Cancer molecular Subtype Classification using Gene Expression Data: A Deep Learning Approach

Submitted By

Md. Jobayed Hossain Rabbi
(193-35-2948)
Department of Software Engineering

Supervised By

Mr. Musabbir Hasan Sammak
Lecturer
Department of Software Engineering

A thesis submitted in partial fulfillment of the requirement for the degree of

Bachelor of Science in Software Engineering

Fall 2023

APPROVAL

This thesis, titled "**Cancer molecular Subtype Classification using Gene Expression Data: A Deep Learning Approach**" submitted by Student Name (**ID: 193-35-2948**) to the Department of Software Engineering at Daffodil International University, has been accepted as satisfactory for partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering, as well as approval of its style and contents.

BOARD OF EXAMINERS

--- **Chairman**

Afsana Begum
Assistant Professor

Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

--- **Internal Examiner 1**

Md Rajib Mia
Lecturer

Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

--- **Internal Examiner 2**

Musabbir Hasan Sammak
Lecturer

Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

--- **External Examiner**

Dr. Md. Manowarul Islam
Associate Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

I hereby certify that, with the exception of equations and sources, this paper is all my own work. This thesis work is ready for submission as part of the B.Sc. in Software Engineering degree program. I further declare that it has not already been submitted for another degree at Daffodil International University.

Supervised by:
Mr. Musabbir Hasan Sammak
Lecturer
Department of Software Engineering
Daffodil International University

Submitted by:
Md. Jobayed Hossain Rabbi
ID: 193-35-2948
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

First of all, I am thankful to Allah, Who provides me the confidence and keeps me well so that I may successfully complete my thesis.

Second, I am grateful to my honorable teacher, **Mr. Musabbir Hasan Sammak**, **Lecturer, Department of Software Engineering, Daffodil International University**, for his boundless calmness, academic guidance, never-ending inspiration, constant and active oversight, organic criticism, valuable advice, reading numerous inferior disasters and fixing them at every level, which enabled me to successfully complete this thesis.

I am also grateful to all of the other faculty and staff members in our department for their participation and assistance.

TABLE OF CONTENTS

LIST OF FIGURE

ABSTRACT

Cancer is a term used to describe a group of diseases that are caused by abnormal cell growth and can spread throughout the body. According to the World Health Organization (WHO), cancer is the second leading cause of death after cardiovascular disease.Lung cancer stands as a major global health concern, with approximately 2.2 million new cases and 1.8 million deaths reported in 2020. Smoking remains the primary risk factor, though non-smokers can also be affected. The disease, often diagnosed at later stages, presents symptoms like persistent cough and chest pain. Treatment involves a combination of surgery, chemotherapy, radiation, targeted therapy, and immunotherapy, with prognosis varying based on factors like cancer stage and overall health. Prevention measures, particularly smoking cessation, are crucial, and early detection through regular screenings can significantly improve outcomes in addressing this serious and prevalent health issue.Lung cancer comprises various subtypes, with adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) being major types. Distinguishing between these subtypes is crucial for improved patient outcomes. LUAD often arises in the outer regions of the lungs and is associated with non-smokers, while LUSC tends to develop centrally and is linked to smoking. Identifying these subtypes aids in tailoring precise treatment strategies. LUAD responds well to targeted therapies, while LUSC may benefit more from traditional treatments like chemotherapy. This personalized approach enhances the efficacy of interventions, fostering better prognosis and overall survival rates. Hence, discerning between LUAD and LUSC is pivotal in advancing lung cancer management and optimizing patient care.Because cancer reflects biochemical processes in tissue and cells as well as an organism's genetic characteristics, gene expression can be useful in the early detection of cancer. Deoxyribonucleic acid (DNA) microarrays and ribonucleic acid (RNA)-sequencing methods for gene expression data allow quantifying gene expression levels and producing valuable data for computational analysis.The purpose of this paper is to look at recent advances in gene expression analysis for cancer classification using a variety of machine learning methods. We propose multiple machine learning approaches based on Neural Network, Decision Tree, KNN, Gradient Descent (XGboost), Regularization, and SVM in this paper. In a comparative analysis of various machine learning algorithms, a Neural Network exhibited superior performance with an accuracy of 95.07246%, surpassing other contenders such as Decision Tree, K-Nearest Neighbors, and

Gradient Descent (XGBoost) which collectively achieved an accuracy of 94.78%. Notably,The Regularization algorithm yielded an accuracy of 91.30435%. Interestingly, the Support Vector Machine algorithm demonstrated the lowest accuracy among the tested methods. This suggests that, in the context of the specific task or dataset under consideration, neural networks prove to be the most effective algorithm, showcasing their capacity for intricate pattern recognition and predictive accuracy. Our approach aims to classify whether and with what accuracy we can identify important genes for each type of cancer and whether good results can be obtained using multimodal data and multiple machine learning and deep learning models.he list of important genes identified in the analysis comprises a set of distinct genetic markers that have demonstrated significant relevance in the prediction model. These genes, denoted by their Ensembl Gene IDs, include ENSG00000134762.17, ENSG00000154227.13,ENSG00000169594.13,ENSG00000170484.10,ENSG00000121552.4,ENSG00000137975.8,ENSG00000073282.14,ENSG00000170465.10,ENSG00000171401.15,ENSG00000163331.12,ENSG00000137699.17,ENSG00000163032.12,ENSG00000125998.8,ENSG00000168453.15,ENSG00000081277.13,ENSG00000086570.12,ENSG00000110400.11,ENSG00000094796.5,ENSG00000094796.5,ENSG00000148600.15,ENSG00000166535.20,ENSG00000114948.13,ENSG00000173805.16,ENSG00000168143.9,ENSG00000128422.17,ENSG00000155918.8. These genes are instrumental in influencing the predictive accuracy of the model, as determined through the importance values assigned to them during the analysis. Understanding the biological functions and implications of these specific genes can provide valuable insights into the molecular mechanisms associated with the subtypes of lung cancer under consideration. Further exploration and validation of these genes may contribute to advancing our understanding of the underlying genetic factors contributing to the classification of lung cancer subtypes, ultimately facilitating more targeted and informed approaches in clinical settings.

CHAPTER 1

INTRODUCTION

1.1 General Topic & Background

Cancer is a broad term that refers to a variety of diseases that can affect any part of the body. Malignant tumors and neoplasms are other terms that are used. Cancer is distinguished by the rapid formation of abnormal cells that grow beyond their normal boundaries and can then invade neighboring parts of the body and spread to other organs; this latter process is known as metastasis. The most common cause of cancer death is widespread metastasis.[1] Cancer-related deaths are estimated to have killed 9.6 million people in 2017. Cancer is responsible for one out of every six deaths worldwide, making it the second leading cause of death, trailing only cardiovascular disease.[2]

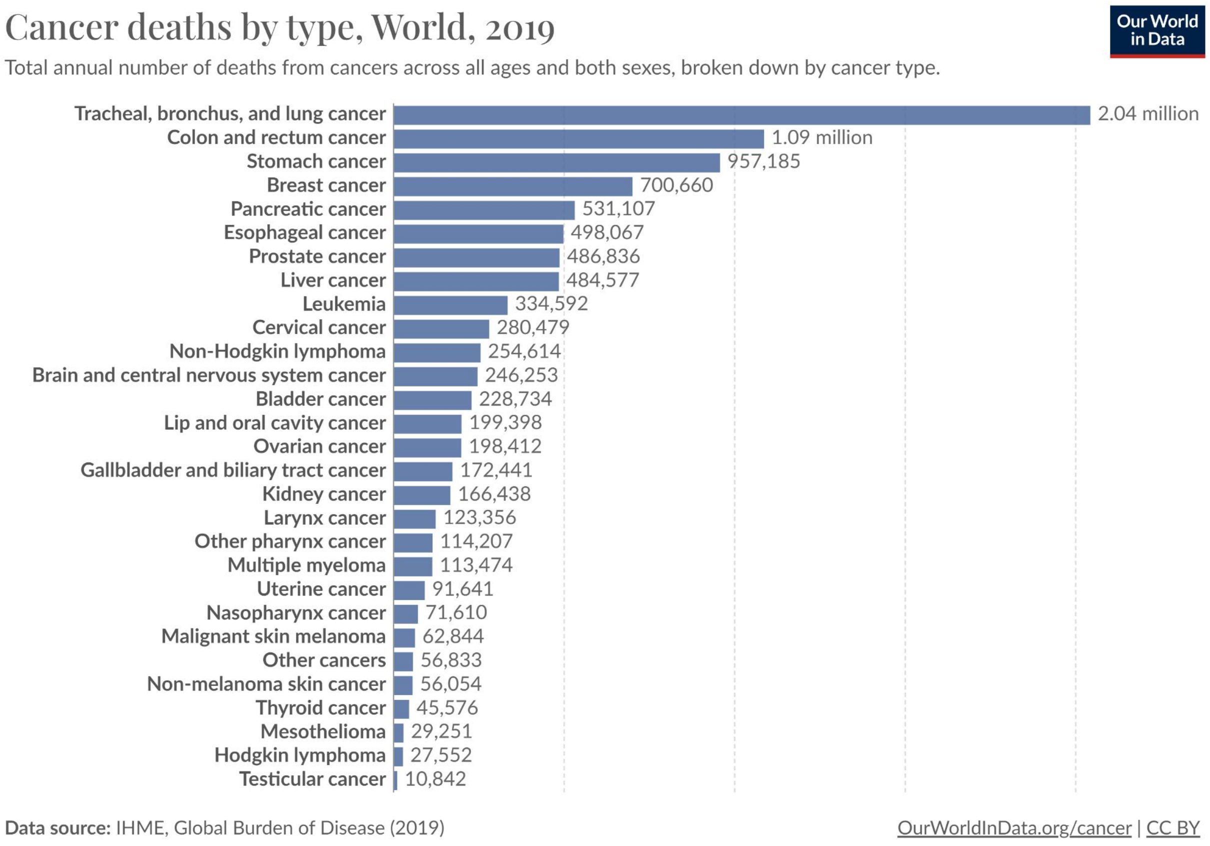


Figure-1: Cancer deaths by type, world 2019

Human cancer is a heterogeneous disease caused by random somatic mutations and by multiple genetic changes resulting from uncontrolled abnormal cell proliferation and spread to other cells and tissues [3,4]; Cancer disrupts an individual's intracellular homeostasis and thus poses a serious threat to human life.Towards personalized treatment plans, tissue-specific cancers can be classified into subtypes based on the molecular characteristics of the primary tumor.These subgroups provide an essential basis for delivering precise and personalized treatment [5,6] to cancer patients and have important implications for etiology, tumor biology, and prognosis in many cancer studies.Gene expression data reflect direct or indirect measurements of the mRNA abundance of gene transcripts in cells.These data can be used to analyze which gene expression characteristics have changed, the correlation between genes, and how gene activity is affected under different conditions.Therefore, these data have important applications in medical clinical diagnosis, assessment of drug effectiveness, and disease detection.mechanism.Therefore, gene expression data can be used in cancer subtype classification studies, and many methods based on gene expression data have been presented.[7]

Cancer is Kazakhstan's third leading cause of premature death. More than 30,000 people in Kazakhstan are diagnosed with cancer each year. This article was created to provide a basis for cancer control programs in Kazakhstan.The most common types of cancer are lung cancer, skin cancer, breast cancer, and stomach cancer. Together, these four cancers account for more than 44% of new cancer cases. Lung cancer is the most common cancer in men, accounting for almost a quarter of cancer cases in men. Breast cancer is the most common cancer in women, accounting for 20% of cases. Cancer remains primarily a disease of older Kazakhs.The largest proportion of cancer deaths in men and women was due to lung cancer, primarily due to smoking. Deaths from lung cancer, stomach cancer, breast cancer, and esophageal cancer together account for almost half (46%) of all cancer deaths. Cancer remains an important public health problem in Kazakhstan, with an estimated 186. 7 new infections and 166. 7 deaths in 2006. The incidence of lung cancer and some other cancers can be reduced by tobacco control and healthy lifestyle improvements[8].

Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) are the most common types of lung cancer. SCLC is less common but often grows quickly, whereas NSCLC is more common and grows slowly[9]. The International Agency for Research on Cancer (IARC) GLOBOCAN 2020 estimates of cancer incidence and mortality show that lung cancer remains the leading cause of cancer death, with an estimated 1.8 million deaths (18%) in 2020[9]. LUAD and LUSC are the two most common types of non-small cell lung cancer (NSCLC). LUAD is an abbreviation for lung adenocarcinoma, and LUSC is an abbreviation for lung squamous cell carcinoma.Every year, over 150,000 patients die from lung cancer, with 154,050 deaths expected in 2018. Lung cancer is one of the most common cancers in the world, and it is caused by factors such as smoking and exposure to toxic chemicals[10].Necessity of classification for lung cancer prediction and gene expression data identified.

In cancer research, ML and DL techniques have been widely used for a variety of purposes, including cancer subtype classification, prognosis prediction, and personalized treatment.

1.2 Problem Statement

Most researchers have done lung cancer prediction using gene expression data, clinical data or DNA methylation data etc. But by combining all the data together, no classification i.e. prediction has been done yet. Moreover, their sample size was small and multiple models were not used.

1.3 Research Questions

Using multimodal data and multiple machine learning and deep learning models to classify whether and with what accuracy we can identify important genes for each type of cancer and whether good results are coming.

1.4 Research Objectives

First, after collecting the data, data preprocessing has to be done. Data visualization, Data missing value fixed, Normalization and Annotation should be done while doing data preprocessing. After model training, selecting the best model and hyperparameter tuning should be done. Important genes must be identified. Finally we have to evaluate the model.

1.5 Research Scope

Gene expression data and clinical data used for cancer subtype classification are only for USA and UK people. Since it is made for USA and UK people then its result will be for them. The results will not be the same for any other country because people from all countries have different genes.

1.6 Paper Organization

In the organization of this paper, seven chapters are included as follows: In chapter 1, general topic & background, problem statement, research questions, research objectives, and research scope of this research are discussed. In chapter 2, Literature review. In chapter 3, description of the dataset implemented in this study. In chapter 4, a brief discussion of the methodology applied in this study. In chapter 5 and 6, the results and discussion of this experiment and conclusion of the study respectively.

CHAPTER 2

LITERATURE REVIEW

Shen et al. et al. "A Deep learning approach for cancer subtype classification using high-dimensional gene expression data." 1-17 in BMC bioinformatics 23.1 (2022). They used the BLCA-TCGA and BLCA-CIT datasets, as well as DCGN approach combined a convolutional neural network (CNN) and bidirectional gated recurrent unit (BiGRU) methodology, and findings combines a (CNN) and (BiGRU) seven other cancer subtype classification and achieved 99.3% accuracy[11].

Nitao Cheng et al. "Prediction of lung cancer metastasis by gene expression." 106490 in Computers in Biology and Medicine 153 (2023). They achieved an accuracy of 82.29% by using Gene Expression data from 226 patient datasets and a Deep Neural Network (DNN) methodology with findings that their method achieved the best precision compared to other methods.Limited sample size..[12]

Liu, Suli, and Wu Yao. "Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection." BMC bioinformatics 23.1 (2022): 175. They are used in LUAD gene expression data from the TCGA and ICGC portals. To achieve 99% accuracy, the sample size is 533 from TCGA and 488 from ICGC using the Deep Neural Network method and findings of this paper This model achieves an AUC of 0.99 which indicates the high generalization performance. The sample size is small. [13]

Bernardo Ramos, PhD, et al. "An interpretable approach for lung cancer prediction and subtype classification using gene expression." IEEE Engineering in Medicine & Biology Society (EMBC) 43rd Annual International Conference, 2021. IEEE, 2021. They achieved 97.1% accuracy using Gene Expression from TCGA datasets and methodology used to Light GBM (tree-based learning)

with findings applying cancerous tissue to normal tissue to predict cancer separates LUAD from LUSC samples for subtype classification.[14].
Joe W. Chen and Joseph Dhahbi. "Using overlapping feature selection methods to classify lung adenocarcinoma and lung squamous cell carcinoma, identify biomarkers, and analyze gene expression." 13323 in Scientific Reports 11.1 (2021).They achieved a 90% accuracy by they are used an external dataset called GSE28582 datasets and Classification statistics of the overlapping method and findings the study also performed gene expression analysis to investigate biological differences between LUAD and LUSC..[15]

Sheetal Rajpal and colleagues. "Deep learning-based model for breast cancer subtype classification." arXiv preprint arXiv:2111.03923 (2021). They used Gene expression data from TCGA for 1218 patients, as well as the DNN and Autoencoder methods, findings The model achieved a mean 10-fold test accuracy of 0.907 on the TCGA breast cancer dataset. to achieve an accuracy of 90.7%.[16]

Kim, Bong-Hyun, Kijin Yu, and Peter CW Lee published "Cancer classification of single-cell gene expression data by neural network." 1360-1366 in Bioinformatics 36.5 (2020). They used 21 different types of cancer gene expression data and the Neural Network, SVM, KNN, and RF methods, finding NN performed consistently better than other methods, which is 94%. improve the accuracy, to achieve a 94% accuracy.[17]

Su, Ran, et al. ""Identification of expression signatures for the classification of non-small-cell lung carcinoma subtypes." Bioinformatics 36.2 (2020): 339-346.". They are used an Gene Expression data from TCGA and Random Forest and Support Vector Machine method to achieve an accuracy of 95.1% .[18]

Xu, Jing, et al. ""A novel deep flexible neural forest model for cancer subtype classification based on gene expression data." IEEE Access 7 (2019): 22086-22095.".They are used collect three type of cancer (BRCA,GBM,LUNG) gene expression data from TCGA and KNN, SVM ,MLP and RF, DFNF method, findings A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data,to achieve an accuracy of 93.6% .[19]

Author	**Year**	Author	Data	Methodology	Findings	Accuracy
Deep learning method for classifying cancer subtypes using high-dimensional gene expression data	2022	Shen, Jiquan, et al	BLCA-TCGA and BLCA-CIT datasets	DCGN approach combined a convolutional neural network (CNN) and bidirectional gated recurrent unit (BiGRU)	Combines a (CNN) and (BiGRU) seven other cancer subtype classification.	99.3%
Gene expression predicts lung cancer metastasis	2023	Cheng, Nitao, et al.	Gene Expression data from 226 patients.	Deep Neural Network (DNN)	Their method achieved the best precision compared to other methods.	82.29%
Lung cancer prediction using gene expression and deep learning with KL divergence gene selection	2022	Liu, Suli, et al.	They are used an external dataset called GSE28582	Classification statistics of the overlapping method.	The study also performed gene expression analysis to investigate biological differences between LUAD and LUSC.	90%

An interpretable approach for lung cancer prediction and subtype classification using gene expression	2019	Ramos, Bernardo, et a Sherafatian, Masih, and Fateme Arjmand	**LUAD-**499 solid tumor and 46 normal tissues. **LUSC-** 478 solid tumors and 45 normal control samples (21,22)	Decision Tree	miRNAs to distinguish lung tumors from normal samples and further classify the tumors into lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) subtypes .	97.1%
Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression	2021	Ramos, Bernardo, et a	They are used an external dataset called GSE28582	Classification statistics of the overlapping method.	The study also performed gene expression analysis to investigate biological differences between LUAD and LUSC.	90%
Deep Learning Based Model for Breast Cancer Subtype Classification	2021	Rajpal, S et al.	Gene expression data from TCGA from 1218 patients	DNN and Autoencoder	The model achieved a mean 10-fold test accuracy of 0.907 on	90.7%

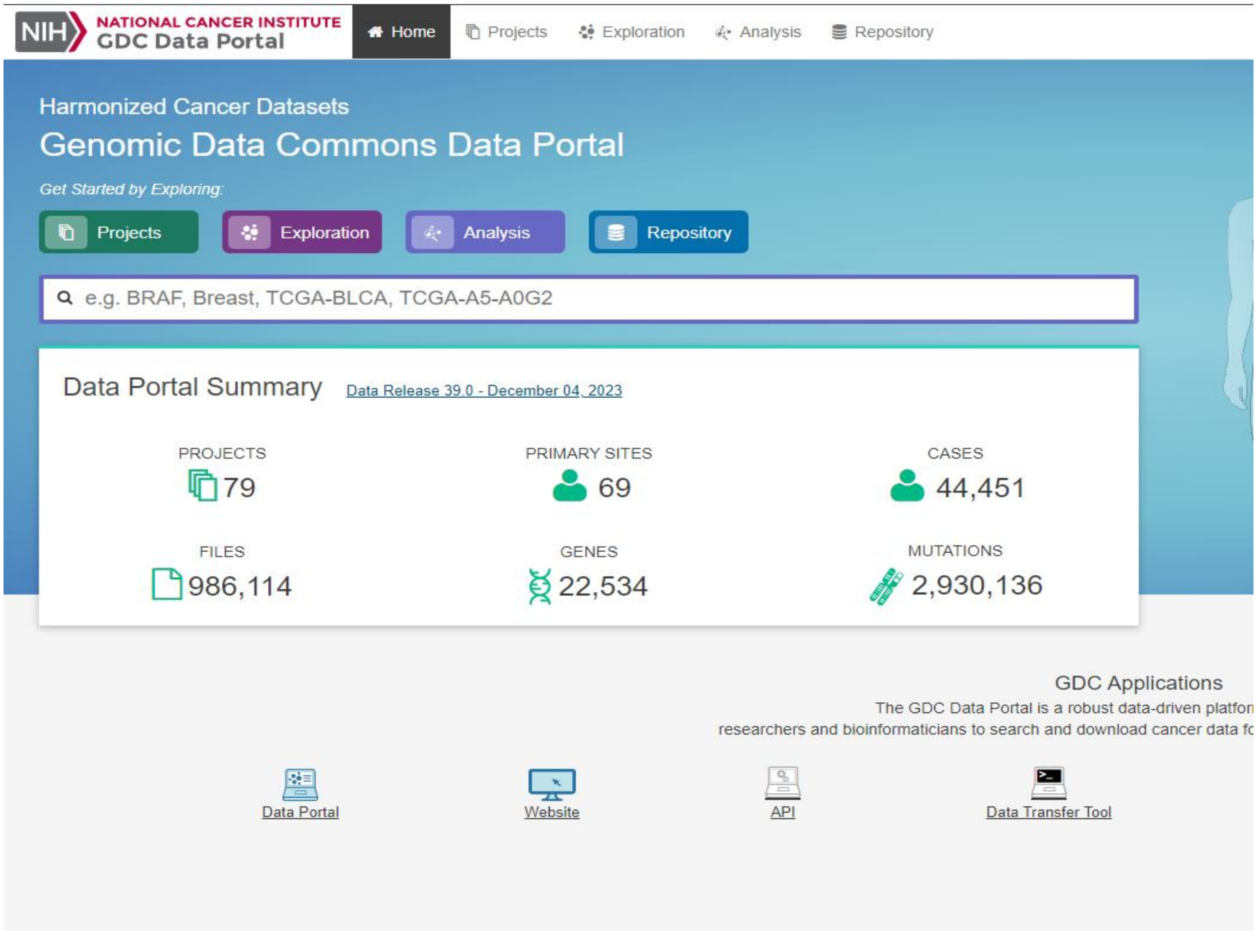
					the TCGA breast cancer dataset.	
Cancer classification of single-cell gene expression data by neural network	2020	Kim, B et al.	21 types of cancer gene expression data.	Neural Network, SVM, KNN, RF	NN performed consistently better than other methods which is 94%.	94%
Identification of expression signatures for non-small-cell lung carcinoma subtype classification	2020	Su, R. et al.	Gene Expression data from TCGA.	Random Forest and Support Vector Machine	Finding the RF 95.1% highest accuracy. Used less model	95.1%
A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data.	2019	Xu, J et al.	Collect three types of cancer (BRCA,GBM,LUNG) gene expression data from TCGA.	DFNF	A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data. Use less cancer type.	93.6%

CHAPTER 3

DATASET

Tabular Data: Lung cancer data can be organized into structured tables, where each row represents a patient or a case, and columns represent various attributes or features related to those patients. These attributes may include patient demographics, medical history, genetic markers, tumor characteristics, and clinical outcomes. Tabular data is commonly used for tasks like lung cancer diagnosis, prognosis prediction, and risk assessment.**Medical Images:** Medical imaging data, such as chest X-rays and CT scans, are crucial for diagnosing and monitoring lung cancer. These images are in the form of pixel values arranged in a grid. Deep learning models, especially Convolutional Neural Networks (CNNs), are often used to analyze and classify these images to detect lung cancer or assess its progression.**Genomic Data:** Genomic data related to lung cancer can include DNA sequences, gene expression profiles, and genetic mutations associated with the disease. These data types are critical for understanding the molecular mechanisms of lung cancer and developing personalized treatment approaches.**Free-text Clinical Notes:** Clinical notes and reports generated by healthcare professionals can contain valuable information about patient history, symptoms, and treatment plans. Natural Language Processing (NLP) techniques are used to extract structured information from unstructured clinical text data.**Time Series Data:** In cases where lung cancer progression is monitored over time, time series data may be used. This can include data points collected at regular intervals, such as tumor size measurements or patient vital signs, to track disease progression or treatment response.**Molecular Pathology Data:** This type of data includes information on the histological characteristics of lung cancer tissue, such as tissue staining patterns and cellular morphology, which can aid in diagnosis and classification[20].Gene expression data and clinical data were downloaded from the TCGA data portal. The National Cancer Institute (NCI) collected this dataset from patients in the United States. Two subtypes of Lung Adenocarcinoma (LAAD) and Lung Squamous Cell Carcinoma

(LUSC) gene expression data and clinical data were collected from the TCGA portal. LUAD has 600 data samples and LUSC has 553 data samples. Among them total primary tissue is 1043 and 110 are normal tissue. So total 1153 both subtypes of lung cancer.[21]



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Data Portal Summary Data Release 39.0 - December 04, 2023

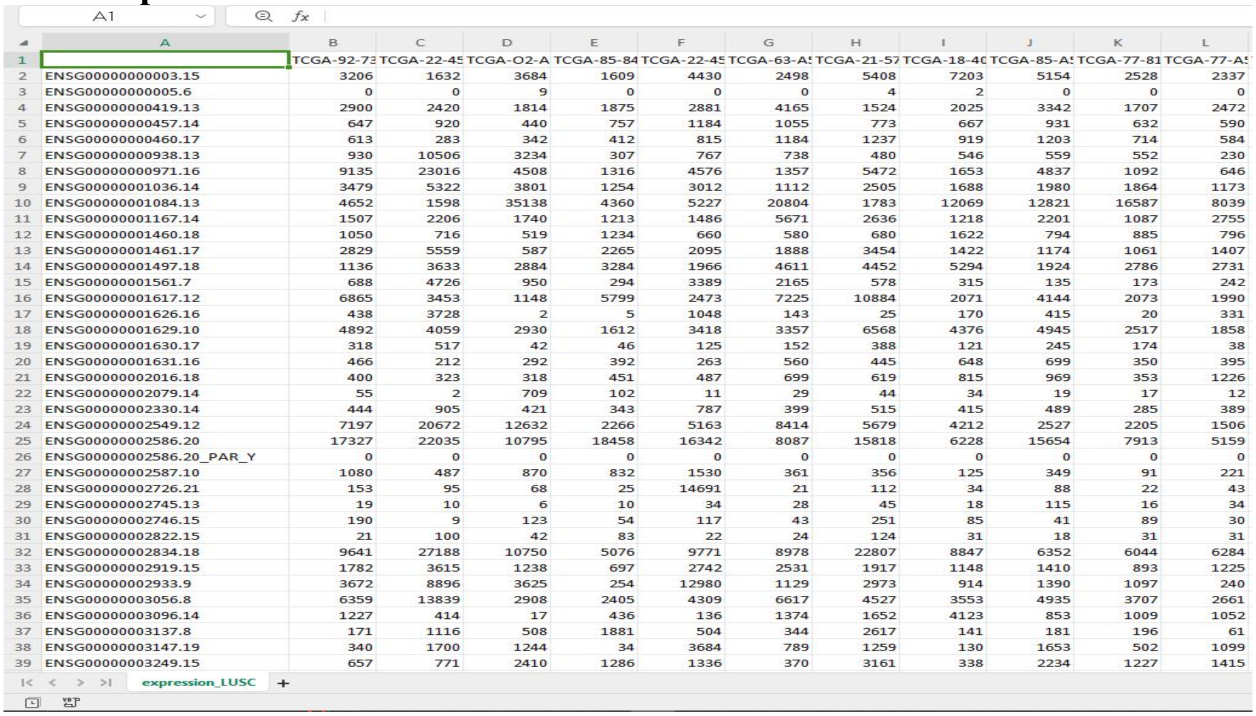
PROJECTS	PRIMARY SITES	CASES
79	69	44,451
FILES	**GENES**	**MUTATIONS**
986,114	22,534	2,930,136

GDC Applications

The GDC Data Portal is a robust data-driven platfon
researchers and bioinformaticians to search and download cancer data fc

Figure-2: Genomic Data Commons Data Portal

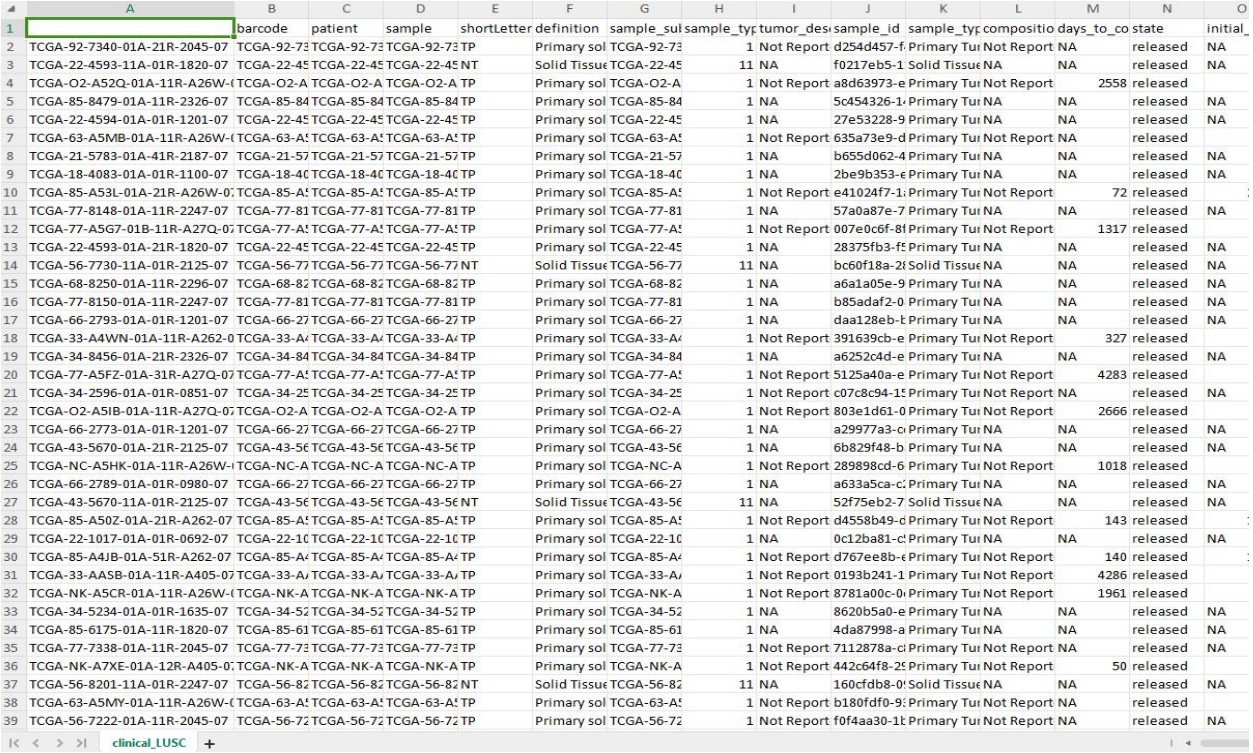
Gene Expression LUSC Data



	A	B	C	D	E	F	G	H	I	J	K	L
1		TCGA-92-73	TCGA-22-45	TCGA-O2-A	TCGA-85-84	TCGA-22-45	TCGA-63-A	TCGA-21-57	TCGA-18-40	TCGA-85-A	TCGA-77-81	TCGA-77-A
2	ENSG00000000003.15	3206	1632	3684	1609	4430	2498	5408	7203	5154	2528	2337
3	ENSG00000000005.6	0	0	9	0	0	0	4	2	0	0	0
4	ENSG00000000419.13	2900	2420	1814	1875	2881	4165	1524	2025	3342	1707	2472
5	ENSG00000000457.14	647	920	440	757	1184	1055	773	667	931	632	590
6	ENSG00000000460.17	613	283	342	412	815	1184	1237	919	1203	714	584
7	ENSG00000000938.13	930	10506	3234	307	767	738	480	546	559	552	230
8	ENSG00000000971.16	9135	23016	4508	1316	4576	1357	5472	1653	4837	1092	646
9	ENSG00000001036.14	3479	5322	3801	1254	3012	1112	2505	1688	1980	1864	1173
10	ENSG00000001084.13	4652	1598	35138	4360	5227	20804	1783	12069	12821	16587	8039
11	ENSG00000001167.14	1507	2206	1740	1213	1486	5671	2636	1218	2201	1087	2755
12	ENSG00000001460.18	1050	716	519	1234	660	580	680	1622	794	885	796
13	ENSG00000001461.17	2829	5559	587	2265	2095	1888	3454	1422	1174	1061	1407
14	ENSG00000001497.18	1136	3633	2884	3284	1966	4611	4452	5294	1924	2786	2731
15	ENSG00000001561.7	688	4726	950	294	3389	2165	578	315	135	173	242
16	ENSG00000001617.12	6865	3453	1148	5799	2473	7225	10884	2071	4144	2073	1990
17	ENSG00000001626.16	438	3728	2	5	1048	143	25	170	415	20	331
18	ENSG00000001629.10	4892	4059	2930	1612	3418	3357	6568	4376	4945	2517	1858
19	ENSG00000001630.17	318	517	42	46	125	152	388	121	245	174	38
20	ENSG00000001631.16	466	212	292	392	263	560	445	648	699	350	395
21	ENSG00000002016.18	400	323	318	451	487	699	619	815	969	353	1226
22	ENSG00000002079.14	55	2	709	102	11	29	44	34	19	17	12
23	ENSG00000002330.14	444	905	421	343	787	399	515	415	489	285	389
24	ENSG00000002549.12	7197	20672	12632	2266	5163	8414	5679	4212	2527	2205	1506
25	ENSG00000002586.20	17327	22035	10795	18458	16342	8087	15818	6228	15654	7913	5159
26	ENSG00000002586.20_PAR_Y	0	0	0	0	0	0	0	0	0	0	0
27	ENSG00000002587.10	1080	487	870	832	1530	361	356	125	349	91	221
28	ENSG00000002726.21	153	95	68	25	14691	21	112	34	88	22	43
29	ENSG00000002745.13	19	10	6	10	34	28	45	18	115	16	34
30	ENSG00000002746.15	190	9	123	54	117	43	251	85	41	89	30
31	ENSG00000002822.15	21	100	42	83	22	24	124	31	18	31	31
32	ENSG00000002834.18	9641	27188	10750	5076	9771	8978	22807	8847	6352	6044	6284
33	ENSG00000002919.15	1782	3615	1238	697	2742	2531	1917	1148	1410	893	1225
34	ENSG00000002933.9	3672	8896	3625	254	12980	1129	2973	914	1390	1097	240
35	ENSG00000003056.8	6359	13839	2908	2405	4309	6617	4527	3553	4935	3707	2661
36	ENSG00000003096.14	1227	414	17	436	136	1374	1652	4123	853	1009	1052
37	ENSG00000003137.8	171	1116	508	1881	504	344	2617	141	181	196	61
38	ENSG00000003147.19	340	1700	1244	34	3684	789	1259	130	1653	502	1099
39	ENSG00000003249.15	657	771	2410	1286	1336	370	3161	338	2234	1227	1415

Figure-3:Gene Expression LUSC Data

Clinical LUSC Data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		barcode	patient	sample	shortLetter	definition	sample_sul	sample_typ	tumor_des	sample_id	sample_typ	compositio	days_to_co	state	initial_
2	TCGA-92-7340-01A-21R-2045-07	TCGA-92-73	TCGA-92-73	TCGA-92-73	TP	Primary sol	TCGA-92-73	1	Not Report	d254d457-f	Primary Tu	Not Report	NA	released	NA
3	TCGA-22-4593-11A-01R-1820-07	TCGA-22-45	TCGA-22-45	TCGA-22-45	NT	Solid Tissue	TCGA-22-45	11	NA	f0217eb5-1	Solid Tissue	NA	NA	released	NA
4	TCGA-O2-A52Q-01A-11R-A26W-(	TCGA-O2-A	TCGA-O2-A	TCGA-O2-A	TP	Primary sol	TCGA-O2-A	1	Not Report	a8d63973-e	Primary Tu	Not Report	2558	released	
5	TCGA-85-8479-01A-11R-2326-07	TCGA-85-84	TCGA-85-84	TCGA-85-84	TP	Primary sol	TCGA-85-84	1	NA	5c454326-1	Primary Tu	NA	NA	released	NA
6	TCGA-22-4594-01A-01R-1201-07	TCGA-22-45	TCGA-22-45	TCGA-22-45	TP	Primary sol	TCGA-22-45	1	NA	27e53228-9	Primary Tu	NA	NA	released	NA
7	TCGA-63-A5MB-01A-11R-A26W-(	TCGA-63-A5	TCGA-63-A5	TCGA-63-A5	TP	Primary sol	TCGA-63-A5	1	Not Report	635a73e9-d	Primary Tu	Not Report	NA	released	
8	TCGA-21-5783-01A-41R-2187-07	TCGA-21-57	TCGA-21-57	TCGA-21-57	TP	Primary sol	TCGA-21-57	1	NA	b655d062-4	Primary Tu	NA	NA	released	NA
9	TCGA-18-4083-01A-01R-1100-07	TCGA-18-40	TCGA-18-40	TCGA-18-40	TP	Primary sol	TCGA-18-40	1	NA	2be9b353-e	Primary Tu	NA	NA	released	NA
10	TCGA-85-A53L-01A-21R-A26W-07	TCGA-85-A5	TCGA-85-A5	TCGA-85-A5	TP	Primary sol	TCGA-85-A5	1	Not Report	e41024f7-1	Primary Tu	Not Report	72	released	
11	TCGA-77-8148-01A-11R-2247-07	TCGA-77-81	TCGA-77-81	TCGA-77-81	TP	Primary sol	TCGA-77-81	1	NA	57a0a87e-7	Primary Tu	NA	NA	released	NA
12	TCGA-77-A5G7-01B-11R-A27Q-07	TCGA-77-A5	TCGA-77-A5	TCGA-77-A5	TP	Primary sol	TCGA-77-A5	1	Not Report	007e0c6f-8	Primary Tu	Not Report	1317	released	
13	TCGA-22-4593-01A-21R-1820-07	TCGA-22-45	TCGA-22-45	TCGA-22-45	TP	Primary sol	TCGA-22-45	1	NA	28375fb3-f	Primary Tu	NA	NA	released	NA
14	TCGA-56-7730-11A-01R-2125-07	TCGA-56-77	TCGA-56-77	TCGA-56-77	NT	Solid Tissue	TCGA-56-77	11	NA	bc60f18a-2	Solid Tissue	NA	NA	released	NA
15	TCGA-68-8250-01A-11R-2296-07	TCGA-68-82	TCGA-68-82	TCGA-68-82	TP	Primary sol	TCGA-68-82	1	NA	a6a1a05e-9	Primary Tu	NA	NA	released	NA
16	TCGA-77-8150-01A-11R-2247-07	TCGA-77-81	TCGA-77-81	TCGA-77-81	TP	Primary sol	TCGA-77-81	1	NA	b85adaf2-0	Primary Tu	NA	NA	released	NA
17	TCGA-66-2793-01A-01R-1201-07	TCGA-66-27	TCGA-66-27	TCGA-66-27	TP	Primary sol	TCGA-66-27	1	NA	daa128eb-b	Primary Tu	NA	NA	released	NA
18	TCGA-33-A4WN-01A-11R-A262-0	TCGA-33-A4	TCGA-33-A4	TCGA-33-A4	TP	Primary sol	TCGA-33-A4	1	Not Report	391639cb-e	Primary Tu	Not Report	327	released	
19	TCGA-34-8456-01A-21R-2326-07	TCGA-34-84	TCGA-34-84	TCGA-34-84	TP	Primary sol	TCGA-34-84	1	NA	a6252c4d-e	Primary Tu	NA	NA	released	NA
20	TCGA-77-A5FZ-01A-31R-A27Q-07	TCGA-77-A5	TCGA-77-A5	TCGA-77-A5	TP	Primary sol	TCGA-77-A5	1	Not Report	5125a40a-e	Primary Tu	Not Report	4283	released	
21	TCGA-34-2596-01A-01R-0851-07	TCGA-34-25	TCGA-34-25	TCGA-34-25	TP	Primary sol	TCGA-34-25	1	Not Report	c07c8c94-1	Primary Tu	Not Report	NA	released	NA
22	TCGA-O2-A5IB-01A-11R-A27Q-07	TCGA-O2-A	TCGA-O2-A	TCGA-O2-A	TP	Primary sol	TCGA-O2-A	1	Not Report	803e1d61-0	Primary Tu	Not Report	2666	released	
23	TCGA-66-2773-01A-01R-1201-07	TCGA-66-27	TCGA-66-27	TCGA-66-27	TP	Primary sol	TCGA-66-27	1	NA	a29977a3-c	Primary Tu	NA	NA	released	NA
24	TCGA-43-5670-01A-21R-2125-07	TCGA-43-56	TCGA-43-56	TCGA-43-56	TP	Primary sol	TCGA-43-56	1	NA	6b829f48-b	Primary Tu	NA	NA	released	NA
25	TCGA-NC-A5HK-01A-11R-A26W-(	TCGA-NC-A	TCGA-NC-A	TCGA-NC-A	TP	Primary sol	TCGA-NC-A	1	Not Report	289898cd-6	Primary Tu	Not Report	1018	released	
26	TCGA-66-2789-01A-01R-0980-07	TCGA-66-27	TCGA-66-27	TCGA-66-27	TP	Primary sol	TCGA-66-27	1	NA	a633a5ca-c	Primary Tu	NA	NA	released	NA
27	TCGA-43-5670-11A-01R-2125-07	TCGA-43-56	TCGA-43-56	TCGA-43-56	NT	Solid Tissue	TCGA-43-56	11	NA	52f75eb2-7	Solid Tissue	NA	NA	released	NA
28	TCGA-85-A50Z-01A-21R-A262-07	TCGA-85-A5	TCGA-85-A5	TCGA-85-A5	TP	Primary sol	TCGA-85-A5	1	Not Report	d4558b49-d	Primary Tu	Not Report	143	released	
29	TCGA-22-1017-01A-01R-0692-07	TCGA-22-10	TCGA-22-10	TCGA-22-10	TP	Primary sol	TCGA-22-10	1	NA	0c12ba81-c	Primary Tu	NA	NA	released	NA
30	TCGA-85-A4JB-01A-51R-A262-07	TCGA-85-A4	TCGA-85-A4	TCGA-85-A4	TP	Primary sol	TCGA-85-A4	1	Not Report	d767ee8b-e	Primary Tu	Not Report	140	released	
31	TCGA-33-AASB-01A-11R-A405-07	TCGA-33-A	TCGA-33-A	TCGA-33-A	TP	Primary sol	TCGA-33-A	1	Not Report	0193b241-1	Primary Tu	Not Report	4286	released	
32	TCGA-NK-A5CR-01A-11R-A26W-(	TCGA-NK-A	TCGA-NK-A	TCGA-NK-A	TP	Primary sol	TCGA-NK-A	1	Not Report	8781a00c-0	Primary Tu	Not Report	1961	released	
33	TCGA-34-5234-01A-01R-1635-07	TCGA-34-52	TCGA-34-52	TCGA-34-52	TP	Primary sol	TCGA-34-52	1	NA	8620b5a0-e	Primary Tu	NA	NA	released	NA
34	TCGA-85-6175-01A-11R-1820-07	TCGA-85-61	TCGA-85-61	TCGA-85-61	TP	Primary sol	TCGA-85-61	1	NA	4da87998-a	Primary Tu	NA	NA	released	NA
35	TCGA-77-7338-01A-11R-2045-07	TCGA-77-73	TCGA-77-73	TCGA-77-73	TP	Primary sol	TCGA-77-73	1	Not Report	7112878a-c	Primary Tu	Not Report	NA	released	NA
36	TCGA-NK-A7XE-01A-12R-A405-07	TCGA-NK-A	TCGA-NK-A	TCGA-NK-A	TP	Primary sol	TCGA-NK-A	1	Not Report	442c64f8-2	Primary Tu	Not Report	50	released	
37	TCGA-56-8201-11A-01R-2247-07	TCGA-56-82	TCGA-56-82	TCGA-56-82	NT	Solid Tissue	TCGA-56-82	11	NA	160cfdb8-0	Solid Tissue	NA	NA	released	NA
38	TCGA-63-A5MY-01A-11R-A26W-(	TCGA-63-A5	TCGA-63-A5	TCGA-63-A5	TP	Primary sol	TCGA-63-A5	1	Not Report	b180fdf0-9	Primary Tu	Not Report	NA	released	
39	TCGA-56-7222-01A-11R-2045-07	TCGA-56-72	TCGA-56-72	TCGA-56-72	TP	Primary sol	TCGA-56-72	1	Not Report	f0f4aa30-1	Primary Tu	Not Report	NA	released	NA

CHAPTER 4

METHODOLOGY

4.1 Data Collection

Gene expression data and clinical data were collected from the TCGA data portal from American patients. [21]

4.2 Data Processing

Data preprocessing is a critical but often overlooked step in the data mining process. Data collection is typically an uncontrolled process, resulting in out-of-range values, for example, impossible data combinations (for example, gender: male; pregnant: yes), missing values, and so on. Data analysis that has not been thoroughly screened for such issues can yield misleading results. Thus, before running an analysis, the representation and quality of the data must be prioritized. Data preprocessing consists of data preparation, which includes data integration, cleaning, normalization, and transformation, as well as data reduction tasks like feature selection, instance selection, and discretization. After a reliable chaining of data preprocessing tasks, the expected result is a final dataset that can be considered correct and useful for further data mining algorithms. [22]

4.3 Data Transforming

The data is first normalized and transformed. Data scale issues must be considered in how experiments are carried out, as well as potential problems during data collection. Each tumor sample should ideally have a similar distribution of gene expression values. Differences in tumor samples must be corrected in a systematic manner. Box plots are used to determine whether such differences exist. To avoid cluttering the figure, only the first 50 tumor samples are plotted.[23]

4.4 Data Visualization

Data visualization is a broad term that refers to any effort to help people understand the importance of data by presenting it in a visual format. Data visualization is the graphical presentation of quantitative information. In other words, data visualizations convert large and small data sets into visuals that the human brain can understand and process more easily. It can be used to uncover previously unknown facts and trends. When communication, data science, and design come together, good data visualizations emerge. [24]

4.5 Missing Value

Missing values are typically represented in databases as "NULL" values or as empty cells in spreadsheet tables. Some flat-file formats use different symbols for missing values, such as the "?" symbol in arff files. These types of missing values are easily detectable. However, missing values can appear as outliers or incorrect data (i.e. outside of boundaries). These data must be removed prior to the intended analysis and are much more difficult to discover. [25]

4.6 Data Normalization

Data normalization is an important pre-processing, mapping, and scaling method that aids in the accuracy of forecasting and prediction models. Using this method, the current data range is transformed into a new, standardized range. When it comes to bringing disparate prediction and forecasting techniques into harmony, normalization is critical. [26]

4.7 Cross Validation

Cross-validation works by randomly splitting the data into k-folds of samples. A 5-fold cross-validation with 100 data points, for example, would provide 5 folds, each with 20 data points. The model is then built and the errors are estimated five times. Four groups are pooled each time (resulting in 80 data points) and used to train the model. The test error is then estimated using the fifth group of 20 observations that were not utilized to develop the model. In the case of 5-fold cross-validation, the error estimates will be 5 and may be averaged to get a more robust estimate of the test error.To equate the k to the number of Data Points field, the number of tumor samples, is an extreme instance of k-fold cross-validation. LOOCV stands for leave-one-out cross-validation. It may be superior to k-fold cross-validation, but training numerous models takes significantly longer when the number of data points is big. For all machine learning algorithms, the caret package has cross-validation features.[27]

4.8 K-Nearest Neighbors

K-closest neighbors (KNN) classification is a kind of nearest neighbor classification based on the notion that the designs closest to a goal pattern x, for which we look for the label and provide essential label information. KNN labels the majority of the K-nearest patterns in data space with a class label. To do this, we must be able to define a similarity measure in data space. It is permissible to use the Minkowski metric (p-norm) in Rq.

$$f_{\mathrm{KNN}}(\mathbf{x}') = \begin{cases} 1 & \text{if } \sum_{i \in N_K(\mathbf{x}')} y_i \geq 0 \\ -1 & \text{if } \sum_{i \in N_K(\mathbf{x}')} y_i < 0 \end{cases}$$

This is equivalent to the Euclidean distance for p = 2. In other data spaces, appropriate distance functions, such as the Hamming distance in Bq, must be chosen. When it comes to binary classification, the label set Y = {1, 1} is used, and KNN is defined as

$$f_{\mathrm{KNN}}(\mathbf{x}') = \begin{cases} 1 & \text{if } \sum_{i \in \mathcal{N}_K(\mathbf{x}')} y_i \geq 0 \\ -1 & \text{if } \sum_{i \in \mathcal{N}_K(\mathbf{x}')} y_i < 0 \end{cases}$$

with neighborhood size K and with the set of indices NK(x') of the K-nearest patterns.[28]

4.9 Decision Tree (Random Forest)

Decision trees are a common solution for a variety of machine learning applications, owing to their great interpretability. A decision tree is a collection of filters applied to predictor variables. A class prediction is the result of the sequence of filters. Each filter is a binary yes/no question, resulting in bifurcations in the sequence of filters and a tree-like structure. The kind of predictor variables influences the filters. If the variables are categorical, such as gender, the filters might include "is gender female" questions. The tree-fitting approach selects the optimum variables at decision nodes based on how effectively they divide the samples into classes after the decision node is applied.Decision trees can deal with both category and numeric predictor variables, are simple to comprehend, and can manage missing information. Tree-based machine learning algorithms come in a variety of flavors. Most algorithms, on the other hand, build decision nodes from the top down. Based on how homogenous the sample sets are after the split, they choose the appropriate variables to utilize in decision nodes. "Gini impurity" is one measure of homogeneity. After the split, this metric is calculated for each subgroup and then totaled as a weighted average. In a two-class problem, a decision node that splits the data perfectly will have a gini impurity of 0, while a node that splits the data into a subset with 50% class A and 50% class B will have an impurity of 0.5%. The Gini impurity, IG(p), of a set of samples with known class labels for K classes is calculated as follows, where Pi represents the probability of observing class i in the subset:[29]

$$I_G(p) = \sum_{i=1}^{K} p_i(1 - p_i) = \sum_{i=1}^{K} p_i - \sum_{i=1}^{K} {p_i}^2 = 1 - \sum_{i=1}^{K} {p_i}^2$$

4.10 Logistic regression and regularization

Logistic regression is a statistical approach for modeling a binary response variable using predictor variables. Although designed for two-class or binary answer issues, this approach may be used to multiclass situations as well. However, because our tumor sample data represents a binary response or two-class problem, we shall skip over the multiclass scenario in this chapter. Logistic regression is conceptually very similar to linear regression, and it can be viewed as a "maximum likelihood estimation" problem in which we try to find statistical parameters that maximize the likelihood that the observed data is drawn from the statistical distribution of interest.. This is also similar to the generic cost/loss function approach seen in supervised machine learning methods. When dealing with binary response variables, a basic linear regression model such as Yi~ β_0+$\beta_1 X1$, would be a poor choice because it can easily generate values outside of the 0 to 1 boundary.The first step in meeting this condition is to rephrase the problem. If y_i can only be 0 or 1, we may define it as a realization of a random variable with probability p_i and 1 p_i , respectively. We may write the issue as $p_i \sim \beta_0 + \beta_1 x_1$ instead of predicting the binary variable because this random variable follows the Bernoulli distribution. However, our original problem remains: simple linear regression will still provide results that are outside of the 0 and 1 range. The logistic equation displayed below is a model that fulfills the boundary condition. [30]

$$p_i = \frac{e^{(\beta_0 + \beta_1 x_i)}}{1 + e^{(\beta_0 + \beta_1 x_i)}}$$

This equation can be linearized using the following transformation.

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

The left side is referred to as the logit, which stands for "logistic unit". It is also called the log odds. Our model generates log scale values, which can be transformed to the 0−1 range using the logistic equation. The best parameter estimates maximize the likelihood that the statistical model will produce the observed data. We can think of this fitting as applying a probability distribution to observed data. The parameters of the probability distribution should maximize the likelihood that the observed data came from the distribution under consideration. If we were using a Gaussian distribution, we would adjust the mean and variance parameters until the observed data was more likely to come from that particular Gaussian distribution.

4.11 Gradient Descent (XGboost)

Gradient boosting is a prediction model that employs an ensemble of decision trees, similar to random forest. However, the decision trees are added sequentially, so these models are also known as "Multiple Additive Regression Trees (MART)" (Friedman and Meulman 2003). In general, "boosting" refers to an iterative learning approach in which each new model attempts to focus on data points where the previous ensemble of simple models did not perform well. Gradient boosting is an improvement over that, in which each new model attempts to focus on the previous model's residual errors. As in random forests, many trees are grown, but in this case, the trees are grown sequentially, with each tree focusing on improving the shortcomings of the previous trees. **Figure 5** illustrates this concept. One of the most popular gradient boosting algorithms is XGboost, which stands for "extreme gradient boosting" (Chen and Guestrin 2016). However, this flexibility has advantages; methods based on XGboost have won numerous machine learning competitions (Chen and Guestrin 2016).**[xx]**

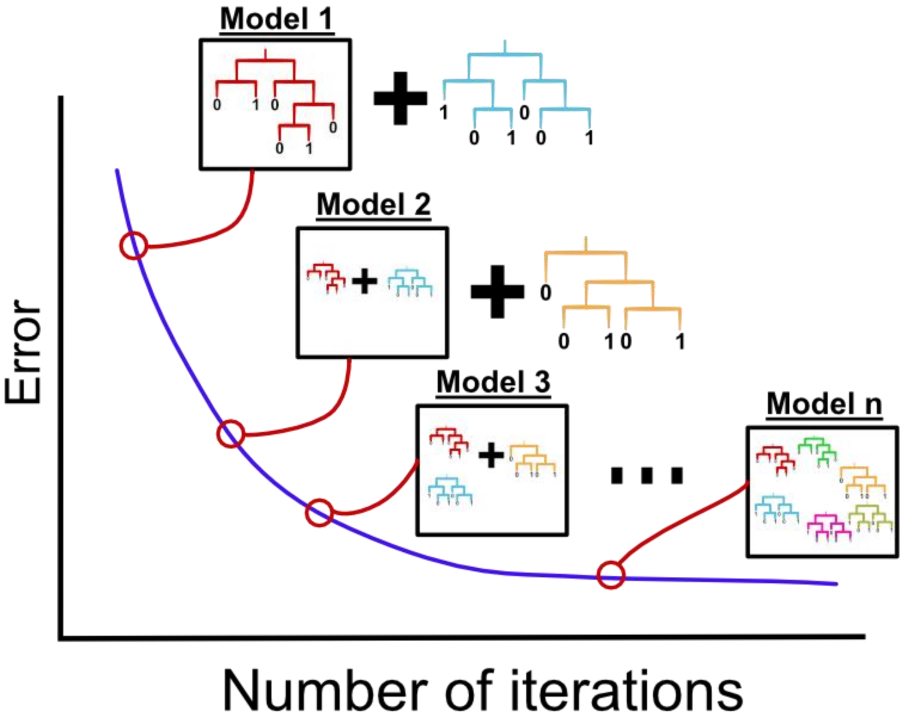


FIGURE 5: Gradient boosting machines concept. Individual decision trees are built sequentially in order to fix the errors from the previous trees.

4.12 Support Vector Machines (SVM)

Support vector machines (SVM) gained popularity in the 1990s due to the algorithm's efficiency and performance (Boser, Guyon, and Vapnik 1992). The method works by determining the best decision boundary for dividing the data points into distinct groups (or classes), and then predicting the class of fresh observations based on this separation boundary. Depending on the circumstances, the various groups may be separated by a linear straight line or a non-linear boundary line or plane. If you look at the k-NN decision boundaries in **Figure 6**, you will notice that they are not linear. SVM can work with both linear and nonlinear decision boundaries.

First, SVM may translate the data to higher dimensions using linear decision boundaries. This is accomplished by applying mathematical functions known as "kernel functions" on the predictor variable space.

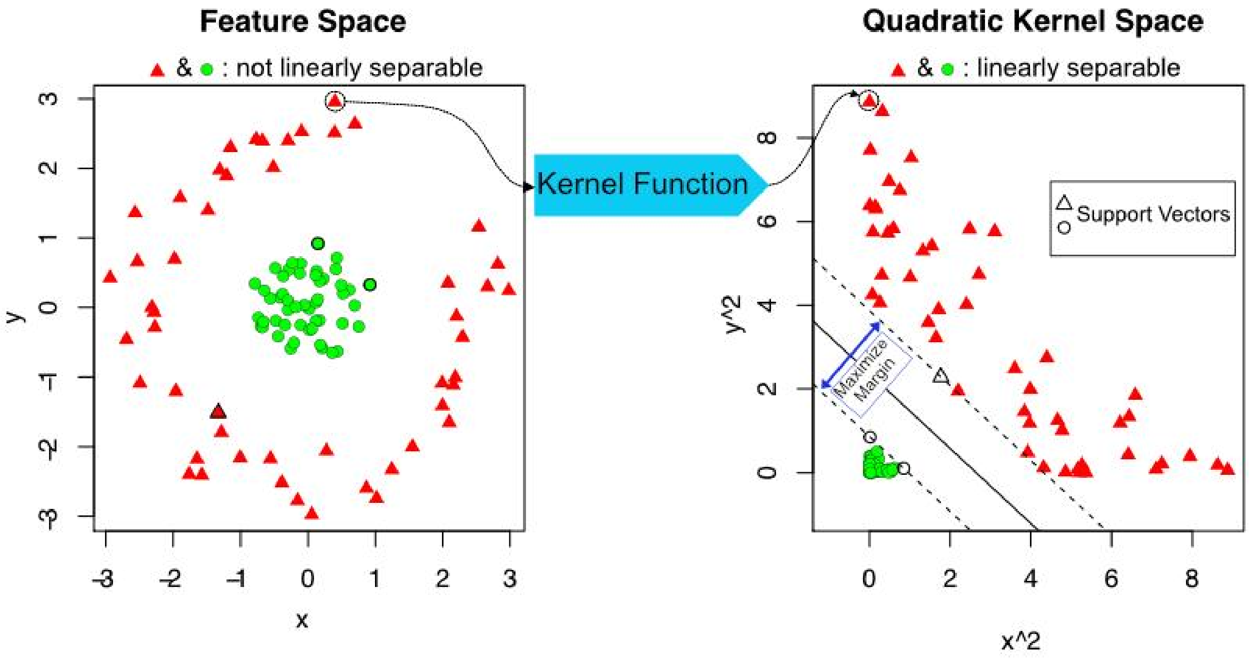


FIGURE 6: Support vector machine concept. With the help of a kernel function,points in feature space are mapped to higher dimensions where linear separation is possible.

Second, SVM seeks not just a decision boundary, but also the border with the biggest buffer zone on both sides of the boundary. A border with a big buffer, or "margin," as it is officially termed, would perform better for fresh data points that were not utilized in model training (margin is highlighted in **Figure-6**). Another important aspect of the algorithm is that SVM determines the decision boundary solely based on the "landmark" data points, also known as "support vectors". These are the points that are closest to the decision boundary and more difficult to classify. By keeping track of such points only for decision boundary creation, the algorithm's computational complexity is reduced.[31]

4.13 Neural Networks

Another famous machine learning technology that has lately gained prominence is neural networks. The algorithm's previous iterations were popularized in the 1980s and 1990s. Neural networks, like SVM, have the benefit of being able to model non-linear decision boundaries. The core concept behind neural networks is to integrate predictor variables to describe the response variable as a nonlinear function. Depending on the number of layers in the network, input variables pass through various layers that combine the variables, alter those combinations, and recombine outputs.In the conceptual example in **Figure-7** the input nodes receive predictor variables and make linear combinations of them in the form of ∑(wixi+b). Simply put, the variables are multiplied with weights and summed up. This is what we call "linear combination".

These values are then sent into another layer termed the hidden layer, which applies an activation function to the sums. And, assuming we're working on a classification method, these findings are sent into an output node that produces class probabilities. There might be many more secret layers that integrate the output of the hidden levels preceding them. Finally, the method has a cost function that is identical to the logistic regression cost function, but it now has to estimate all of the weight parameters: wi. Because of the amount of parameters to be estimated, this is a more harder issue than logistic regression, however neural networks can also fit complex functions due to their parameter space flexibility.

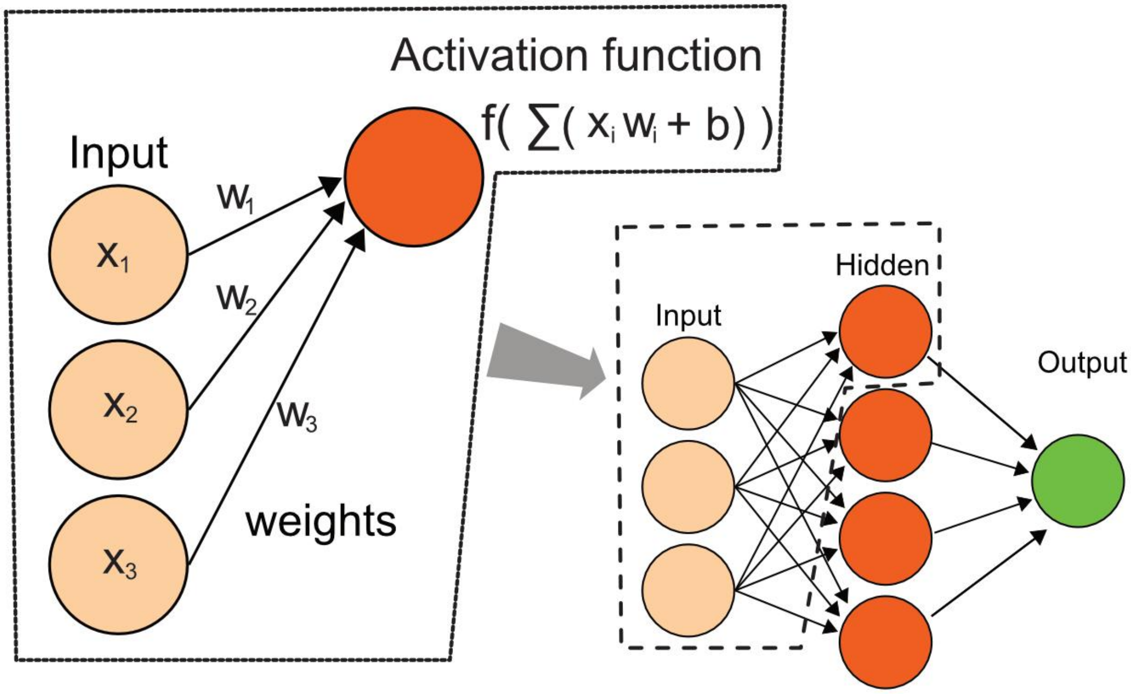


FIGURE 7: A basic neural network diagram, with their combinations passing via hidden layers and being mixed again for output. The network is fed predictor variables, and weights are modified to maximize the cost function.

In practice, the number of nodes in the hidden layer (size) and some weight regularization can be used to prevent overfitting. This is known as the calculated (decay) parameter controls for overfitting.[32]

CHAPTER 5

RESULT AND DISCUSSION

K-Nearest Neighbors

Process	Training Accuracy	Test Accuracy
K Neighbor Neighbors	0.9493	0.9478

Table-1: K-Nearest Neighbors Training & Testing Accuracy

In the presented machine learning analysis, the K Nearest Neighbors (KNN) algorithm was employed to model a dataset. The model's performance was evaluated based on training and test accuracies, with the training accuracy yielding a value of 0.9493 or 94.93%, and the test accuracy demonstrating a closely comparable value of 0.9478 or 94.78%. These accuracy metrics serve as indicators of the model's ability to generalize well to new, unseen data, with the high values suggesting a robust performance.**Figure-8** displays the top importance genes identified through the K Nearest Neighbors Algorithm. The determination of gene importance was facilitated by the caret::varImp() function, which assesses the significance of features in the model. Subsequently, the importance values were visualized using the plot() function. The top 25 genes, as revealed by this analysis, play a crucial role in influencing the model's predictions.This information not only provides insights into the predictive capacity of the KNN model but also highlights the specific genetic factors that contribute significantly to the model's decision-making process. The identification and understanding of these top 25 important genes can be instrumental in gaining insights into the biological or clinical relevance of the studied dataset, contributing to the interpretability of the machine learning model's outcomes.

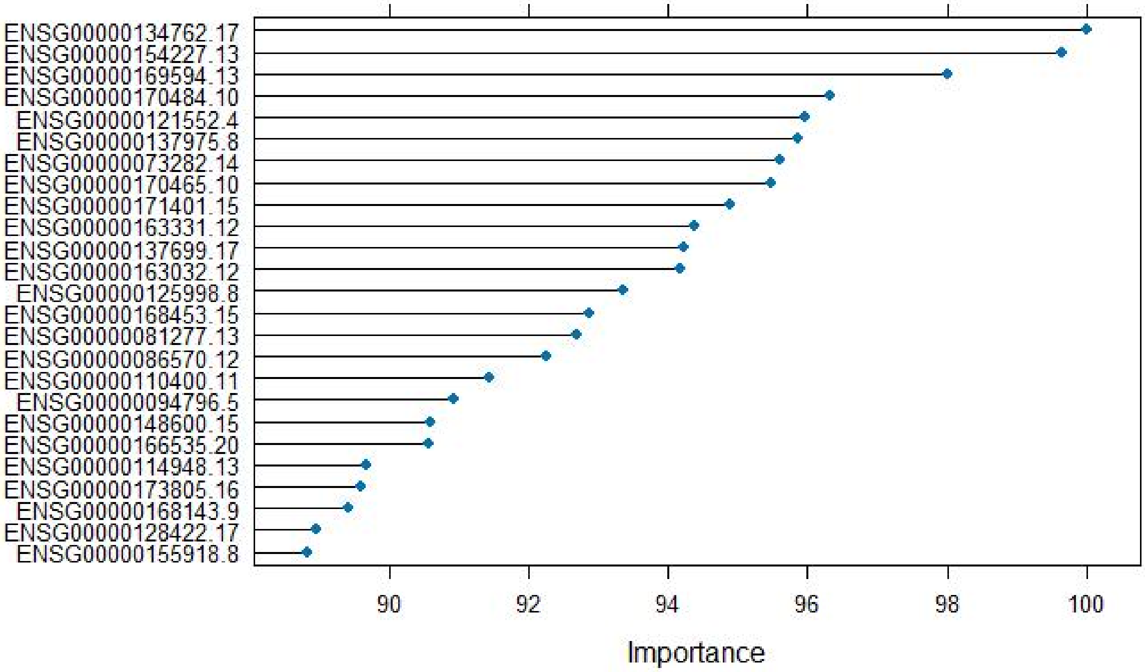


FIGURE 8: Top 25 important variables for the K-Nearest Neighbors(KNN).

Decision Tree (Random Forest)

Process	**Training Accuracy**	**Test Accuracy**
Decision Tree	1	0.9478261

Table-2: Decision Tree (Random Forest) Training & Testing Accuracy

In this machine learning analysis, a Decision Tree model implemented as a Random Forest was utilized. The model exhibited perfect training accuracy, achieving a score of 1, indicating an ideal fit to the training data. However, the test accuracy, measuring the model's performance on new data, was slightly lower but still commendable at 94.78261%. While the high training accuracy suggests a strong grasp of the training set, attention should be given to potential overfitting, where the model may struggle to generalize to unseen data. The use of Random Forest, an ensemble method, enhances the model's robustness by aggregating insights from multiple trees. Further investigation and fine-tuning may be needed to optimize the model's generalization capabilities beyond the training dataset.Random forests include variable importance metrics by default. One metric is similar to the "variable dropout metric" in that the predictor variables are permuted. OOB samples are used in this case, and the variables are permuted one at a time. Every time, the permuted variable samples are fed into the network, and the decrease in accuracy is measured. The variables can be ranked using this number.Random forests include variable importance metrics by default. One metric is similar to the "variable dropout metric" in that the predictor variables are permuted. OOB samples are used in this case, and the variables are permuted one at a time. Every time, the permuted variable samples are fed into the network, and the decrease in accuracy is measured. The variables can be ranked using this number.Let's plot the shift-based importance metric below. This metric is computed during the execution of the preceding model. The importance values were accessed using the caret::varImp() function and plotted using the plot() function; the results are shown in **Figure 9**.

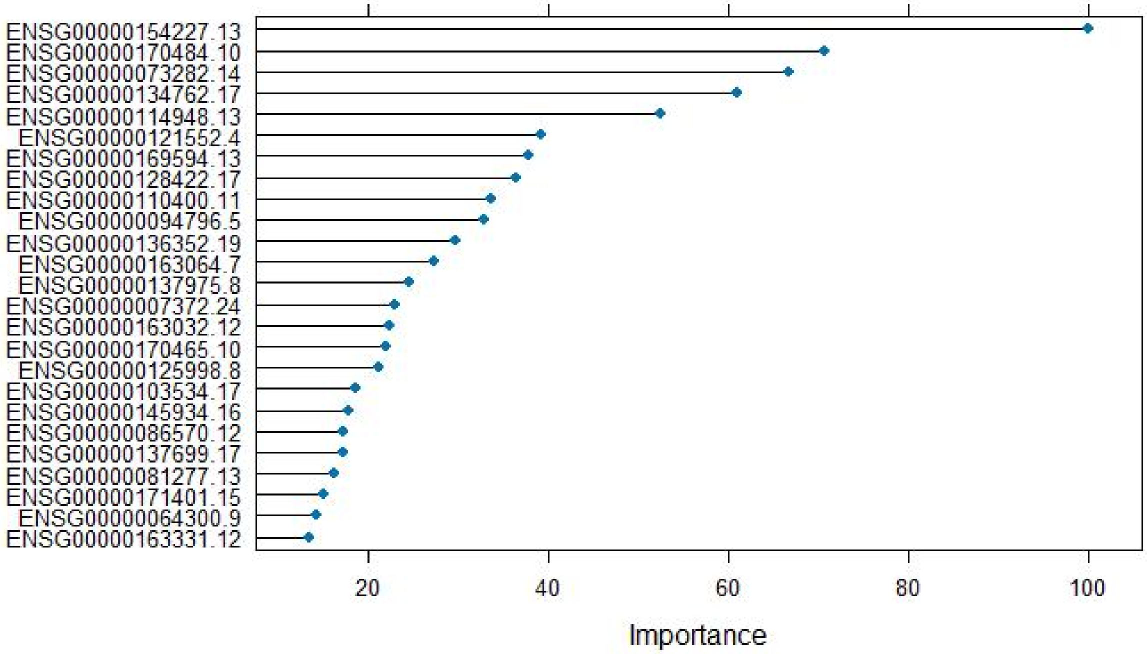


FIGURE-9: Top 25 important variables based on permutation-based method for the random forest classification.

Regularization

Process	**Training Accuracy**	**Test Accuracy**
Regularization	0.9319307	0.9130435

Table-3: K-Nearest Neighbors Training & Testing Accuracy

In this machine learning analysis, Regularization techniques were employed to train a model. The training accuracy yielded a respectable score of 0.9319307 or 93.19307%, indicating the model's proficiency on the training data. The test accuracy, assessing the model's performance on new and unseen data, was slightly lower but still substantial at 91.30435%. Regularization methods are crucial for preventing overfitting and enhancing a model's ability to generalize. The slightly lower test accuracy suggests a balance between fitting the training data well and maintaining good performance on external datasets. Overall, the use of Regularization underscores a thoughtful approach to model training, emphasizing a robust and generalizable performance beyond the training set. The variable importance of penalized regression, especially for lasso and elastic net, is surprising. These methods, as previously stated, will set the regression coefficient to zero for variables that are irrelevant. It provides a system for selecting and ranking important variables. A method for ranking predictor variables based on the size of regression coefficients; however, if the data is not normalized, different scales will be obtained for different variables. In this case, the data has been normalized, and we know that the variables had the same scale prior to training. This fact can be used to sort them according to their regression coefficients. The caret::variImp() function uses coefficients to rank the variables from the elastic net model.We'll plot the top 25 most important variables below, normalized to the importance of the most important variable.

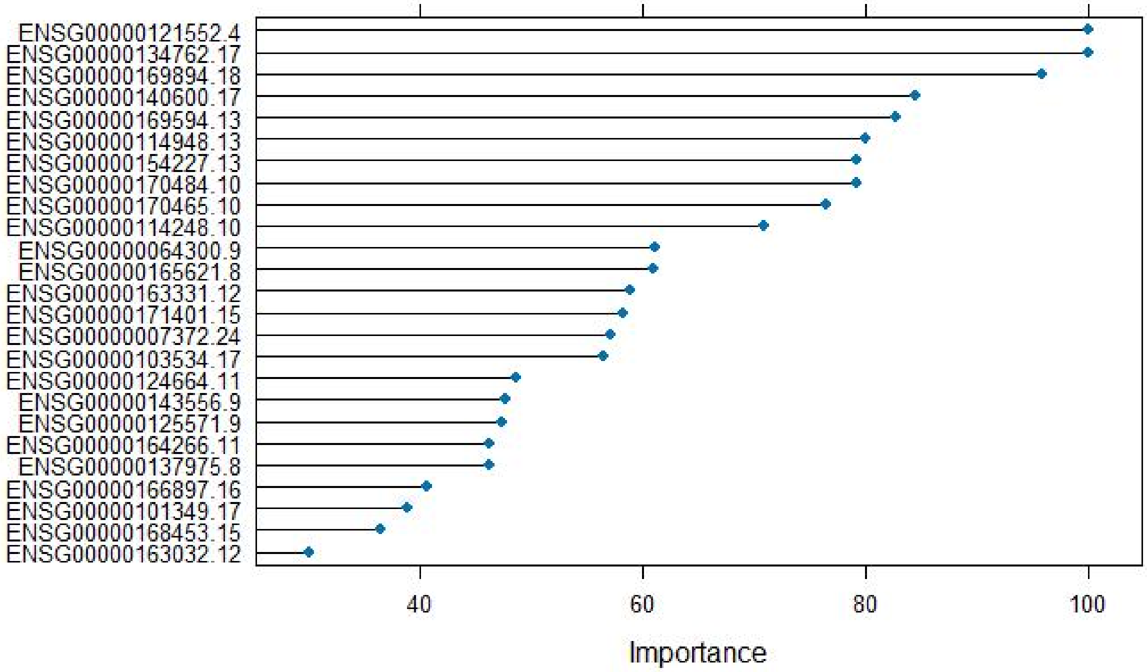


FIGURE 10: Top 25 important Variable importance metric for elastic net. This metric uses regression coefficients as importance.

Gradient Descent (XGboost)

Process	**Training Accuracy**	**Test Accuracy**
Gradient Descent (XGboost)	1	0.9478261

Table-4: Gradient Descent (XGboost) Training & Testing Accuracy

In this machine learning analysis, the Gradient Descent algorithm, specifically implemented as XGBoost, was employed to model a dataset. Notably, the training accuracy achieved a perfect score of 1, indicating an optimal fit to the training data. However, it's essential to approach perfect training accuracy with caution, as it may suggest potential overfitting, where the model may struggle to generalize to new, unseen data. The test accuracy, a crucial metric gauging the model's performance on new data, was found to be 0.9478261 or 94.78261%, demonstrating a strong capability to generalize beyond the training set. The use of XGBoost, a gradient boosting algorithm, typically enhances predictive performance by iteratively improving the model's accuracy. **Figure-11** visually presents the top importance genes identified through the Gradient Descent (XGBoost) Algorithm. The importance values were accessed using the caret::varImp() function and effectively plotted using the plot() function. This visualization provides insights into the genetic factors that significantly influence the model's decision-making process. The identification of the top 25 important genes not only contributes to the interpretability of the model's outcomes but also sheds light on the biological or clinical relevance of these specific genetic features. Overall, the utilization of Gradient Descent with XGBoost underscores a powerful approach to predictive modeling, balancing high training accuracy with strong generalization capabilities on new and diverse datasets.

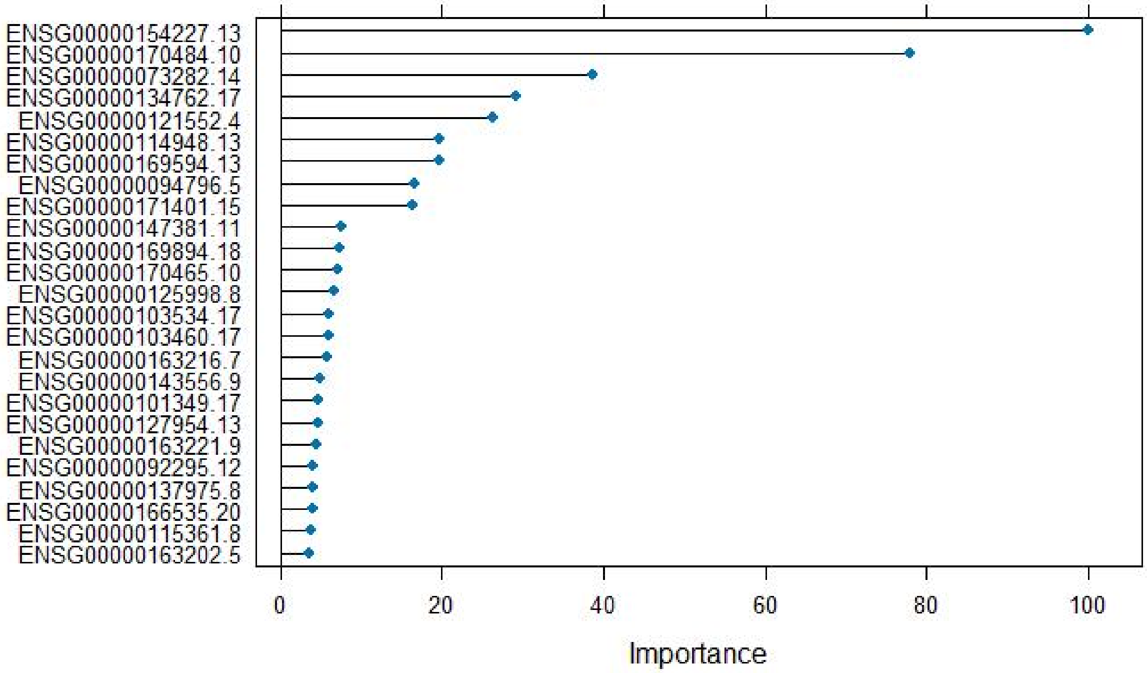


FIGURE 11: Top 25 important variables for the Gradient Descent (XGboost).

Support Vector Machine(SVM)

Process	**Training Accuracy**	**Test Accuracy**
Support Vector Machine	0.519802	0.5217391

Table-5: Support Vector Machine(SVM) Training & Testing Accuracy

In this machine learning analysis, the Support Vector Machine (SVM) algorithm was employed to model a given dataset. The training accuracy was determined to be 0.519802 or 51.9802%, while the test accuracy was slightly higher at 0.5217391 or 52.17391%. The relatively modest accuracy values indicate that the SVM model may face challenges in accurately classifying instances within both the training and test datasets. **Figure-12** presents a visual representation of the top importance genes identified through the Support Vector Machine (SVM) Algorithm. The importance values associated with these genes were accessed using the caret::varImp() function and effectively visualized using the plot() function. Understanding the significance of these genes is crucial for interpreting the SVM model's decision boundaries and gaining insights into the features that contribute most to its predictive capabilities. The fact that SVM achieved similar accuracies on both training and test sets suggests a degree of consistency in its performance across different data subsets. However, the relatively low accuracy values highlight the need for further exploration and potential refinement of the model, such as parameter tuning or feature engineering, to improve its predictive accuracy. Overall, the application of SVM underscores the importance of careful model evaluation and refinement to enhance its performance on the given dataset.

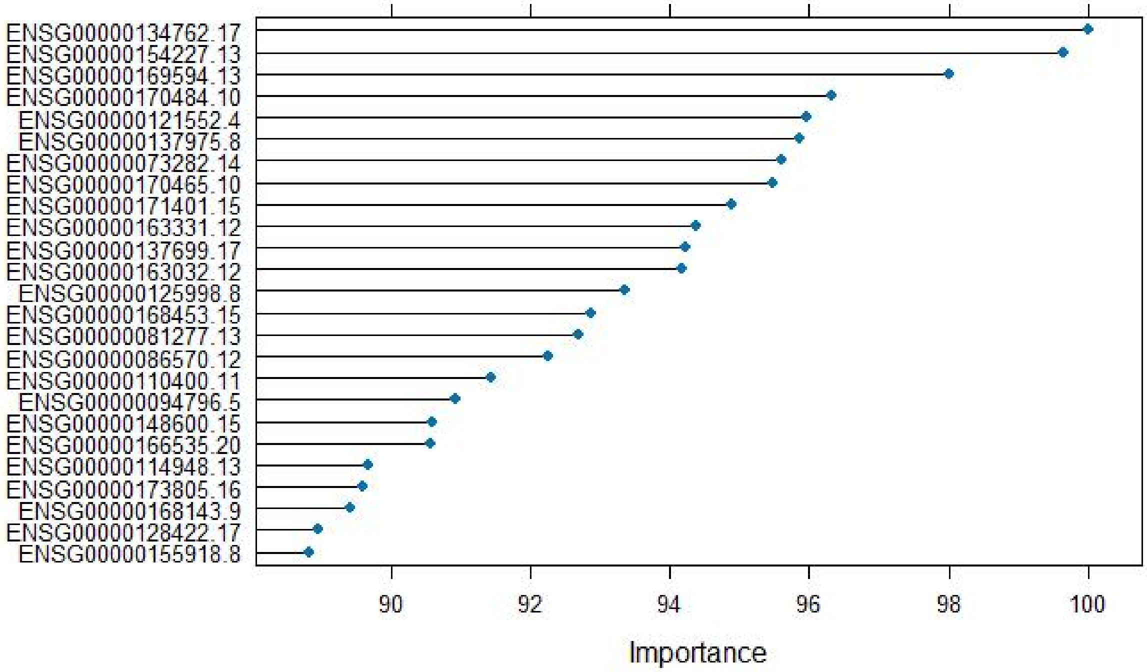


FIGURE 12: Top 25 important variables for the Support Vector Machine(SVM).

Neural Network

Process	Training Accuracy	Test Accuracy
Neural Network	0.9777228	0.9507246

Table-6: Neural Network Training & Testing Accuracy

In this machine learning analysis, a Neural Network algorithm was employed to model a given dataset. The training accuracy was determined to be 0.9777228 or 97.77228%, indicating a highly accurate fit to the training data. The test accuracy, measuring the model's performance on new and unseen data, was slightly lower but still substantial at 0.9507246 or 95.07246%. The high training accuracy suggests that the neural network has effectively learned patterns within the training set, but it's essential to ensure that this performance extends to diverse datasets. **Figure-13** visually represents the top importance genes identified through the Neural Network Algorithm, with importance values accessed using the caret::varImp() function and plotted using the plot() function. This visualization provides valuable insights into the genes that significantly influence the neural network's decision-making process, contributing to the interpretability of its outcomes. The combination of a strong training accuracy and a commendable test accuracy suggests that the neural network model exhibits robust generalization capabilities. However, further scrutiny and validation may be necessary to assess its performance on various datasets and potential fine-tuning to optimize its predictive accuracy. Overall, the application of Neural Network algorithms demonstrates a powerful capacity for capturing complex relationships within the data, with the importance of genes shedding light on the biological or clinical relevance of specific genetic features in the context of the model's predictions.

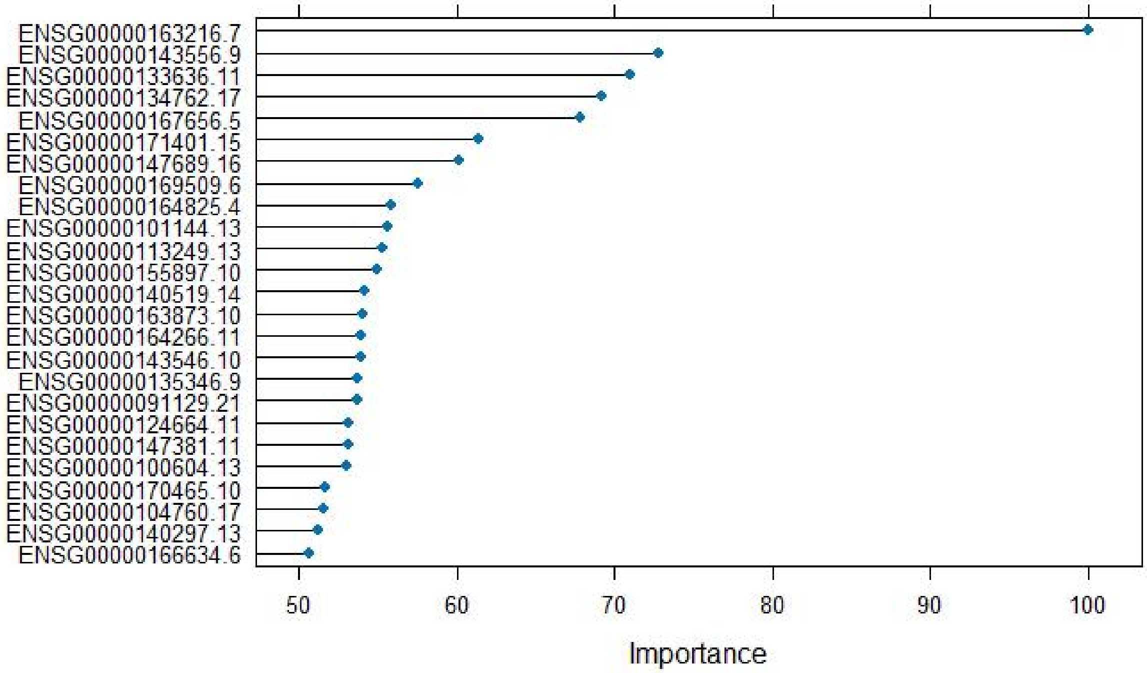
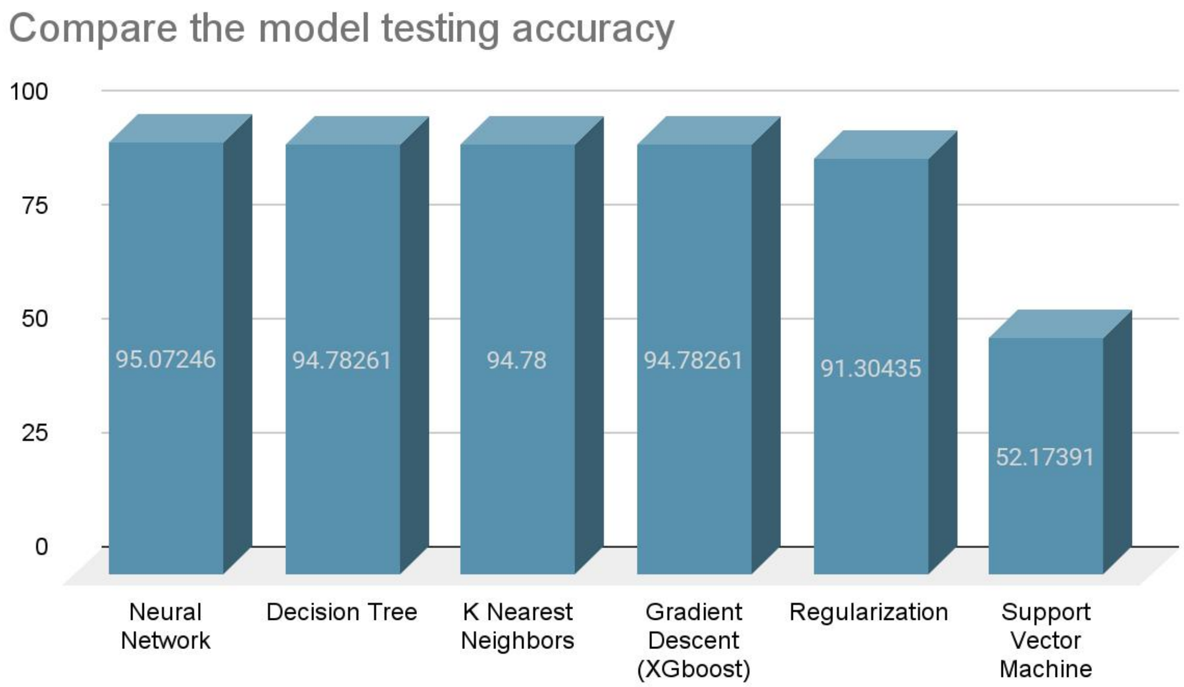


FIGURE 13: Top 25 important variables for the Neural Network.



In this comparative analysis of various machine learning algorithms for predicting subtypes of lung cancer using gene expression and clinical data, several models were evaluated based on their accuracy metrics. The K Nearest Neighbors (KNN) algorithm, Decision Tree algorithm, and Gradient Descent (XGBoost) algorithm demonstrated remarkably similar accuracies, with all three achieving around 94.78%. This suggests that these methods perform consistently well in capturing patterns within the dataset.Contrastingly, the Regularization algorithm yielded a slightly lower accuracy of 91.30435%, implying that its approach to preventing overfitting may result in a trade-off with predictive accuracy. Notably, the Neural Network algorithm outperformed other models, achieving a higher accuracy of 95.07246%. This suggests that the neural network's ability to capture intricate relationships within the data contributes to superior predictive performance in this context.On the other hand, the Support Vector Machine (SVM) algorithm exhibited a significantly lower accuracy of 52.17391%, indicating challenges in accurately classifying the subtypes of lung cancer based on the given features. In summary, these findings suggest that, for the specific task of predicting lung cancer subtypes, the Neural Network algorithm proves to be the most effective among the evaluated methods. This conclusion underscores the importance of selecting an algorithm tailored to the characteristics of the dataset and task at hand, with the neural network demonstrating promising results for this particular application.

CHAPTER 5

CONCLUSION

In this comprehensive study, diverse machine learning and deep learning approaches were employed to predict subtypes of lung cancer, including Neural Network, Decision Tree, K Nearest Neighbors (KNN), Gradient Descent (XGBoost), Regularization, and Support Vector Machine. The evaluation of these models revealed varying levels of accuracy, with the Neural Network demonstrating the highest performance at 95.07246%, followed closely by Decision Tree and KNN with accuracies of 94.78261% and 94.78%, respectively. The Gradient Descent (XGBoost) and Regularization methods achieved accuracies of 94.78261% and 91.30435%, while the Support Vector Machine exhibited a lower accuracy of 52.17391%.The study leveraged publicly available gene expression data and clinical data obtained from the TCGA data portal, focusing specifically on American cancer patients. The use of real-world data enhances the applicability of the findings to clinical settings. Notably, the dataset served as a foundation for the evaluation of diverse machine learning and deep learning models, providing insights into their effectiveness in subtype prediction for lung cancer.As part of future work, the study aims to expand its scope by incorporating multiple cancer datasets and employing an array of advanced deep learning models. This expansion seeks to enhance the generalizability of the models across different cancer types and populations, with the ultimate goal of achieving highly improved model accuracy. The comprehensive approach and future plans outlined in this study contribute to the ongoing efforts in leveraging machine learning and deep learning for more accurate and robust cancer subtype predictions.

CHAPTER 7

REFERENCE

[1]. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; 2020 (https://gco.iarc.fr/today, accessed February 2021).

[2]. Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2019 global survey. Geneva: World Health Organization; 2020.

[3]. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.

[4]. Sun Y, Yao J, Yang L, Chen R, Nowak NJ, Goodison S. Computational approach for deriving cancer progression roadmaps from static sample data. Nucleic Acids Res. 2017;45(9):e69.

[5]. Curtis, Christina, et al. "The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups." Nature 486.7403 (2012): 346-352.

[6]. Parker, Joel S., et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes." *Journal of clinical oncology* 27.8 (2009): 1160.

[7]. Deep learning approach for cancer subtype classification using high-dimensional gene expression data.

[8]. Baizhumanova, Ainur, and Junichi Sakamoto. "Cancer in Kazakhstan: Present situation on Cancer." *Annals of Cancer Research and Therapy* 18.2 (2010): 65-68.

[10]. Coudray, Nicolas, et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." Nature medicine 24.10 (2018): 1559-1567.

[11]. Shen, Jiquan, et al. "Deep learning approach for cancer subtype classification using high-dimensional gene expression data." BMC bioinformatics 23.1 (2022): 1-17.

[12]. Cheng, Nitao, et al. "Prediction of lung cancer metastasis by gene expression." Computers in Biology and Medicine 153 (2023): 106490.

[13]. Liu, Suli, and Wu Yao. "Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection." BMC bioinformatics 23.1 (2022): 175.

[14]. Ramos, Bernardo, et al. "An interpretable approach for lung cancer prediction and subtype classification using gene expression." 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021

[15]. Chen, Joe W., and Joseph Dhahbi. "Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods." Scientific reports 11.1 (2021): 13323.

[16]. Rajpal, Sheetal, et al. "Deep learning based model for breast cancer subtype classification." arXiv preprint arXiv:2111.03923 (2021).

[17]. Kim, Bong-Hyun, Kijin Yu, and Peter CW Lee. "Cancer classification of single-cell gene expression data by neural network." Bioinformatics 36.5 (2020): 1360-1366.

[18]. Su, Ran, et al. "Identification of expression signatures for non-small-cell lung carcinoma subtype classification." Bioinformatics 36.2 (2020): 339-346.

[19]. Xu, Jing, et al. "A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data." IEEE Access 7 (2019): 22086-22095.

[20]. Wistuba, Ignacio I., and Adi F. Gazdar. "Lung cancer preneoplasia." Annu. Rev. Pathol. Mech. Dis. 1 (2006): 331-348.

[21]. https://portal.gdc.cancer.gov/

[22]. García, Salvador, Julián Luengo, and Francisco Herrera. Data preprocessing in data mining. Vol. 72. Cham, Switzerland: Springer International Publishing, 2015.

[23]. Manikandan, S. "Data transformation." Journal of Pharmacology and Pharmacotherapeutics 1.2 (2010): 126.

[24]. Islam, Mohaiminul, and Shangzhu Jin. "An overview of data visualization." 2019 International Conference on Information Science and Communications Technologies (ICISCT). IEEE, 2019.

[25]. Kaiser, Jiří. "Dealing with Missing Values in Data." Journal Of Systems Integration (1804-2724) 5.1 (2014).

[26]. Ali, Peshawa Jamal Muhammad, et al. "Data normalization and standardization: a technical report." Mach Learn Tech Rep 1.1 (2014): 1-6.

[27]. Browne, Michael W. "Cross-validation methods." Journal of mathematical psychology 44.1 (2000): 108-132.

[28]. Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4.2 (2009): 1883.

[29]. Breiman. 2001. “Random Forests.” Machine Learning 45 (1): 5–32.

[30]. Zou, and Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2): 301–20.

[31]. Chen, and Guestrin. 2016. “Xgboost: A Scalable Tree Boosting System.” In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 785–94. ACM.

[32]. Friedman. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” Annals of Statistics, 1189–1232.

[33]. Friedman. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” Annals of Statistics, 1189–1232.