cancer." *Johns Hopkins Medicine*, www.hopkinsmedicine.org/health/conditions-and-diseases/lung-cancer/manage-shortness-of-breath-with-lung-cancer#:~:text=Blocked%20airways%3A%20Lung%20tumors%20can,wall%2C%20called%20the%20pleural%20space.
5. "Lung Cancer Risk Factors." *The Centers for Disease Control and Prevention*, 15 October 2024, www.cdc.gov/lung-cancer/risk-factors/index.html#:~:text=Smoking-,Cigarette%20smoking%20is%20the%20number%20one%20risk%20factor%20for%20lung,of%20more%20than%207%2C000%20chemicals.

Figure 1



Example of a healthy lung and a cancerous lung

Figure 2: The Training Accuracy, Testing Accuracy, and Weights of all Models

Average and Weights of all Models	CLF	Decision Tree	Random Forest	Logistics Regression

Train	**91.3%**	**99.52%**	**100%**	**96.62%**
Test	**91.2%**	**92.2%**	**95.1%**	**65.1%**
Weights	**0.75964**	**0.18991**	**0.04748**	**0.00297**

Figure 3: The Training and Testing Accuracies of All Models

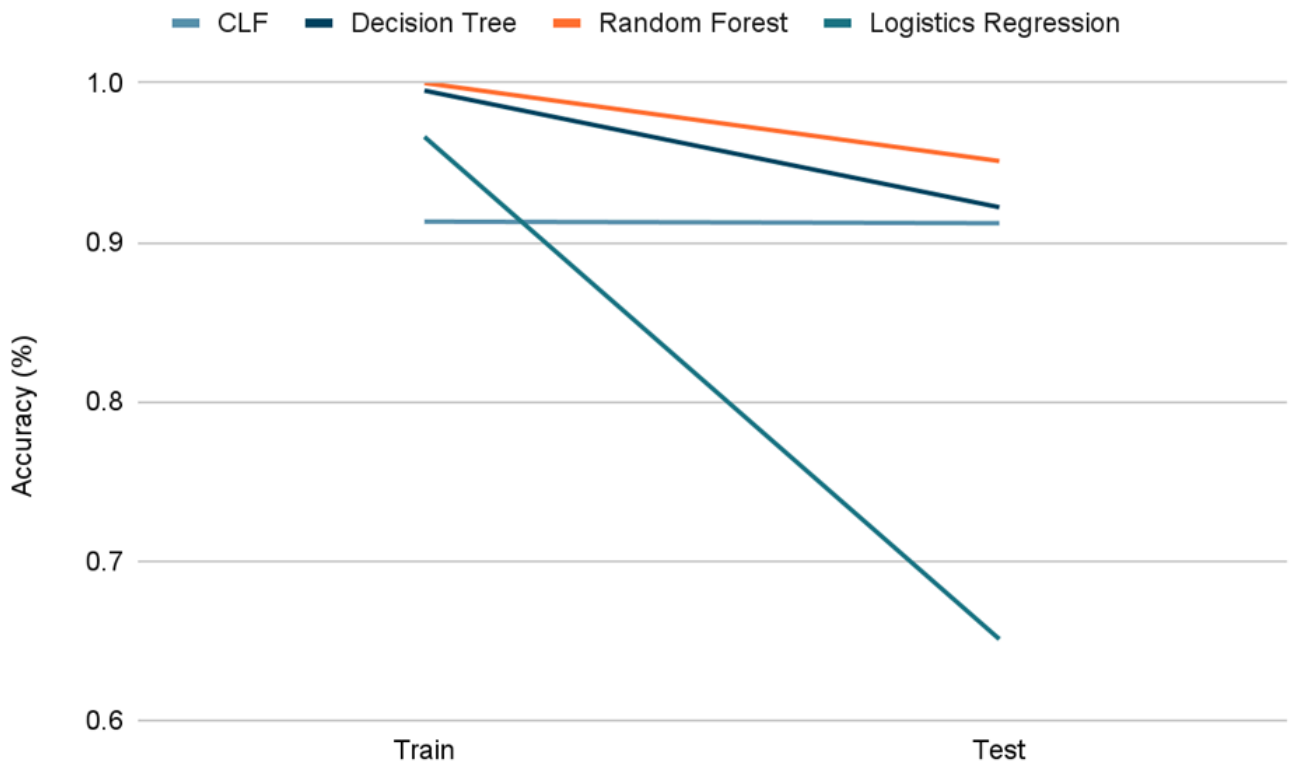


Figure 4: CLF Maximum Iteration

Maximum Iteration	CLF Test	CLF Train	CLF Average (testing-training)
5	0.79	0.91	0.85
10	0.81	0.90	0.85
15	0.85	0.88	0.86
20	0.86	0.87	0.87
25	0.85	0.88	0.86
30	0.87	0.87	0.87
35	0.86	0.87	0.87
40	0.85	0.88	0.86
45	0.88	0.86	0.87
50	0.88	0.86	0.87
55	0.87	0.87	0.87
60	0.91	0.85	0.88

65	0.85	0.88	0.86
70	0.89	0.86	0.87
75	0.90	0.85	0.88
80	0.91	0.85	0.88
85	0.86	0.87	0.87
90	0.91	0.85	0.88
95	0.91	0.85	0.88
100	0.89	0.86	0.87

Figure 5: Decision Tree

Maximum Depth	Decision Tree Test	Decision Tree Train	Decision Tree Average
5	0.84	0.96	0.90

10	0.90	0.99	0.94
15	0.90	1	0.95
20	0.86	0.99	0.92
25	0.84	1	0.92
30	0.90	0.99	0.94
35	0.89	1	0.94
40	0.88	1	0.94
45	0.90	0.99	0.94
50	0.87	0.99	0.93
55	0.87	1	0.93
60	0.88	1	0.94
65	0.85	1	0.92
70	0.86	1	0.93

75	0.86	1	0.93
80	0.92	0.99	0.95
85	0.84	1	0.92
90	0.86	1	0.93
95	0.90	1	0.95
100	0.87	1	0.93

Figure 6: Random Forest Classifier

Maximum Depth	Random Forest Test	Random Forest Train	Random Forest Average
5	0.81	0.96	0.88
10	0.90	0.99	0.94
15	0.93	1	0.96

20	0.89	0.99	0.94
25	0.94	1	0.97
30	0.90	0.99	0.94
35	0.92	1	0.96
40	0.89	1	0.94
45	0.90	0.99	0.94
50	0.91	0.99	0.95
55	0.90	1	0.95
60	0.94	1	0.97
65	0.92	1	0.96
70	0.90	1	0.95
75	0.90	1	0.95
80	0.95	0.99	0.97

85	0.89	1	0.94
90	0.95	1	0.97
95	0.94	1	0.97
100	0.93	1	0.96

Figure 7

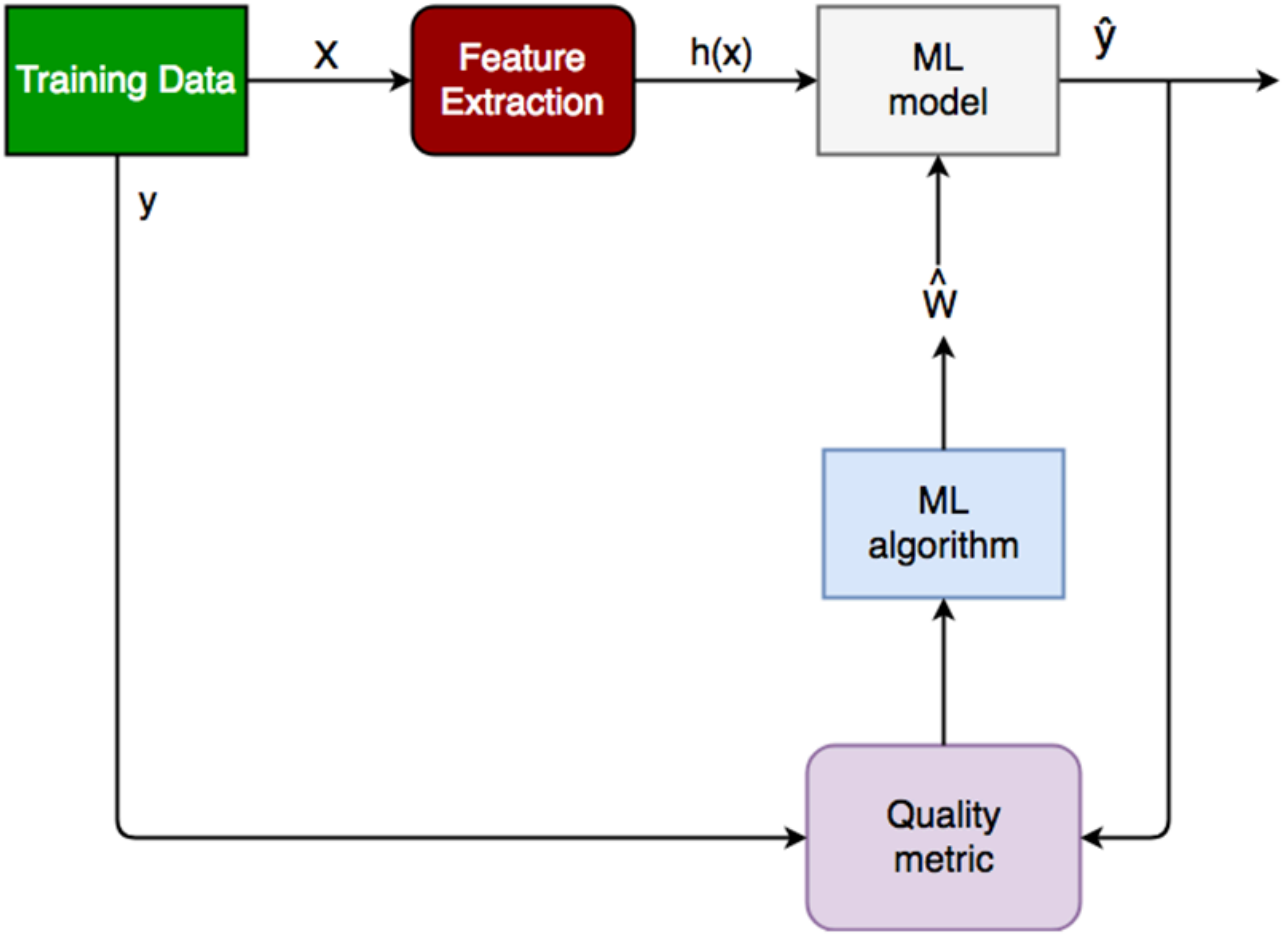

Training Data
X
Feature Extraction
h(x)
ML model
ŷ
y
Ŵ
ML algorithm
Quality metric

Figure 8

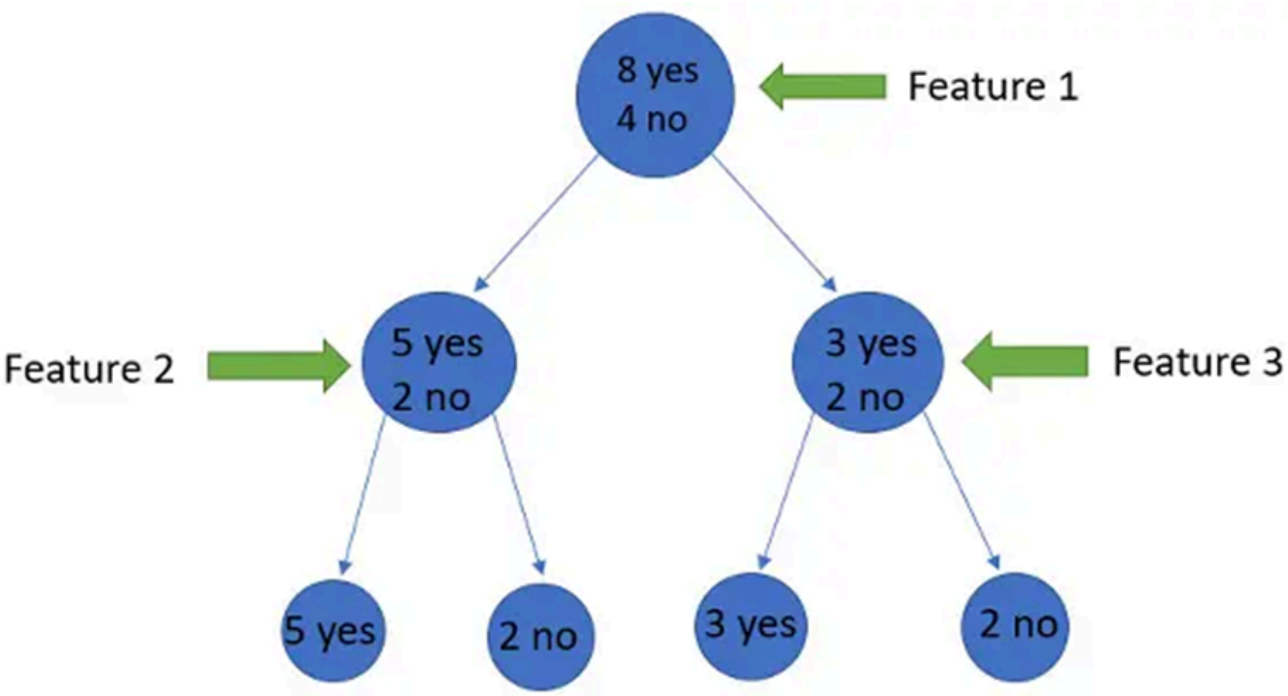

8 yes
4 no
Feature 1
Feature 2
5 yes
2 no
3 yes
2 no
Feature 3
5 yes
2 no
3 yes
2 no

Figure 9

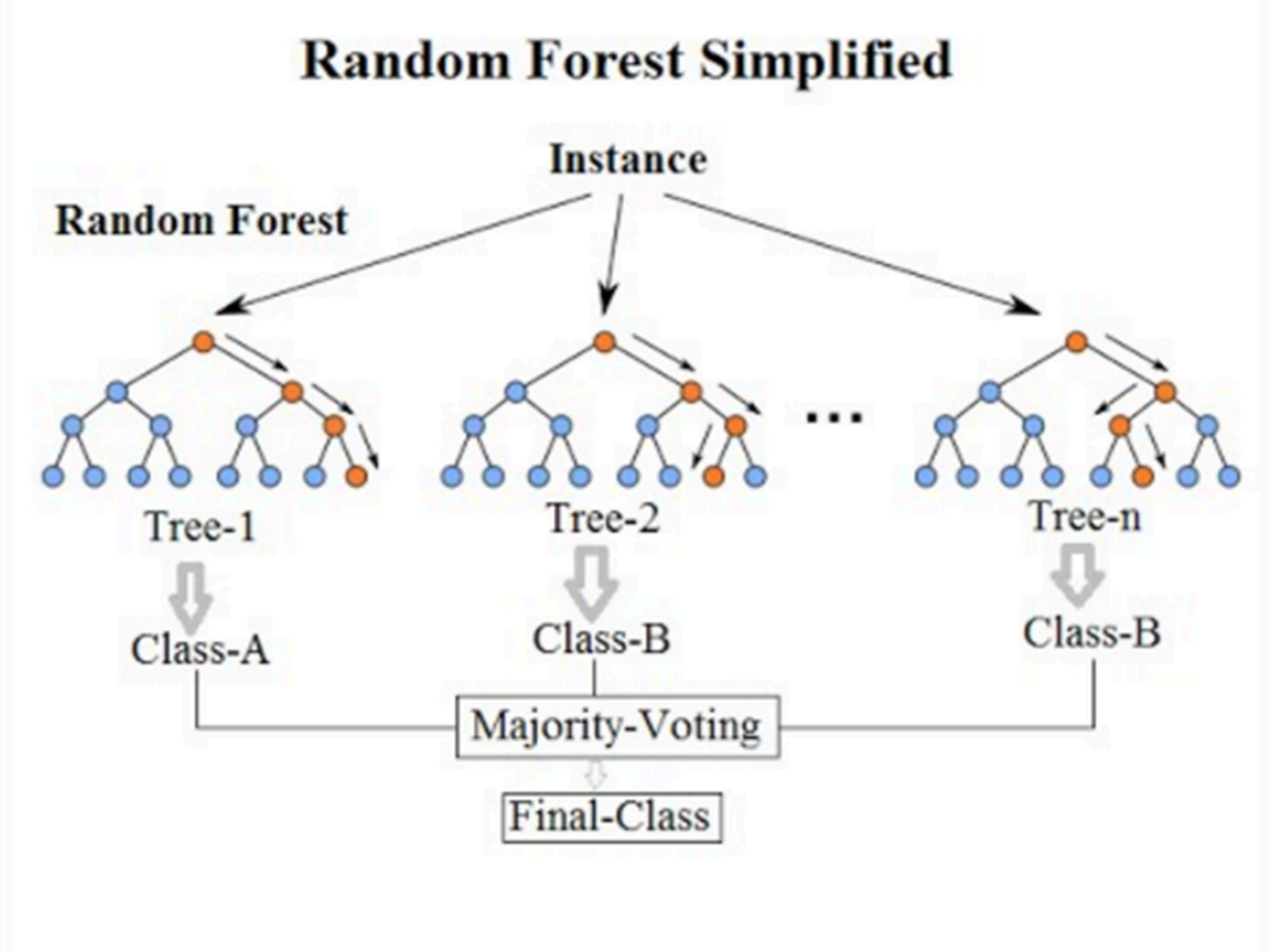
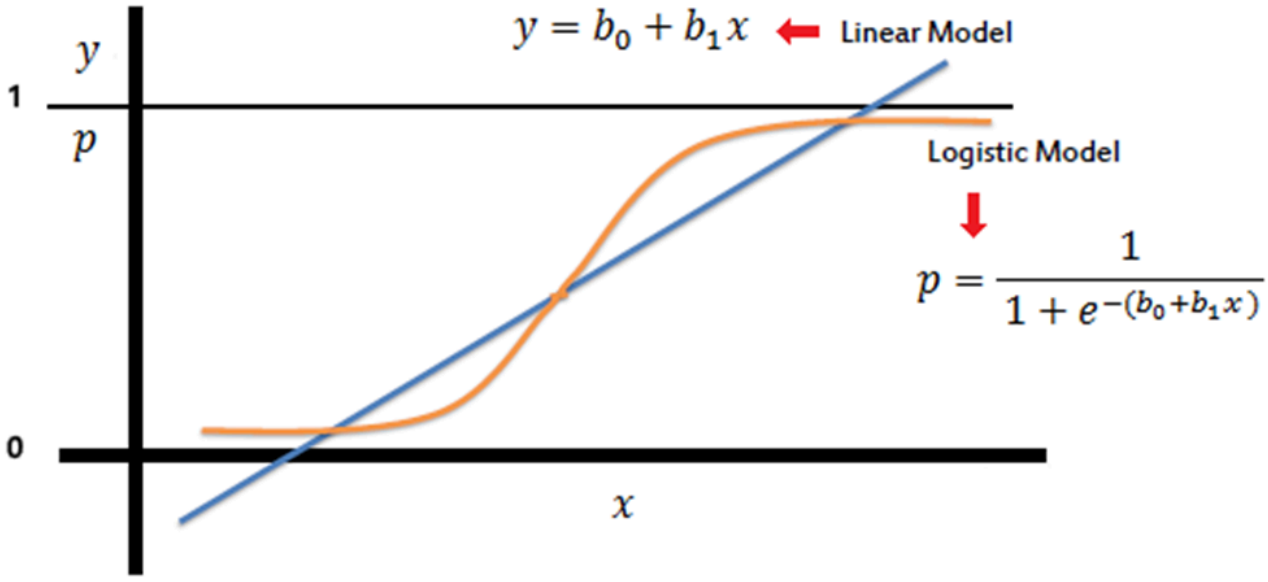

Random Forest Simplified
Instance
Random Forest
Tree-1
Tree-2
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class

Figure 10



APPENDIX

Link to code:

https://github.com/omg29/Utilizing-the-Power-of-Artificial-Intelligence-for-Early-Detection-of-Lung-Cancer