

A MACHINE LEARNING PIPELINE WITH SWITCHING ALGORITHMS TO
PREDICT LUNG CANCER AND IDENTIFY TOP FEATURES

by

Anika Tasnim

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Computer Science

at

The University of Wisconsin-Milwaukee

August 2021

ABSTRACT

A MACHINE LEARNING PIPELINE WITH SWITCHING ALGORITHMS TO PREDICT LUNG CANCER AND IDENTIFY TOP FEATURES

by

Anika Tasnim

The University of Wisconsin-Milwaukee, 2021

Under the Supervision of Professor Jake Luo and Professor Tian Zhao

Lung cancer is the leading cause of cancer-related death around the world. Early detection is a critical factor for its effective treatment. To facilitate early-stage prediction, a Machine Learning (ML) pipeline has been built that uses inpatient admission data to train four ML models. The data is dynamically loaded into a database, cleaned, and passed through the SelectKBest selector to identify the top features influencing the prognosis, which are then fed into the pipeline and fitted to the ML models to make the forecast. Among the models used, Decision Tree provides the highest accuracy (97.09%), followed by Random Forest (94.07%). MLP and Logistic Regression reach an accuracy of 84.58% and 77.65% respectively. Some of the top 50 features include chronic obstructive pulmonary disease, pleural effusion, secondary and unspecified malignant neoplasm of intrathoracic lymph nodes, syndrome of inappropriate secretion of antidiuretic hormone, and neoplasm-related acute, chronic pain.

© Copyright by Anika Tasnim, 2021
All Rights Reserved

TABLE OF CONTENTS

List of Figures	v
List of Tables	vi
1. Introduction	1
2. Background Study	3
3. Data Description	5
3.1. Data source	5
3.2. Pre-processing	7
3.2.1. Loading into Database	7
3.2.2. Handling invalid and missing values	8
3.2.3. Indexing ICD codes	11
3.3. Feature Engineering	12
4. Machine Learning and Deep Learning Models	15
4.1. Decision Tree	15
4.2. Logistic regression	16
4.3. Random Forest	16
4.4. Multi-Layer Perceptron	17
5. Machine Learning pipeline	19
6. Results	21
6.1. Important Features	21
6.1.1. Age	21
6.1.2. Gender	22
6.1.3. Race	22
6.1.4. Disease conditions	24
6.2. Comparison among ML models	26
6.2.1. Accuracy	26
6.2.2. Training Time	28
6.2.3. Feature Saturation Curve	29
6.3. Limitations and Future Work	32
7. Conclusion	33
References	34

LIST OF FIGURES

Figure 1: Annual number of new lung cancer cases, USA, 1999-2017.....	2
Figure 2: ICD-10-CM codes and disease conditions.....	6
Figure 3: File specification of Core File data.....	7
Figure 4: Dataset before data cleaning.....	9
Figure 5: Dataset after data cleaning.....	10
Figure 6: ICD dictionary.....	11
Figure 7: Decision Tree architecture.....	15
Figure 8: Logistic Regression.....	16
Figure 9: Random Forest architecture.....	17
Figure 10: MLP model architecture.....	18
Figure 11: Machine Learning Pipeline.....	19
Figure 12: Lung cancer patient's age.....	21
Figure 13: Lung cancer patient's gender.....	22
Figure 14: Lung cancer patient's race.....	23
Figure 15: Lung cancer case percentage among racial groups.....	23
Figure 16: Analysis of disease conditions.....	24
Figure 17: Best accuracy among models.....	27
Figure 18: Training time among models.....	28
Figure 19: Decision Tree saturation curve.....	29
Figure 20: Logistic Regression saturation curve.....	29
Figure 21: Random Forest saturation curve.....	30
Figure 22: MLP saturation curve.....	30
Figure 23: Saturation curve comparison.....	31

LIST OF TABLES

Table 1: Column names and expected values.....	8
Table 2: Top 50 features identifying lung cancer.....	12
Table 3: Comparison among the Machine Learning models' accuracy.....	26
Table 4: Comparison among the Machine Learning models' training time.....	28

1. Introduction

The human body is made up of trillions of cells, the growth of which is generally controlled by genetic factors. However, when there is uncontrolled abnormal cell growth, we have what is commonly referred to as cancerous growth. Lung cancer is a form of cancer in which such uncontrolled cell growth starts in the lungs. It is the 2nd most diagnosed cancer as well as the leading cause of cancer mortality for both men and women in America [1]. It is responsible for about 25% of all cancer-related deaths, killing more people than colon, prostate, and breast cancers combined [2].

While the symptoms of lung cancer vary widely from person to person, some of the most common symptoms include cough (which worsens over time and does not go away), hoarseness, chronic chest pain, shortness of breath, repeated lung infections including bronchitis or pneumonia and hemoptysis (the coughing up of blood). Although symptoms may start developing from the early stages of cancerous cell formation, many people experience them only when cancer reaches an advanced stage [3].

There are two major types of lung cancer: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung cancer (NSCLC). 80% to 85% of the lung cancer cases are identified as NSCLC [4]. NSCLC treatment uses surgery, radiation therapy, chemotherapy, targeted therapy, or a combination of the four. On the other hand, radiation therapy and chemotherapy are widely used for SCLC treatment. Over the last few decades, the percentage of new lung cancer cases among the total population in the USA has decreased. Nevertheless, **Figure 1** suggests that the number of new lung cancer cases detected in America during the years 1999 to 2017 is still considerably high [5].

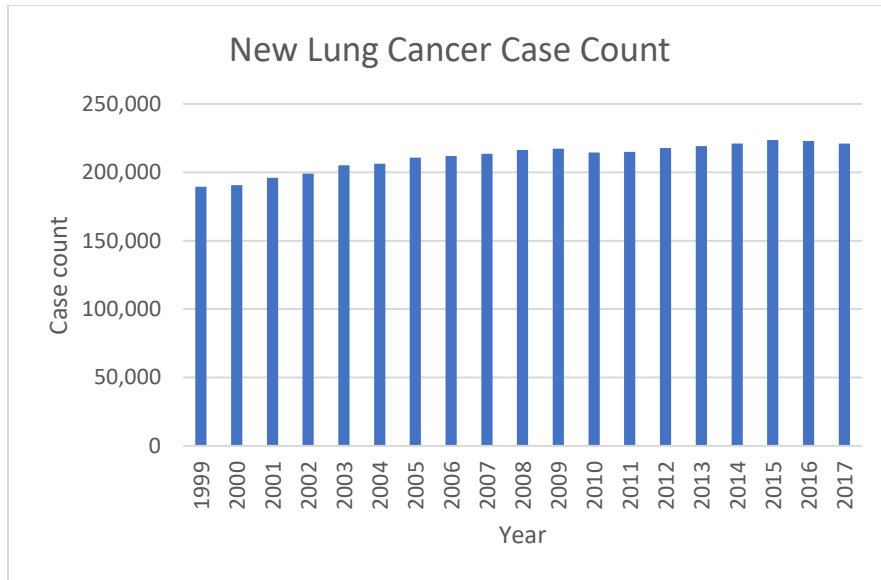


Figure 1: Annual number of new lung cancer cases, USA, 1999-2017 [5].

The dissertation titled “A Machine Learning pipeline with switching algorithms to predict lung cancer and identify top features” applies Machine Learning and Deep Learning techniques on healthcare data acquired from Nationwide Inpatient Sample (NIS) Dataset [6]. It uses models such as Decision Tree, Logistic Regression, Random Forest, and Multi-Layer Perceptron to predict lung cancer, and identify the top features contributing to the prognosis. The project is implemented in Python using the Scikit-Learn package. It utilizes the Jupyter Notebook platform as it provides an interactive development environment.

2. Background Study

Zubi and Saad (2014) proposed a Computer-Aided Diagnosis (CAD) system which consisted of two fundamental steps: feature extraction from chest x-rays and employment of a Backward Propagating Neural Network (NN) classifier [7]. The classification accuracy for normal, benign, and malignant images were 100%, 95%, and 85% respectively.

Kohad and Ahire (2015) aimed to detect malignant nodules from lung computerized tomography (CT) scans and proposed a CAD system with 4 steps: preprocessing, feature extraction, feature selection, and classification [8]. To improve accuracy, the Ant colony optimization algorithm was utilized as a feature selection strategy. Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms were used to categorize normal and abnormal lung images. Between the algorithms, SVM had a 93.2% accuracy, whereas ANN had a 98.4% accuracy.

Hussein et al. (2019) used supervised and unsupervised ML techniques on a lung nodules dataset to characterize lung tumors [9]. For the supervised approach, a 3D Convolutional Neural Network (CNN) was used which provided an accuracy of 78.06%. The unsupervised approach achieved a 91.26% accuracy categorizing benign and malignant data using the SVM algorithm.

Ganggayah et al. (2019) used a breast cancer dataset collected from hospitals containing 8066 records and 23 features [10]. The natural language data was fed to ML models such as Decision Tree, Random Forest, Neural Networks, Extreme Boost, Logistic Regression, and SVM to predict the survival rate and top features of prognosis. Among these models, Decision Tree provided the lowest accuracy of 79.8% while Random Forest provided the highest accuracy of

82.7%. Some of the top features influencing the prediction were cancer stage, tumor size, and the number of axillary lymph nodes removed.

A significant number of existing lung cancer predictive models work with CT scan images and use image processing as well as ML techniques to estimate the forecast. However, this dissertation focuses on natural language hospital inpatient admission data to predict lung cancer as inpatient data is more available, and accessible than CT scan images. Since the model is not dependent on any specific type of image, the disease condition may be switched from lung cancer to any other disease condition. The use of ML pipeline also enables changing ML models as well as parameters such as the size of selected data, the ratio of training and testing dataset sizes, and feature numbers without changing the main code.

ML models such as Decision Tree, Logistic Regression, Random Forest, and MLP are employed as they are compatible with natural language data, and there exists a wide range of resources to provide ease of implementation. Models such as Linear Regression and Naive Bayes are not used since Linear Regression works with continuous data and Naive Bayes is known to be a bad estimator. Since Naive Bayes requires the predictors to be independent, the probability calculations are known to be less accurate [11]. Deep Learning models such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) were avoided as they would require more computational resources than MLP.

3. Data Description

3.1. Data Source

National (Nationwide) Inpatient Sample (NIS) database is a healthcare database that contains information about hospital inpatient admissions. Since 1988 the NIS database has been storing records for more than 7 million hospital stays each year. Sponsored by Agency for Healthcare Research and Quality (AHRQ) and developed by Healthcare Cost and Utilization Project (HCUP), the NIS database is one of the largest healthcare databases.

This project uses the 2016 NIS dataset where over 7 million (7,135,090) records are included. Each entry consists of a total of 98 data elements including clinical data such as procedures, treatment types, diagnosis categories, diagnosis codes as well as non-clinical data such as patient demographics, admission date, total cost, zip code, hospital id, duration of stay. Patient demographics also include information such as age, sex, and race.

One of the most important clinical elements used in this project is the set of 30 diagnosis codes. These codes are represented in ICD-10-CM format, which is elaborated to International Classification of Diseases, 10th revision, Clinical Modification. However, the database holds some ICD-9 format codes as well. These medical classification lists have been designed by the World Health Organization (WHO) [12]. ICD codes have a specific format, and they provide a unique code for all possible disease conditions, signs and symptoms, adverse observations, complaints, and external causes of death. Hence, health care personnel, insurance providers, and other services use these codes as a standard to identify health conditions.

ICD-10 codes include 71,924 procedure codes (ICD-10-PCS) and 69,823 diagnosis codes (ICD-10-CM). ICD-10-CM codes are 3 to 7 characters long, where the 1st character is an

alphabet, the 2nd character is a number, and characters 3 to 7 may be alphanumeric [12]. **Figure 2** depicts some disease conditions along with their ICD-10-CM codes.

A000	Cholera due to <i>Vibrio cholerae</i> 01, biovar cholerae
A001	Cholera due to <i>Vibrio cholerae</i> 01, biovar eltor
A009	Cholera, unspecified
A0100	Typhoid fever, unspecified
A0101	Typhoid meningitis
A0102	Typhoid fever with heart involvement
A0103	Typhoid pneumonia
A0104	Typhoid arthritis
A0105	Typhoid osteomyelitis
A0109	Typhoid fever with other complications
A011	Paratyphoid fever A
A012	Paratyphoid fever B
A013	Paratyphoid fever C
A014	Paratyphoid fever, unspecified
A020	Salmonella enteritis
A021	Salmonella sepsis
A0220	Localized salmonella infection, unspecified

Figure 2: ICD-10-CM codes and disease conditions.

Once the data is collected, it is dynamically loaded into a database. Then the data goes through some preprocessing and feature engineering techniques to identify the top features contributing to the prognosis. Afterward, the ML pipeline is built, and the top features are passed to the pipeline. Using a custom transformer, the data is transformed and then fitted to the appropriate ML model to calculate the predictions. After the results are generated, some of the top features and the performance of the models are analyzed. Finally, some saturation curves are compared, and a conclusion is drawn from the results.

3.2. Data Pre-processing

3.2.1. Loading into Database

The 2016 dataset has a size of 15GB and is split into 3 ASCII files: “Core File”, “Hospital Weights File” and “Severity Measures File”. **Figure 3** reveals the “File Specification” file describing how data is represented in the “Core File”. It lists the database name, discharge year, file name, data element number, data element name, the starting and ending position of each data element in ASCII file, data element type, and data element label.

```
Data Set Name: NIS_2016_CORE
Number of Observations: 7135090
Total Record Length: 497
Total Number of Data Elements: 98

Columns  Description
=====  =====
 1-  3   Database name
 5-  8   Discharge year of data
10- 35   File name
37- 39   Data element number
41- 69   Data element name
71- 73   Starting column of data element in ASCII file
75- 77   Ending column of data element in ASCII file
79- 79   Non-zero number of digits after decimal point for numeric data element
81- 84   Data element type (Num=numeric; Char=character)
86-185  Data element label

NIS 2016 NIS_2016_Core      1 AGE                1  3   Num  Age in years at admission
NIS 2016 NIS_2016_Core      2 AGE_NEONATE        4  5   Num  Neonatal age (first 28 days after birth) indicator
NIS 2016 NIS_2016_Core      3 AMONTH              6  7   Num  Admission month
NIS 2016 NIS_2016_Core      4 A WEEKEND           8  9   Num  Admission day is a weekend
NIS 2016 NIS_2016_Core      5 DIED                10 11  Num  Died during hospitalization
NIS 2016 NIS_2016_Core      6 DISCWT              12 22 7  Num  NIS discharge weight
NIS 2016 NIS_2016_Core      7 DISPUNIFORM         23 24  Num  Disposition of patient (uniform)
NIS 2016 NIS_2016_Core      8 DQTR                25 26  Num  Discharge quarter
NIS 2016 NIS_2016_Core      9 DRG                 27 29  Num  DRG in effect on discharge date
NIS 2016 NIS_2016_Core     10 DRGVER              30 31  Num  DRG grouper version used on discharge date
NIS 2016 NIS_2016_Core     11 DRG_NoPOA           32 34  Num  DRG in use on discharge date, calculated without POA
NIS 2016 NIS_2016_Core     12 DXVER               35 36  Num  Diagnosis Version
NIS 2016 NIS_2016_Core     13 ELECTIVE            37 38  Num  Elective versus non-elective admission
NIS 2016 NIS_2016_Core     14 FEMALE              39 40  Num  Indicator of sex
NIS 2016 NIS_2016_Core     15 HCUP_ED             41 43  Num  HCUP Emergency Department service indicator
NIS 2016 NIS_2016_Core     16 HOSP_DIVISION       44 45  Num  Census Division of hospital
NIS 2016 NIS_2016_Core     17 HOSP_NIS            46 50  Num  NIS hospital number
NIS 2016 NIS_2016_Core     18 I10_DX1             51 57  Char  ICD-10-CM Diagnosis 1
NIS 2016 NIS_2016_Core     19 I10_DX2             58 64  Char  ICD-10-CM Diagnosis 2
NIS 2016 NIS_2016_Core     20 I10_DX3             65 71  Char  ICD-10-CM Diagnosis 3
NIS 2016 NIS_2016_Core     21 I10_DX4             72 78  Char  ICD-10-CM Diagnosis 4
NIS 2016 NIS_2016_Core     22 I10_DX5            79 85  Char  ICD-10-CM Diagnosis 5
```

Figure 3: File specification of Core File data.

A PostgreSQL database named “NIS_2016_Core” is created using a dataset parsed from the “Core File” following the information from “File Specification”. The parsed data is then dynamically loaded into the database.

3.2.2. Handling invalid and missing values

Each column of the “NIS_2016_Core” database represents a specific type of data and has its definition of what constitutes valid data. **Table 1** provides some of the columns and their consecutive definitions of valid data. The database holds a large amount of invalid and missing values. To produce good results using the ML models these invalid and missing values need to be ignored or transformed into valid data. This process is known as data cleaning. The most time-consuming part of using ML data is cleaning the dataset or developing an error-free dataset. Data cleaning includes procedures such as filling in missing information, deleting rows, and lowering data size [13].

Type of Data Element	HCUP Name	Coding Notes
Admission information		
Admission day	AWEEKEND	Admission on weekend: (0) admission on Monday-Friday, (1) admission on Saturday-Sunday
Admission month	AMONTH	Admission month coded from (1) January to (12) December
Transferred into hospital	TRAN_IN	Transfer In Indicator: (0) not a transfer, (1) transferred in from a different acute care hospital [ATYPE NE 4 & (ASOURCE=2 or POO=4)], (2) transferred in from another type of health facility [ATYPE NE 4 & (ASOURCE=3 or POO=5, 6)]
Indicator of emergency department service	HCUP_ED	Indicator that discharge record includes evidence of emergency department (ED) services: (0) Record does not meet any HCUP Emergency Department criteria, (1) Emergency Department revenue code on record, (2) Positive Emergency Department charge (when revenue center codes are not available), (3) Emergency Department CPT procedure code on record, (4) Admission source of ED, (5) State-defined ED record; no ED charges available
Admission type	ELECTIVE	Indicates elective admission: (1) elective, (0) non-elective admission
Patient demographic and location information		
Age at admission	AGE	Age in years coded 0-124 years
	AGE_NEONATE	Neonatal age (first 28 days after birth) indicator: (0) non-neonatal age (1) neonatal age
Sex of patient	FEMALE	Indicates gender for NIS beginning in 1998: (0) male, (1) female
Race of patient	RACE	Race, uniform coding: (1) white, (2) black, (3) Hispanic, (4) Asian or Pacific Islander, (5) Native American, (6) other (For 2016, Race contains missing values on about 5 percent of the records.)
Location of patient's residence	PL_NCHS	Patient Location: NCHS Urban-Rural Code. This is a six-category urban-rural classification scheme for U.S. counties: (1) "Central" counties of metro areas of >=1 million population, (2) "Fringe" counties of metro areas of >=1 million population, (3) Counties in metro areas of 250,000-999,999 population, (4) Counties in metro areas of 50,000-249,999 population, (5) Micropolitan counties, (6) Not metropolitan or micropolitan counties

Table 1: Column names and expected values [6].

The non-ICD columns are expected to have only positive numeric values. The missing, incomplete, invalid, and negative values are replaced with 0. **Figure 4** shows a part of the original dataset before handling missing values and **Figure 5** shows the same dataset after handling invalid and missing values.

Dataset before clean up										
	age	age_neonate	amonth	aweekend	died	discwt	dispuniform	dqtr	drg	
0	75	-9	10	0	0	5.0000303		1	4	175
1	80	-9	1	0	0	5.0000303		5	1	189
2	81	-9	9	0	0	5.0000303		5	3	456
3	71	-9	9	0	0	5.0000303		6	4	181
4	67	-9	10	0	0	5.0000303		6	4	181
5	58	-9	3	1	0	5.0000303		6	1	281
6	75	-9	3	0	0	5.0000303		5	1	177
7	64	-9	5	0	0	5.0000303		6	2	252
8	77	-9	8	0	0	5.0000303		5	3	641
9	82	-9	10	0	0	5.0000303		6	4	682
10	58	-9	4	1	0	5.0000303		1	2	191
11	60	-9	6	0	0	5.0000303		5	2	470
12	50	-9	9	0	0	5.0000303		1	3	603
13	40	-9	8	0	0	5.0000303		7	3	638
14	83	-9	6	0	0	5.0000303		6	3	293
15	90	-9	6	0	0	5.0000303		5	2	690
16	65	-9	1	0	0	5.0000303		5	1	189
17	25	-9	12	0	0	5.0000303		1	4	203
18	53	-9	10	0	0	5.0000303		1	4	281
19	73	-9	3	0	0	5.0000303		1	1	419

Figure 4: Dataset before data cleaning.

Dataset after clean up

	age	age_neonate	amonth	aweekend	died	discwt	dispuniform	dqtr	drg
0	75	0	10	0	0	5.0000303	1	4	175
1	80	0	1	0	0	5.0000303	5	1	189
2	81	0	9	0	0	5.0000303	5	3	456
3	71	0	9	0	0	5.0000303	6	4	181
4	67	0	10	0	0	5.0000303	6	4	181
5	58	0	3	1	0	5.0000303	6	1	281
6	75	0	3	0	0	5.0000303	5	1	177
7	64	0	5	0	0	5.0000303	6	2	252
8	77	0	8	0	0	5.0000303	5	3	641
9	82	0	10	0	0	5.0000303	6	4	682
10	58	0	4	1	0	5.0000303	1	2	191
11	60	0	6	0	0	5.0000303	5	2	470
12	50	0	9	0	0	5.0000303	1	3	603
13	40	0	8	0	0	5.0000303	7	3	638
14	83	0	6	0	0	5.0000303	6	3	293
15	90	0	6	0	0	5.0000303	5	2	690
16	65	0	1	0	0	5.0000303	5	1	189
17	25	0	12	0	0	5.0000303	1	4	203
18	53	0	10	0	0	5.0000303	1	4	281
19	73	0	3	0	0	5.0000303	1	1	419

Figure 5: Dataset after data cleaning.

Columns such as “HOSP_NIS” (NIS hospital number) and “HOSP_DIVISION” (Census Division of the hospital) have a direct correlation with the identification of lung cancer. On the other hand, the “YEAR” column holding the same value for all patients has no impact on the identification. Hence, these columns are dropped.

3.2.3. Indexing ICD codes

30 columns are storing ICD-10-CM codes in the database. These columns may hold up to 69,823 possible codes. At each iteration, a subset of the dataset is randomly selected and fed to the model. The selected subset holds a portion of the possible codes. The ICD codes present in the subset are inserted into a sorted set. Each of the alphanumeric ICD codes is then indexed to consecutive integers and an ICD dictionary is built. The ICD code for lung cancer is C34 which is not inserted into the dictionary. **Figure 6** denotes a part of the ICD dictionary. There exists a small size of numeric ICD codes which are represented in ICD-9 format.

ICD : Map	ICD : Map	ICD : Map
'000' : 1,	'023' : 11,	'083' : 21,
'001' : 2,	'029' : 12,	'084' : 22,
'002' : 3,	'03' : 13,	'090' : 23,
'010' : 4,	'030' : 14,	'091' : 24,
'011' : 5,	'031' : 15,	'092' : 25,
'012' : 6,	'032' : 16,	'099' : 26,
'019' : 7,	'07' : 17,	'A0100': 27,
'02' : 8,	'080' : 18,	'A020' : 28,
'020' : 9,	'081' : 19,	'A039' : 29,
'021' : 10,	'082' : 20,	'A040' : 30,

Figure 6: ICD dictionary.

While feeding the data to the model each of the ICD codes is treated as a distinct feature. So, the initial 30 columns transform into X number of columns where X represents the number of unique ICD codes present in the selected dataset. These columns store binary values (0/1) depending on if the patient has any record of a specific ICD code's disease condition.

3.3. Feature Engineering

After generating the ICD dictionary, around 70,000 unique ICD codes may be extracted. Each of the codes needs to be treated as a feature. However, many of these features may be irrelevant to predicting lung cancer. For example, someone may have a history of breaking a leg that has no relation to lung cancer. So, the relevant features need to be identified.

Selecting important features provides advantages such as reduced overfitting (less redundant data results in the lower possibility of making decisions based upon redundant data or noise), improved accuracy (fewer misleading data enhances modeling accuracy), and reduced training time (a lesser amount of data reduces algorithm complexity and helps the models train faster) [14].

The **SelectKBest** selector is used to identify the top features that affect the prediction. Each model is run with different feature numbers. Initially, the model is trained with 5 features and the feature number is gradually increased to up to 50. **Table 2** lists the top 50 features impacting lung cancer prognosis arranged in descending order of importance. Some of these top features are analyzed in section **6.1 Important Features**.

No.	Feature	Feature Definition
1	HCUP_ED	HCUP Emergency Department service indicator (0) Record does not meet any HCUP Emergency Department criteria, (1) Emergency Department revenue code on the record, (2) Positive Emergency Department charge (when revenue center codes are not available), (3) Emergency Department CPT procedure code on record, (4) Admission source of ED, (5) State-defined ED record; no ED charges available
2	RACE	Race (1) white, (2) black, (3) Hispanic, (4) Asian or Pacific Islander, (5) Native American, (6) other

3	MDC_NOPOA	Major Diagnostic Category in use on the discharge date, calculated without Present on Admission indicators
4	DRGVER	Diagnosis-Related Group version used on the discharge date
5	DQTR	Discharge quarter (1) 1 st quarter, Jan - Mar, (2) 2 nd quarter, Apr - Jun, (3) 3 rd quarter, Jul - Sep, (4) 4 th quarter, Oct - Dec
6	ZIPINC_QRTL	Median household income national quartile for patient (1) \$1 - \$42,999; (2) \$43,000 - \$53,999; (3) \$54,000 - 70,999; and (4) \$71,000 or more.
7	090	Congenital syphilis
8	DISCWT	NIS discharge weight
9	LOS	Length of stay
10	MDC	Major Diagnostic Category in effect on the discharge date
11	C7931	Secondary malignant neoplasm of brain
12	C7951	Secondary malignant neoplasm of bone
13	011	Pulmonary tuberculosis
14	012	Other respiratory tuberculosis
15	091	Early syphilis symptomatic
16	C787	Secondary malignant neoplasm of liver and intrahepatic bile duct
17	J910	Malignant pleural effusion
18	FEMALE	Indicator of sex
19	J449	Chronic obstructive pulmonary disease, unspecified
20	Z515	Encounter for palliative care
21	031	Diseases due to other mycobacteria
22	092	Early syphilis latent
23	J189	Pneumonia, unspecified organism
24	Z66	Do not resuscitate
25	032	Diphtheria
26	Z87891	Personal history of nicotine dependence
27	Z923	Personal history of irradiation
28	Z9221	Personal history of antineoplastic chemotherapy
29	C771	Secondary and unspecified malignant neoplasm of intrathoracic lymph nodes
30	G893	Neoplasm-related pain (acute) (chronic)

31	T451X5A	Adverse effect of antineoplastic and immunosuppressive drugs, initial encounter
32	AWEEKEND	Admission day is a weekend
33	J90	Pleural effusion, not elsewhere classified
34	AGE	Age in years at admission
35	J441	Chronic obstructive pulmonary disease with (acute) exacerbation
36	Z370	Single live birth
37	G936	Cerebral edema
38	D6481	Anemia due to antineoplastic chemotherapy
39	D61810	Antineoplastic chemotherapy induced pancytopenia
40	D630	Anemia in neoplastic disease
41	J95811	Postprocedural pneumothorax
42	C7989	Secondary malignant neoplasm of other specified sites
43	R64	Cachexia
44	002	Typhoid and paratyphoid fevers
45	Z681	Body mass index (BMI) 19.9 or less, adult
46	Z9981	Dependence on supplemental oxygen
47	E222	Syndrome of inappropriate secretion of antidiuretic hormone
48	Z902	Acquired absence of lung [part of]
49	Z3800	Single liveborn infant, delivered vaginally
50	J9621	Acute and chronic respiratory failure with hypoxia

Table 2: Top 50 features identifying lung cancer.

4. Machine Learning and Deep Learning Models

Machine Learning is a subfield of Artificial Intelligence that allows machines to learn like human beings through the utilization of data and algorithms. It also provides the ability to learn from the machines' prior experience and consequently improve the accuracy without having to be explicitly programmed [15, 16]. The ML models used in the project are briefly described below.

4.1. Decision Tree

Decision Trees are a type of supervised Machine Learning in which data is continually separated based on a specific parameter. Two entities: decision nodes, and leaves, can be used to explain the tree. The decisions or results are represented by the leaves as shown in **Figure 7**. The data is separated at the decision nodes. Decision Trees can handle categorical as well as continuous input and output variables [17].

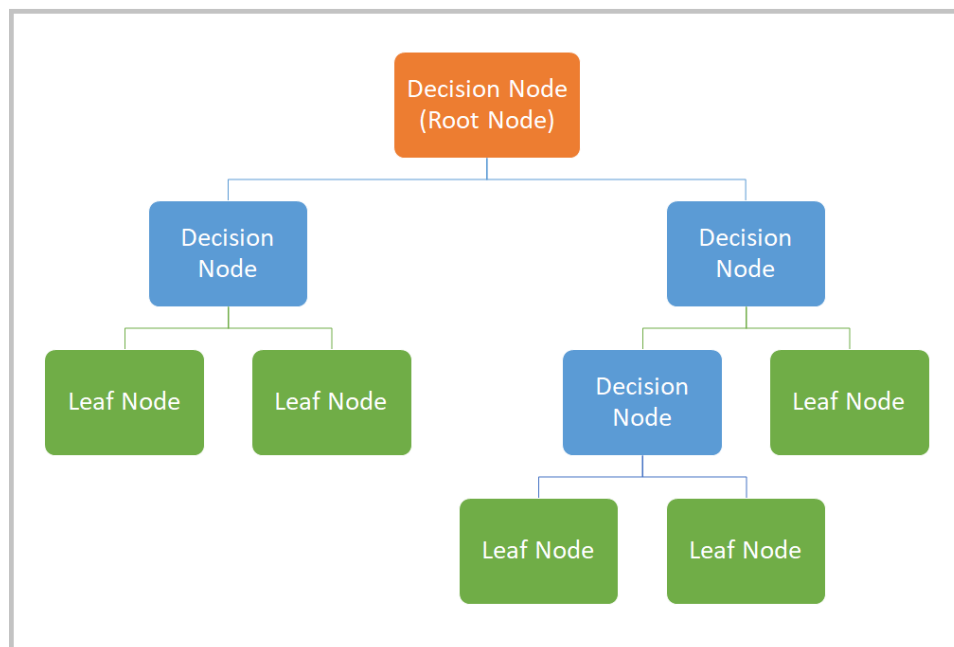


Figure 7: Decision Tree architecture [18].

4.2. Logistic Regression

Logistic Regression is a Supervised Learning Classification algorithm that calculates an estimation of distinct binary values (true/false, yes/no, 0/1) from a group of independent variables. It fits data to a logit function to forecast the probability of an event occurring as illustrated in **Figure 8**. As a result, Logistic Regression is also called Logit Regression. Its output falls in the range $[0,1]$ because it forecasts probability [17].

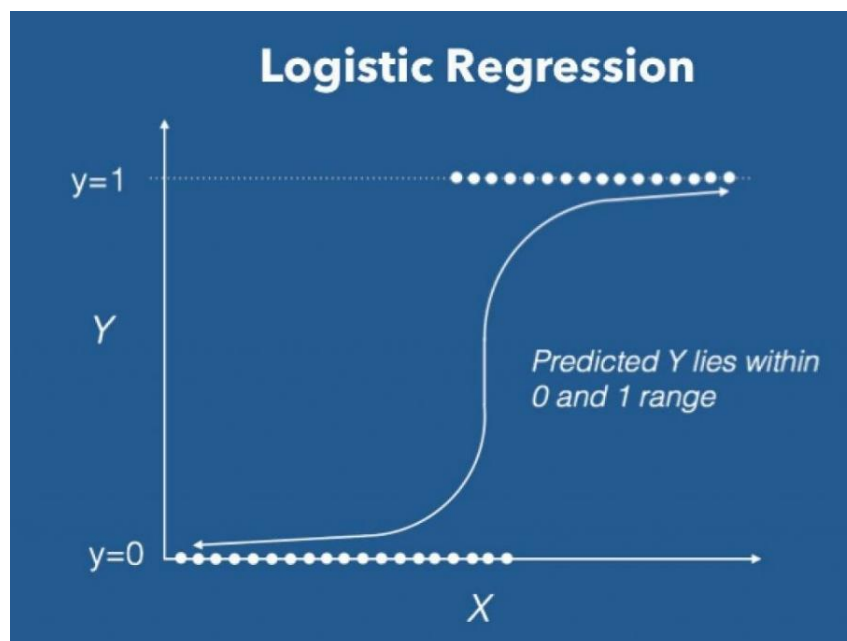


Figure 8: Logistic Regression [19].

4.3. Random Forest

Random Forest or Random Decision Forest is a Supervised Learning algorithm that creates a "forest" out of an ensemble of Decision Trees, which are commonly trained using the "bagging" method. The bagging method's basic premise is that combining several learning models improves the overall output. In simple terms, a Random Forest combines many Decision Trees to produce a more accurate and stable prediction [20]. This model may be used for both classification and regression. **Figure 9** illustrates a Random Forest model architecture.

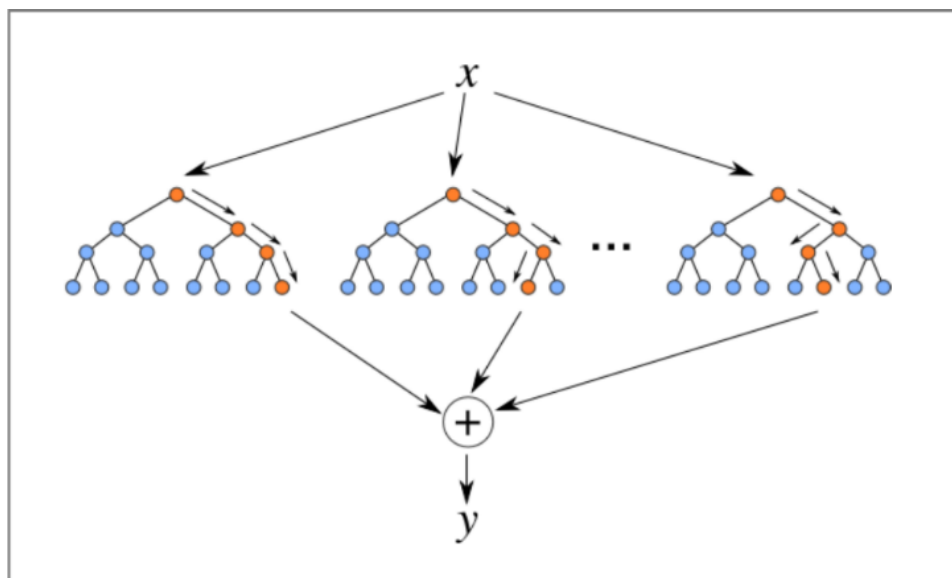


Figure 9: Random Forest architecture [21].

4.4. Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) is a Deep Learning model. It is a Feed-Forward Neural Network that consists of three types of layers: input layer, output layer, and hidden layer. The input signal that needs to be processed is received by the input layer. While the output layer is responsible for tasks such as prediction and classification, the hidden layers perform the main computation [22]. **Figure 10** depicts how data flows from the input to the output layer in the forward direction. The Back-Propagation Learning algorithm is used to train the neurons in the MLP. Pattern categorization, recognition, prediction, and approximation are some of MLP's most common applications [22].

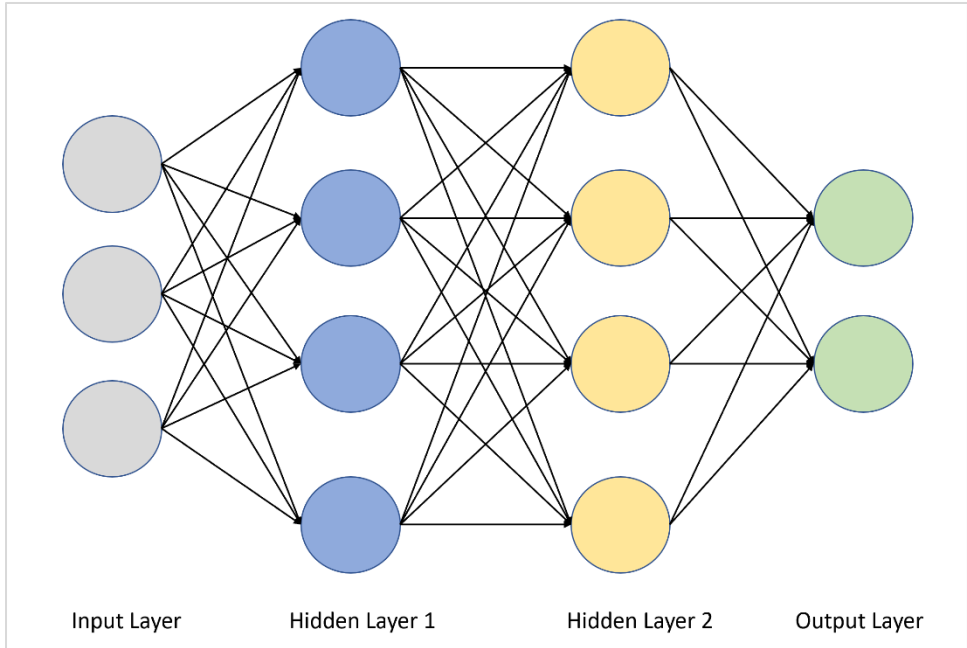


Figure 10: MLP model architecture.

5. Machine Learning Pipeline

A Machine Learning Pipeline is a method that allows automation of the steps required for creating an ML model. ML pipelines are made up of a series of steps that handle everything including data extraction, preprocessing, model training, model testing, and deployment [23]. The behavior of each step within the pipeline can be generalized, and every step can be built as a reusable component. The sequence in which the components are executed, as well as how inputs and outputs flow through the pipeline can be defined [23]. The pipeline makes the code flexible to work with various selectors, ML models, and estimators. **Figure 11** represents the flow of an ML pipeline.

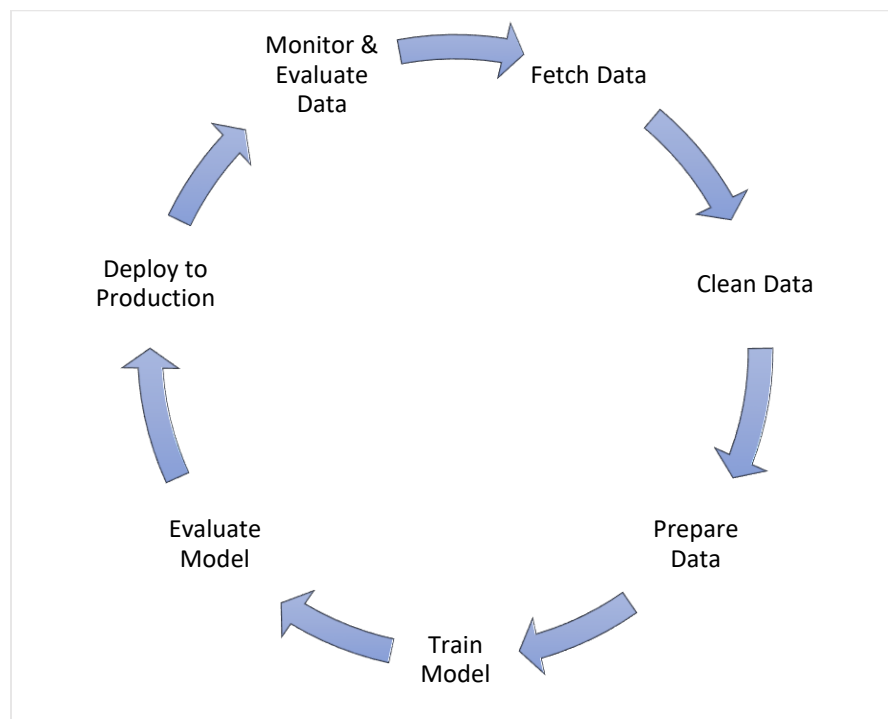


Figure 11: Machine Learning Pipeline [24].

After cleaning the data and identifying the top features, the dataset is passed to the Machine Learning pipeline, where it is split into training and testing datasets. 75% of the data is chosen for the training set and the remaining 25% is used for the testing set. Afterward, a custom

transformer is used on both the training and testing sets. The non-ICD features are indexed to integer values, just like the ICD dictionary. The transformer converts the dataset containing the top features into the LIBSVM format. Since the dataset contains many zeros (specifically the ICD columns), converting the dataset to the LIBSVM format makes better use of memory and makes the data easier to read.

In the LIBSVM format, each entry is represented by a line that starts with a label corresponding to the value of the target (in this case 0 or 1 depending on if the patient has lung cancer) followed by the index, value pairs separated by a colon. The index represents the feature number (indexed integer equivalent of the feature), and the value represents the nonzero value the feature holds. Features containing zero are omitted in the line. An example of a LIBSVM format entry is noted below:

```
<label> <index1>:<value1> <index2>:<value2> ...
```

The LIBSVM data is converted into a sparse matrix named “X_transformed”, which holds all the feature information, and a matrix “Y” which holds the target. The pipeline provides an option to switch among ML models. Depending on which model is passed to the pipeline, the transformed dataset is fitted into the appropriate model where predictions and accuracy are both calculated. The pipeline also keeps track of the training time for each iteration of the model. The pipeline makes the code flexible and generic as different disease prediction conditions, data size, and ML models can be used without having to change the main code.

6. Results

6.1. Important features

The database holds records of a total of 7,135,090 patients among which there exist 79,096 lung cancer patients. In this section, the lung cancer patient data is investigated and a combination of ICD features and non-ICD features including age, gender, and race are analyzed.

6.1.1. Age

The age of the lung cancer patients in the dataset lies between 0 to 90. The patients are divided into 5 age groups and the percentage of patients falling under each age group is depicted in **Figure 12**. Only a minority of the patients (4.20%) are below 50 years. On the other hand, the highest number of patients (about one-third) belong to the age group 61 to 70, followed closely by the age group 71 to 80 at 30.94%. These results match with the statistics reported by the American Cancer Society that most lung cancer patients are 65 years old or above [2].

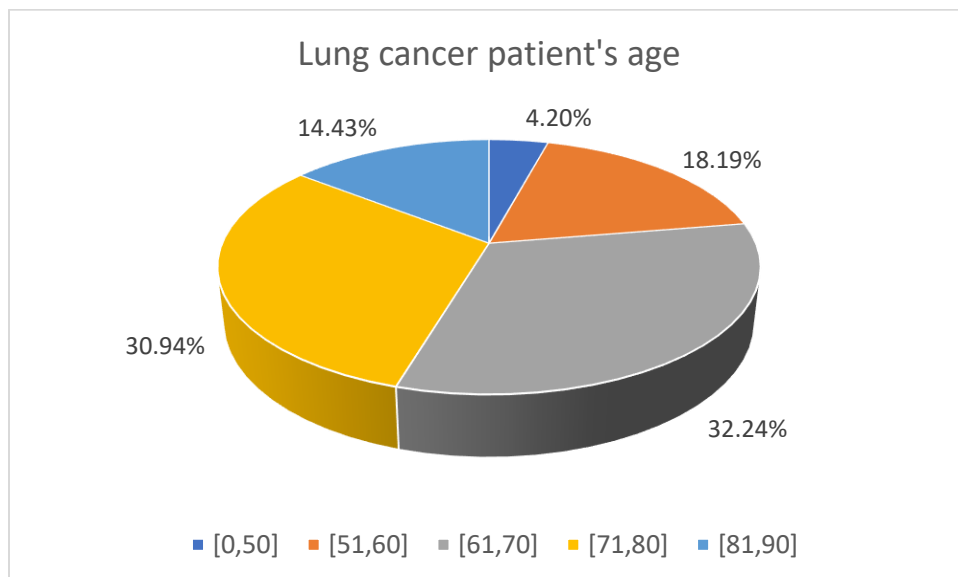


Figure 12: Lung cancer patient's age.

6.1.2. Gender

Figure 13 reveals that more than half of the lung cancer patients (51.46%) are male and the remaining 48.54% of patients are female. The total population consists of about 43% men and 57% women. 1.34% of the male population and 0.96% of the female population suffer from lung cancer. This indicates that men are at a slightly higher risk of developing lung cancer.

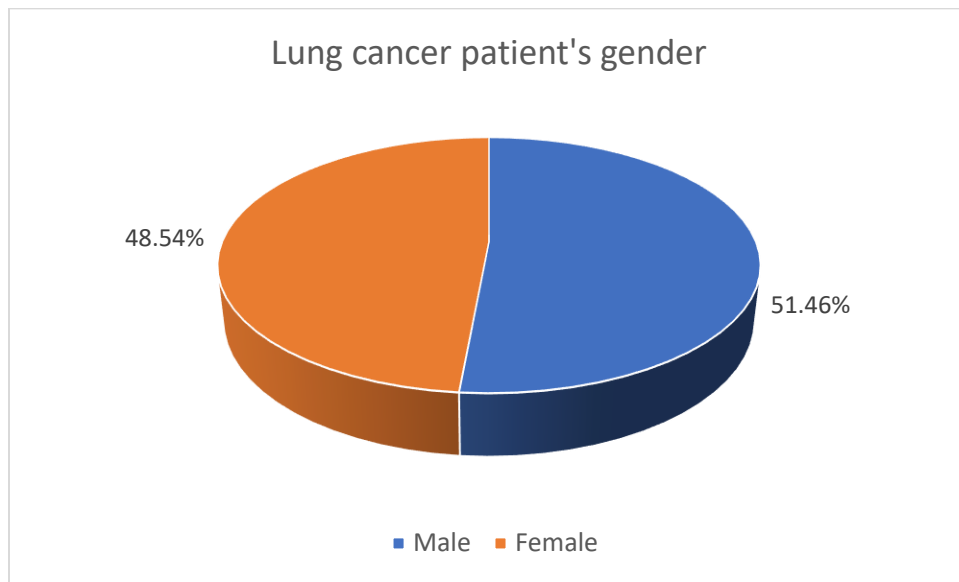


Figure 13: Lung cancer patient's gender

6.1.2. Race

The NIS database categorizes race into the following 6: White, Black, Hispanic, Asian, Native American, and others. **Figure 14** illustrates the distribution of lung cancer patients belonging to each of these racial groups. While 3/4th of the positive cases belong to the White population, they also account for a larger group among the total sample size. Due to the uneven racial makeup of the sample population, a more significant data indication may be the individual positive case percentage within each racial group.

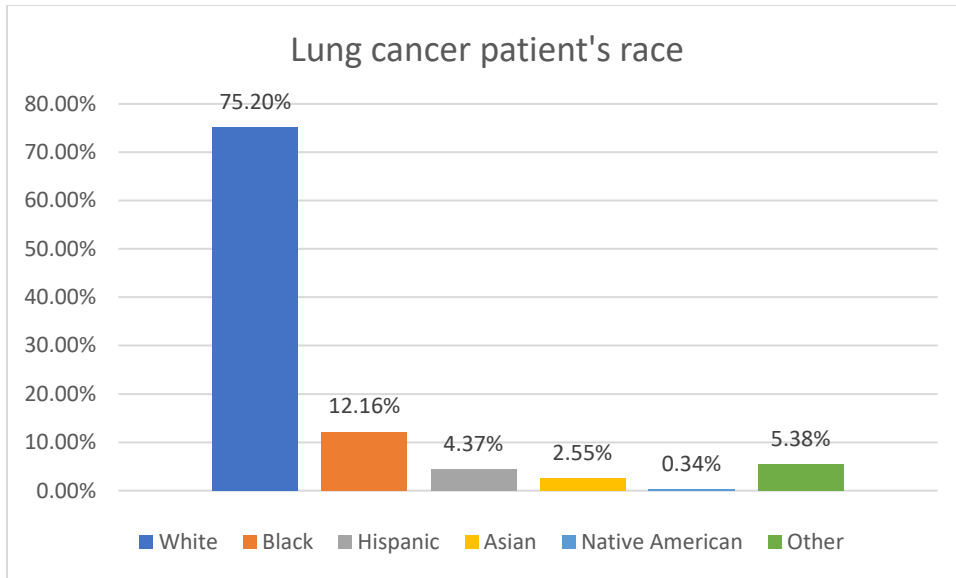


Figure 14: Lung cancer patient's race.

Figure 15 charts the percentage of lung cancer cases within each racial group. The data shows the highest rate of lung cancer cases (1.34%) among the white population. In contrast, the Hispanic population has the lowest rate of positive cases (0.42%). For Blacks and Asians, the rate of positive cases is somewhat similar (0.93% and 0.97% respectively) while Native Americans have a slightly lower rate of 0.62%.

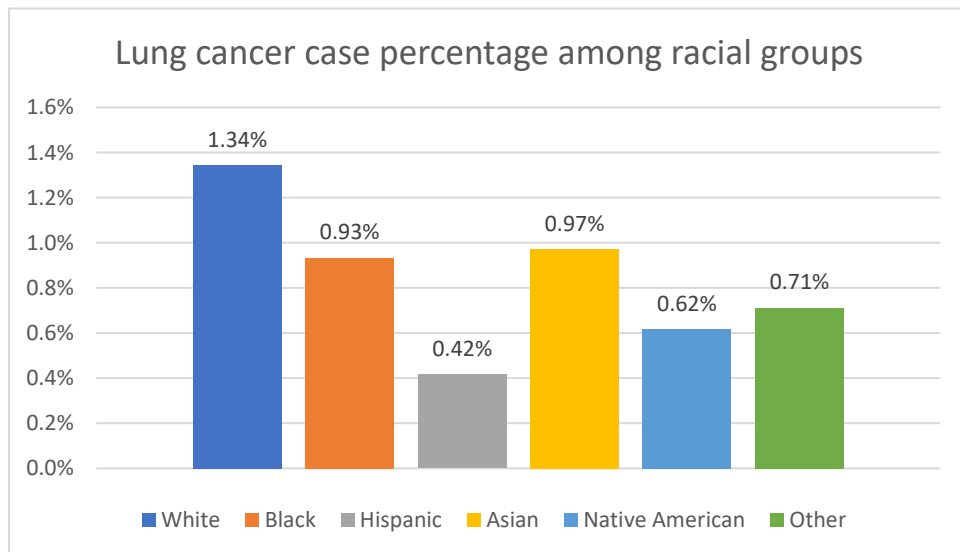


Figure 15: Lung cancer case percentage among racial groups.

6.1.4. Disease Conditions

In this section, some disease conditions are analyzed to find how these diseases influence lung cancer positive and negative patients. **Figure 16** indicates that Chronic Obstructive Pulmonary Disease (J449) affects 8.05% of the negative patients and 32.01% of the positive patients, which is roughly a 4 times higher rate than the negative rate. Secondary Malignant Neoplasm of Bone (C7951) is reported in 18.05% of the positive cases but just 0.63% in negative cases, which represents a major difference between the two groups. Pleural Effusion (J90) and BMI less than 20 in adults (Z681) have similar frequency where it is found in roughly 8.5% and 7% of the positive cases respectively and a little over 1.5% of the negative cases.

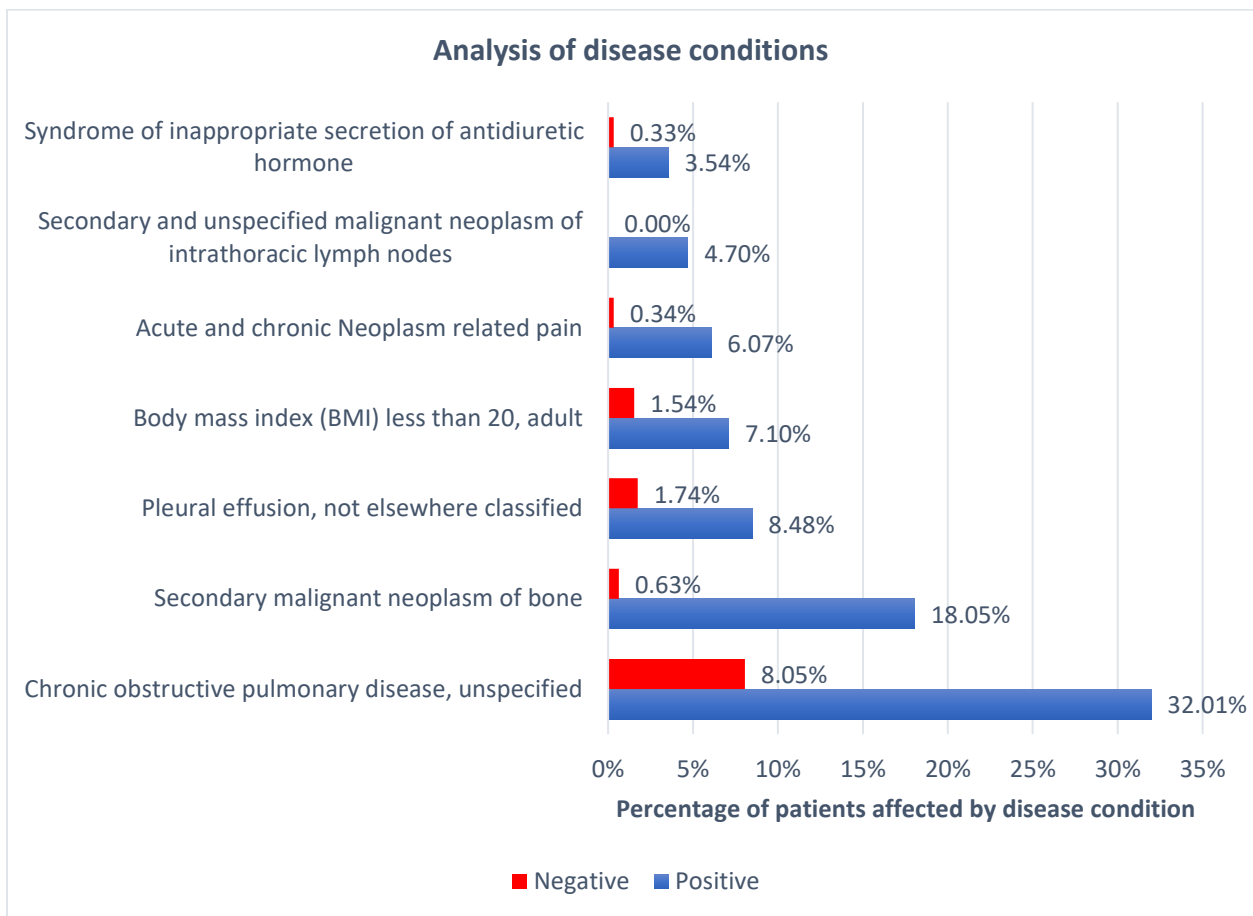


Figure 16: Analysis of disease conditions.

Lung cancer patients have about 18 times higher rate of suffering from acute and chronic Neoplasm-related pain (G893) compared to the negative patients. Secondary and unspecified Malignant Neoplasm of Intrathoracic Lymph Nodes (C771) is one of the few diseases that affects 4.70% of the positive patients but no negative patients. Lastly, the syndrome of inappropriate Secretion of Antidiuretic Hormone is recorded in 3.54% of the lung cancer patients compared to 0.33% of the negative patients.

Lung cancer patients are found to have a significantly higher probability of suffering from each of these 7 disease conditions in comparison to the negative patients. Hence, they are selected as important features contributing to lung cancer prediction.

6.2. Comparison among ML models

6.2.1. Accuracy

A portion of the original dataset is employed to calculate the accuracy of the four Machine Learning models. 10 iterations are performed for each model. In each iteration, about 320,000 records are chosen at random. Then about 80,000 records are chosen for the testing, while the remainder is used for training. **Table 3** lists the accuracy of the following models: Decision Tree, Logistic Regression, Random Forest, and Multi-Layer Perceptron with feature numbers varying from 5 to 50. The accuracy achieved with each feature number is taken by calculating the mean accuracy of 10 iterations.

Feature Number	Decision Tree Accuracy (%)	Logistic Regression Accuracy (%)	Random Forest Accuracy (%)	MLP Accuracy (%)
5	83.07	76.26	83.28	82.87
10	83.16	77.65	83.72	84.58
15	89.37	76.49	87.19	80.73
20	92.29	75.43	89.56	79.31
25	92.29	75.03	89.60	79.31
30	97.01	75.03	91.88	78.31
35	97.09	75.03	93.73	77.54
40	97.09	75.03	93.93	77.33
45	97.09	75.03	93.95	79.66
50	97.09	75.03	94.07	79.99
Best Accuracy	97.09	77.65	94.07	84.58

Table 3: Comparison among the Machine Learning models' accuracy.

The performance of Logistic Regression and MLP starts deteriorating as feature numbers increase from 10. This will be discussed in detail in section **6.2.3 Feature Saturation Curve**.

Figure 17 compares the best accuracy achieved by each of the models.

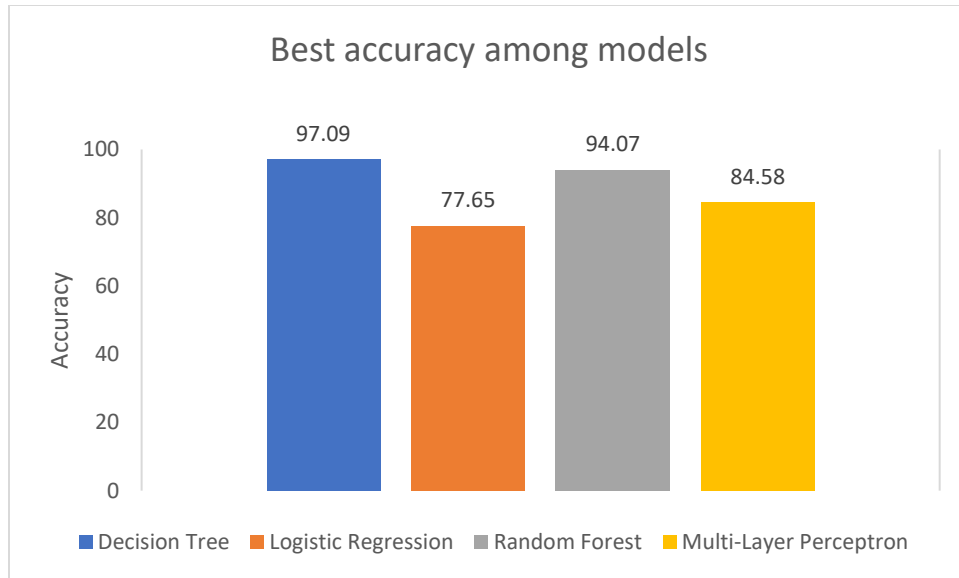


Figure 17: Best accuracy among models.

Among the four models, Decision Tree provides the best accuracy of 97.09%, followed by Random Forest having a 94.07% accuracy. MLP is 84.58% accurate whereas, Logistic Regression makes the least accurate forecast with an accuracy of 77.65%.

6.2.2. Training Time

Figure 18 provides a visual comparison of the training time for each model with different feature numbers while **Table 4** records the exact training time values.

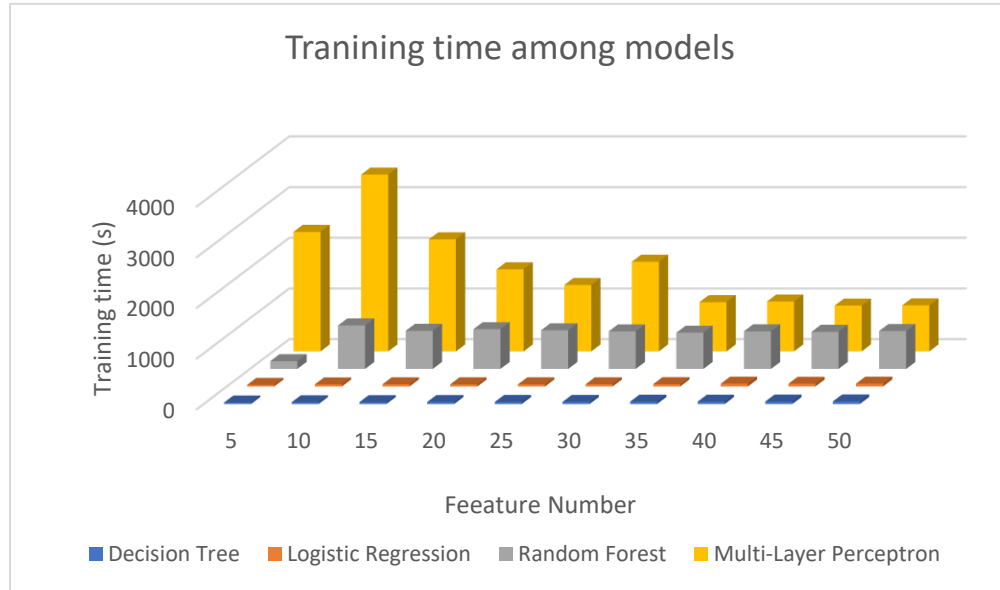


Figure 18: Training time among models.

Feature Number	Decision Tree Training Time (s)	Logistic Regression Training Time (s)	Random Forest Training Time (s)	MLP Training Time (s)
5	28.63	32.11	153.07	2357.99
10	34.83	41.64	858.35	3487.70
15	36.13	43.57	752.04	2211.19
20	38.83	41.74	784.22	1619.45
25	39.89	41.97	762.57	1311.03
30	40.98	43.14	741.17	1769.07
35	45.08	45.05	714.74	971.77
40	46.72	55.14	742.81	984.97
45	47.84	53.20	733.19	903.68
50	50.98	55.23	747.62	908.20

Table 4: Comparison among the Machine Learning models' training time.

Decision Tree and Logistic Regression are the two fastest models. Random Forest is quite slower than the previous two, but it is considerably faster than the most time-consuming MLP.

6.2.3. Feature Saturation Curve

As the feature number is changed the performance of each model varies as well. A feature saturation curve is drawn to visualize the change of accuracy with respect to the feature numbers. The goal is to identify the feature number which produces the highest accuracy for a model. **Figure 19** denotes the accuracy of Decision Tree saturates with 30 features. Logistic Regression provides the best accuracy with 10 features as identified from **Figure 20**. The accuracy undergoes a slow fall from feature number 10 to 25 and stays constant afterward.

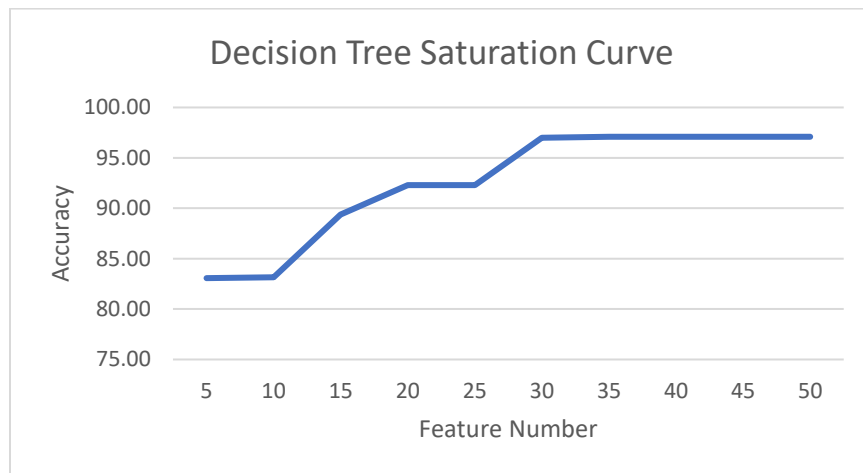


Figure 19: Decision Tree saturation curve.

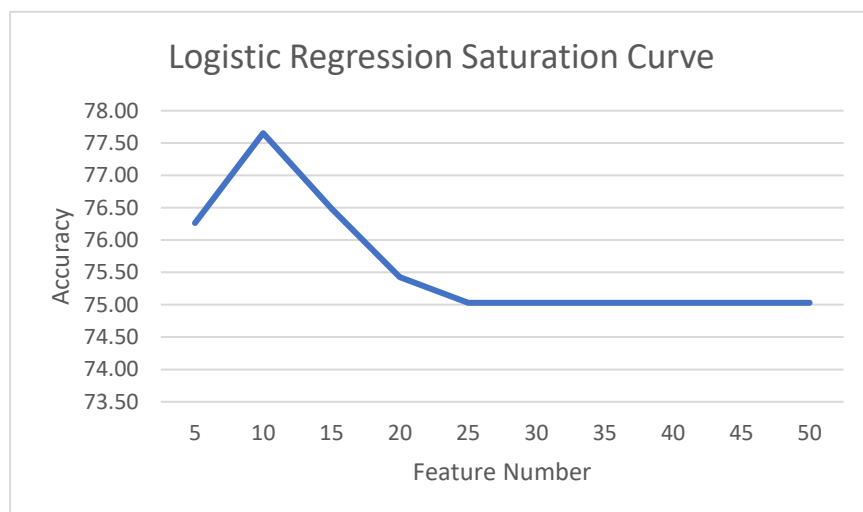


Figure 20: Logistic Regression saturation curve.

As shown in **Figure 21** Random Forest performs the best when the feature number is around 50. The exact saturating feature number could not be pinpointed as there is an upwards trend of accuracy even at feature number 50.

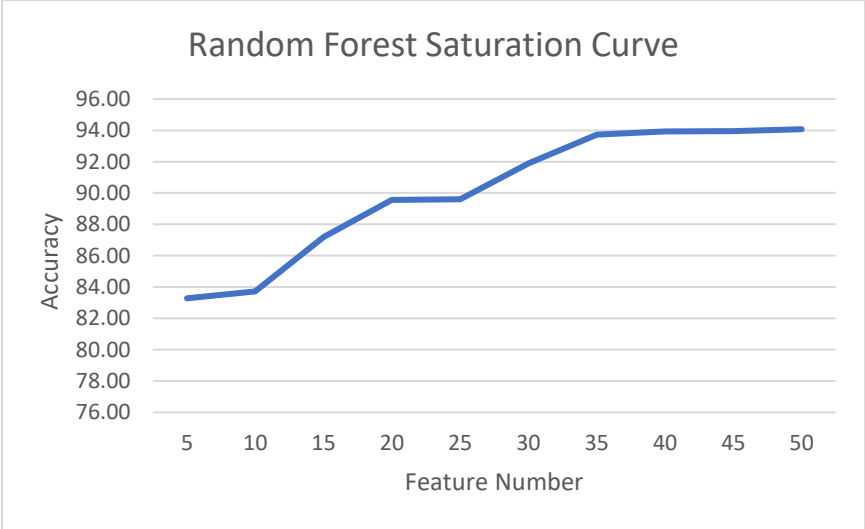


Figure 21: Random Forest saturation curve.

MLP’s accuracy initially increases with the feature number. A downward trend of accuracy is detected as the feature number increases from 10 to 40. After 40 features, there is a gradual rise in the accuracy. The best performance is observed with 10 features as illustrated in **Figure 22**.

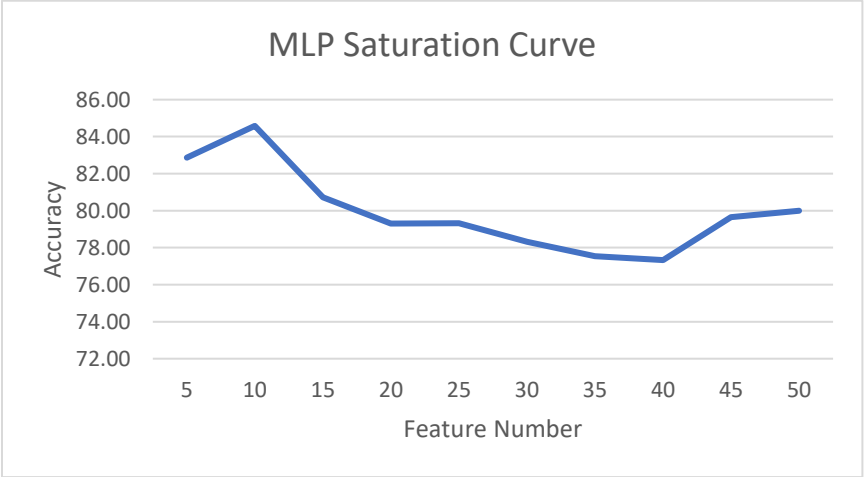


Figure 22: MLP saturation curve.

Figure 23 compares the results of the four ML models. Logistic Regression and MLP perform best with 10 features, where their performance comes up short against the other models. On the other hand, Decision Tree, and Random Forest saturate with 30, and 50 features, respectively.

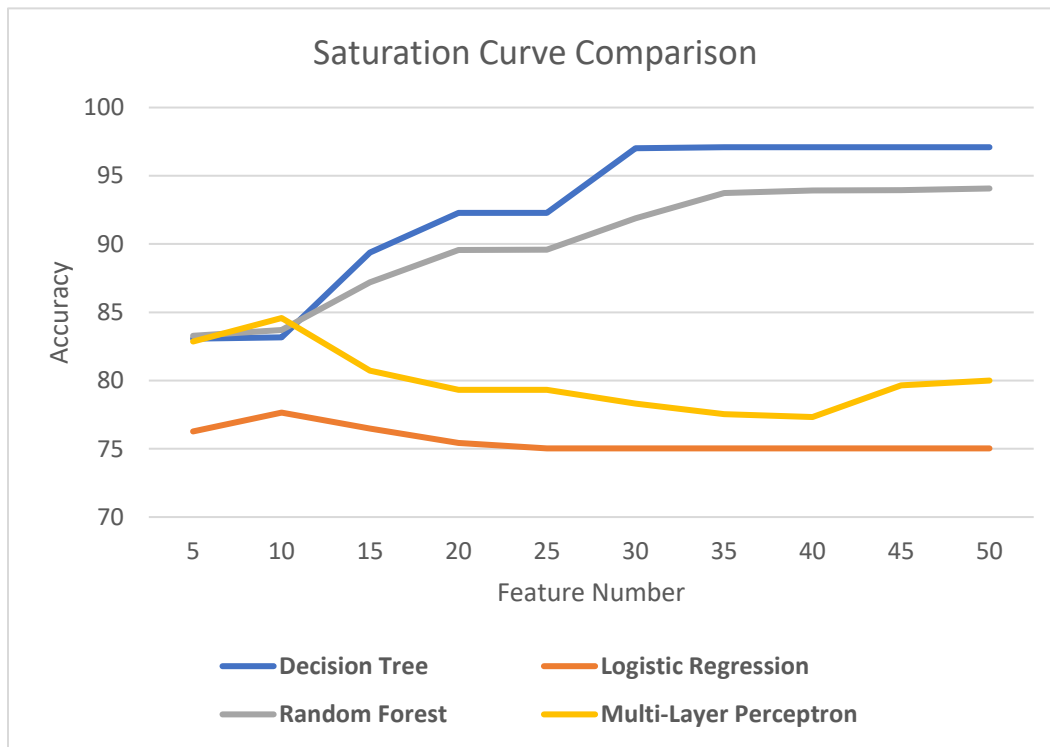


Figure 23: Saturation curve comparison.

6.3. Limitations and Future Work

Shortcomings of this project include the inferior performance of Logistic Regression and Multi-Layer Perceptron models. However, these models can be stabilized to produce better forecasts by fine-tuning the parameters such as dataset size, number of iterations, training time, training to testing ratio, and feature number. Many of these parameters are kept constant to ensure the four models' results are comparable.

Training Deep Learning models are generally resource-exhausting as they require a huge amount of computational power. Hence, parallel computing is essential to train models such as MLP. Due to resource limitations, MLP could not be explored thoroughly in this project. Tweaking MLP architecture by adding more hidden layers and increasing the layer size may result in alleviating the model.

In addition, some of the top features such as Anemia due to antineoplastic chemotherapy (D6481) are directly related to cancer treatment. The important disease conditions need to be analyzed comprehensively to understand the medical cause of the disease, and consequences they may have on a lung cancer patient. Diseases directly correlated to lung cancer treatment should be identified and removed from existing disease conditions to prepare the model for more realistic scenarios.

Future work may focus on these aspects to overcome the shortcomings. Also preparing a report that explains the reasoning behind the prediction would make the project more useful as physicians can use the reasoning to investigate further.

Conclusion

Lung cancer is responsible for most of the cancer-related deaths around the globe. This is partly because specific forms of lung cancer have significantly high metastasis (the development of secondary malignant growths at a distance from a primary site of cancer) rates, making early detection critical for its successful treatment. It is in the interest of expediting early-stage lung cancer prediction, that a Machine Learning pipeline is constructed that also identifies important features influencing the prediction. Admission level healthcare data has been used to train four ML models such as Decision Tree, Logistic Regression, Random Forest, and Multi-Layer Perceptron. These models have been chosen for their compatibility with natural language data, reliable predicting capability, and ease of implementation due to widely available resources.

The 10 most important features contributing to the lung cancer forecast include (1) HCUP Emergency Department service indicator, (2) race, (3) Major Diagnostic Category in use on the discharge date, calculated without Present on Admission indicators, (4) Diagnosis-Related Group version used on the discharge date, (5) Discharge quarter, (6) Median household income national quartile for patient, (7) Congenital syphilis, (8) NIS discharge weight, (9) Length of stay, and (10) Major Diagnostic Category in effect on the discharge date.

Decision Tree predicted lung cancer with the highest accuracy of 97.09% and the least training time among the four models used. Random Forest is the 2nd most precise (94.07%) although it is slower than Decision Tree and Logistic Regression. MLP is the slowest model which produces 84.58% accurate predictions. Logistic Regression is the 2nd fastest model, yet it has the least accuracy of 77.65%.

References

1. “Basic Information About Lung Cancer,” *Centers for Disease Control and Prevention*, 22-Sep-2020. [Online]. Available: https://www.cdc.gov/cancer/lung/basic_info/. [Accessed: 08-Jun-2021].
2. “Lung Cancer Statistics: How Common is Lung Cancer?,” *American Cancer Society*. [Online]. Available: <https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html>. [Accessed: 08-Jun-2021].
3. “Lung Cancer Symptoms,” *American Lung Association*. [Online]. Available: <https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/learn-about-lung-cancer/symptoms>. [Accessed: 08-Jun-2021].
4. “What Is Lung Cancer?: Types of Lung Cancer,” *American Cancer Society*, 01-Oct-2019. [Online]. Available: <https://www.cancer.org/cancer/lung-cancer/about/what-is.html>. [Accessed: 11-Jun-2021].
5. “USCS Data Visualizations - CDC,” *Centers for Disease Control and Prevention*. [Online]. Available: <https://gis.cdc.gov/Cancer/USCS/DataViz.html>. [Accessed: 08-Jun-2021].
6. “Introduction to the HCUP National Inpatient Sample (NIS), 2016,” *Healthcare Cost and Utilization Project (HCUP)*, 26-Oct-2018. [Online]. Available: https://hcup-us.ahrq.gov/db/nation/nis/NIS_Introduction_2016.jsp. [Accessed: 08-Jun-2021].
7. Z. S. Zubi and R. A. Saad, “Improves Treatment Programs of Lung Cancer Using Data Mining Techniques,” *Journal of Software Engineering and Applications*, vol. 07, no. 02, pp. 69–77, 2014, doi: 10.4236/jsea.2014.72008.
8. R. Kohad and V. Ahire, “Application of Machine Learning Techniques for the Diagnosis of Lung Cancer with ANT Colony Optimization,” *International Journal of Computer Applications*, vol. 113, no. 18, pp. 34–41, 2015, doi: 10.5120/19928-2069.
9. S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci, “Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1777–1787, Aug. 2019, doi: 10.1109/TMI.2019.2894349.
10. M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019, doi: 10.1186/s12911-019-0801-4.
11. G. Chauhan, "All about Naive Bayes", *Towards Data Science*, 2021. [Online]. Available: <https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf>. [Accessed: 18- Jun- 2021].
12. “ICD - ICD-10-CM - International Classification of Diseases, (ICD-10-CM/PCS Transition,” *Centers for Disease Control and Prevention*, 06-Nov-2015. [Online]. Available: https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm. [Accessed: 09-Jun-2021].

13. S. Thomas, "Data Cleaning in Machine Learning: Best Practices and Methods," *eInfochips*, 11-Dec-2019. [Online]. Available: <https://www.einfochips.com/blog/data-cleaning-in-machine-learning-best-practices-and-methods/>. [Accessed: 09-Jun-2021].
14. R. Shaikh, "Feature Selection Techniques in Machine Learning with Python", *Towards Data Science*, 2018. [Online]. Available: <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>. [Accessed: 20-Jun-2021].
15. "Machine Learning," *IBM*, 15-Jul-2020. [Online]. Available: <https://www.ibm.com/cloud/learn/machine-learning>. [Accessed: 09-Jun-2021].
16. "Machine Learning," *GeeksforGeeks*, 29-Jun-2020. [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/>. [Accessed: 09-Jun-2021].
17. S. Ray, "Commonly Used Machine Learning Algorithms: Data Science," *Analytics Vidhya*, 23-Dec-2020. [Online]. Available: <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>. [Accessed: 09-Jun-2021].
18. "Machine Learning Decision Tree Classification Algorithm - Javatpoint," *www.javatpoint.com*. [Online]. Available: <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>. [Accessed: 10-Jun-2021].
19. S. Prabhakaran, "Logistic Regression - A Complete Tutorial with Examples in R," *Machine Learning Plus*, 04-Jun-2021. [Online]. Available: <https://www.machinelearningplus.com/logistic-regression-tutorial-examples-r>. [Accessed 17-June-2021].
20. N. Donges, "A complete guide to the random forest algorithm," *Built In*, 16-Jun-2019. [Online]. Available: <https://builtin.com/data-science/random-forest-algorithm>. [Accessed: 09-Jun-2021].
21. C. Bakshi, "Random Forest Regression," *gitConnected*, 09-Jun-2020. [Online]. Available: <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>. [Accessed: 10-Jun-2021].
22. S. Abirami and P. Chitra, "Energy-efficient edge based real-time healthcare support system," *Advances in Computers*, pp. 339–368, 2020, doi: 10.1016/bs.adcom.2019.09.007.
23. *What is a Machine Learning Pipeline?* [Online]. Available: <https://valohai.com/machine-learning-pipeline/>. [Accessed: 09-Jun-2021].
24. *particle*, 13-Jun-2018. [Online]. Available: <https://www.particle-data.com/example-machine-learning-pipeline/>. [Accessed: 10-Jun-2021].