IDENTIFICATION OF IRRELEVANT FEATURES IN CLINICAL NON-SMALL CELL LUNG CANCER DATA USING REVERSE FEATURE ELIMINATION MODEL

Sumalatha Mani

Research Scholar, Periyar University,
Salem, Tamil Nadu.
latha7suma@gmail.com

Dr. Latha Parthiban

Head In-charge, Department of Computer Science,
Pondicherry University, Community College Group, Puducherry
lathaparthiban@yahoo.com

Abstract

Non-Small Cell Lung Cancer (NSCLC) developed as a malignant cell cancer in the major regions of lungs thus identified as one of the life-threatening diseases. Many factors constitute to the development of this type of cancer in human lungs in both active or passive form. The major objective of this research paper is to identify the irrelevant factors or symptoms from a high dimensional clinical NSCLC dataset with 31 features using a novel Reverse Feature Elimination Dimensionality Reduction (RFEDR) algorithm designed for identifying worst features through Best Cost ranking technique applied in the dataset. In the initial stage pre-processing, 3 features (Admission-ID, Age and Gender) were removed in three filter stages and Feature Elimination method is applied. At the end of the implementation process with 40 iterations, the line of convergence was obtained Iteration 7 to 27. The best cost was determined to be 1.0042e-26 and the worst features through reverse elimination algorithm was identified as rare Alcohol Intake (Feature 2), Work Threats (Feature-4), Smoking Habit (Feature-9), Cold (Feature-20), Snoring (Feature-22) and Headache (Feature-27) respectively. The research sublimated the importance of eliminating irrelevant features apart from selecting best features for better prediction.

***Keywords*: Non-Small Cell Lung Cancer; Reverse Feature Elimination Dimensionality Reduction; clinical NSCLC dataset; Best Cost Ranking Technique; line of convergence.**

1. Introduction

Non-Small Cell Lung Cancer (NSCLC) is developed without early signs or symptoms and advances to the next stage at a rapid pace. NSCLC is also found complicated as it is moved to different forms of cancer like Adenocarcinoma, Squamous Carcinoma and Large Cell Carcinoma [1] since its inception into the human cells. Hence it cannot be predicted by analyzing the symptoms alone. The general habits and other features [2] also play a significant role in bringing a successful prediction at the earlier stage of NSCLC cancer. Hence the best features should be identified [3]. It is more important to spot the irrelevant features and eliminate them in the beginning of the prediction to generate good performance [4] as well. The major objective of this research work is to rank the features using a novel Reverse Feature Elimination Dimensionality Reduction (RFEDR) algorithm to identify the worst ranked features in a NSCLC clinical dataset to eliminate them at the beginning stage of prediction using classifiers. The scope of the research was confined to the lung cancer patient's responses collected from hospital through questionnaire and formed as a dataset for evaluation. The dataset was expected to contain impure data that are physically and logically irrelevant to the prediction of an expert engine designed using the classifiers.

2. Literature Review

The Literature reviews are performed to counteract the fact that the numerical analysis from questionnaire could effectively predict the NSCLC cancer rather than an Image Processing or Biopsy or image diagnosis [5] data from medical scans. Various studies were conducted based on image processing [6] and biopsy feature [7] analysis

DOI : 10.21817/indjcse/2022/v13i5/221305187

techniques where the prediction methods are defined and were able to predict the disease. Some of them are analysed and presented under Table-1.

Ref. No	Author & Year	Methodology	Techniques used	Results and drawbacks
[8]	Coudray, N et.al. (2018)	Collected individual data based on biopsies, frozen and formalin fixed paraffin tissues.	Deep learning methods to investigate images	Predicted STK11, EGFR, FAT1, SETBP1, KRAS and TP53 as predictable features
[9]	Bracht, J. W. P et.al. (2018)	Investigated tumor-educated platelets (TEPs), liquid bio sources, cell-free DNA (cfDNA), circulating tumor cells (CTCs), and extracellular vesicles (EVs) for predicting NSCLC.	Genetic Biopsy Analysis	Predicted Adenocarcinoma type cancer at earlier stage.
[10]	Liang, H., et.al. (2018)	E-Databases for NSCLC studied hazards ratio (HR) for the disease-free survival (DFS) between CTCs/ctDNA positive and negative groups.	Statistical and Medical recurrence Analysis.	Both CTCs and ctDNA were capable of predicting biomarkers of NSCLC disease.
[11]	Haragan, A., et.al. (2019)	The nodal metastases expression analysis for PD-L1evaluated using SP263 from 107 clone resected primary NSCLC data.	Squares method using Digital Image Processing Technique.	Heterogeneity tumour was predictable in 53% cases.
[12]	Khorrami, M et.al. (2020)	Analysed 139 patients with NSCLC at two institutions, separated into a discovery set (D1 = 50) and two independent validation sets (D2 = 62, D3 = 27)	Linear Discriminant Analysis (LDA) classifier trained with DelRADx features	DelRADx found effective in early prediction of NSCLC
[13]	Li, Y., et.al. (2018)	Tested optimal features of CT images to predict NSCLC cancer	Machine Learning Classifiers and Area Under Curve	Predicted heterogeneity of cancer cells at earliest stage.
[14]	Zhang, J., et.al. (2020)	Analysed clinical and radiomic features of F-FDG PET/CT to detect cancer cell growth status.	Used Least Absolute Shrinkage and Selection Operator (LASSO) algorithm	Clinical and radiomic features were predicted with improved performance.
[15]	Karlsson, A., et.al. (2019)	Examined NSCLC histology for combined, simultaneous, histological classification and fusion gene detection in minimal formalin fixed paraffin embedded (FFPE) tissue for NSCLC prediction	Gene Expression and SSP Measure	SSP Gene Expression proved successful in diagnostic lung cancer
[16]	Guibert, N., et.al. (2019)	Analysed 36 genes of NSCLC data and its kinetics observed.	Immuno score Analysis technique	Clinical outcome analysis was predicted with good immune score.

Table-1: Feature Analysis reviews based on image processing biomarkers and biopsy data

The literature surveys in Table-1 showed that the numerical analysis was essential for prediction of NSCLC cancer in many situations. Based on various reviews on NSCLC prediction, few research gaps were identified as follows:

- The image processing techniques were cost effective [11] [18] and were found in most of the research works completed,
- The technology was very expensive for initial screening as it was hard for common people to examine at regular intervals,
- The data were mostly image [13][14] and biopsy [8][18] based that could only be handled by experts and doctors for prediction,
- The commonly used clinical records [21] [22] had not combined with symptoms and habit-based tests in Lung Cancer predictions and
- The feature handling for complex medical datasets focused on the selection of best features rather than elimination of irrelevant features.

Thus, based on the identification of gaps in feature analysis process, the proposed model to find the irrelevant features was determined.

3. Materials and Methods

The dataset for Feature Analysis was created based on the Questionnaire conducted with 1000 respondents from Lung Cancer Institutes based diagnosis and their predicted stage of cancer affected. The Questionnaire collected from radiology departments of cancer institute was based on the symptoms, medical history and general habits of the patients. Also, their clinical diagnosed stage of NSCLC infection was obtained from hospital as class data. The features collected based on various categories of responses from patients are summarised in Table-2.

General Details	**Physical Exposure**	**Genetic Problems**	
1. Admission Number 2. Age 3. Gender	4. Pollution in Air 5. Alcohol Intake 6. Allergic to Dust 7. Work Threats	8. Genetic Risk 9. Chronic Lung Disease 10. Balanced Diet 11. Obesity	
Common Habits	**General Symptoms**	**Frequent Symptoms**	**NSCLC specific Symptoms**
12. Smoking Habit 13. Exposure to Smoking 14. Occasional Chest Pain 15. Coughing with Blood	16. Tiredness 17. Sudden Weight Loss 18. Reduced SPO2 19. Breathing Difficulty	20. Wheezing Nature 21. Difficult to Swallow 22. Finger Nails Clubbing 23. Cold 24. Dry Cough 25. Snoring	26. Sudden Weight Loss 27. Appetite 28. Hoarseness 29. Haemoptysis 30. Headache 31. Bone Pain
32. Stage of Cancer diagnosed clinically (Class Feature)			
Beginner	Moderate	Advanced	

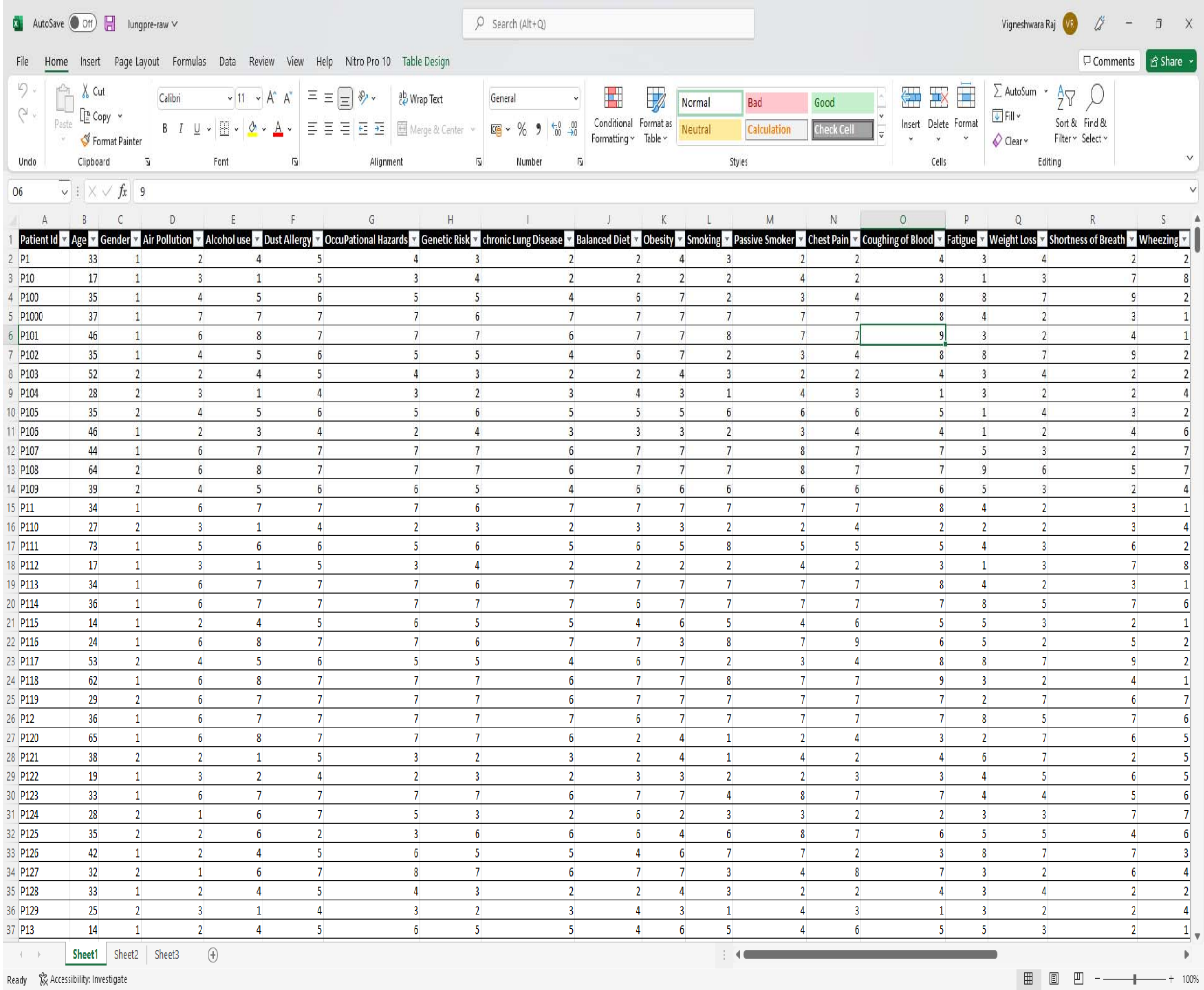
Table-2: Clinical Features collected from medical diagnosis to prepare dataset.

The initial dataset was prepared in Excel format with 32 features that has 31 predictive features and 1 class feature. The questionnaire values ranged as shown in Table-3.

Feature Value	Interpretation	Numeric Value
1	Extremely low	1
2	Very Low	2
3	Low	3
4	Moderately Low	4
5	Moderate	5
6	Moderately High	6
7	High	7
8	Very High	8
9	Extremely High	9

Table-3: The interpretation of feature values in the Questionnaire

It was evident from Table-3 that the feature value with 1 is regarded as the value with Extremely low, 2 as Very Low, 3 as Low, 4 as Moderately Low, 5 as Moderate, 6 as Moderately High, 7 as High, 8 as Very High and 9 as Extremely High in terms of all the features stated in Table-2. The responses were developed as table and class attribute was coded with NSCLC at beginner stage as 1, moderate as 2 and advanced as 3 respectively. All these values represent the level of infection acquired by the patients on a particular symptom of Non-Small Cell Lung Cancer. Also, it signifies the nature of answer given by the respondents during the data collection process. The data collected in the form of questionnaire are converted into Excel form as shown in Figure-1.



O6 | 9

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	Smoking	Passive Smoker	Chest Pain	Coughing of Blood	Fatigue	Weight Loss	Shortness of Breath	Wheezing
2	P1	33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2
3	P10	17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8
4	P100	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2
5	P1000	37	1	7	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1
6	P101	46	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1
7	P102	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2
8	P103	52	2	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2
9	P104	28	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4
10	P105	35	2	4	5	6	5	6	5	5	5	6	6	6	5	1	4	3	2
11	P106	46	1	2	3	4	2	4	3	3	3	2	3	4	4	1	2	4	6
12	P107	44	1	6	7	7	7	7	6	7	7	7	8	7	7	5	3	2	7
13	P108	64	2	6	8	7	7	7	6	7	7	7	8	7	7	9	6	5	7
14	P109	39	2	4	5	6	6	5	4	6	6	6	6	6	6	5	3	2	4
15	P11	34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1
16	P110	27	2	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4
17	P111	73	1	5	6	6	5	6	5	6	5	8	5	5	5	4	3	6	2
18	P112	17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8
19	P113	34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1
20	P114	36	1	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6
21	P115	14	1	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1
22	P116	24	1	6	8	7	7	6	7	7	3	8	7	9	6	5	2	5	2
23	P117	53	2	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2
24	P118	62	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1
25	P119	29	2	6	7	7	7	7	6	7	7	7	7	7	7	2	7	6	7
26	P12	36	1	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6
27	P120	65	1	6	8	7	7	7	6	2	4	1	2	4	3	2	7	6	5
28	P121	38	2	2	1	5	3	2	3	2	4	1	4	2	4	6	7	2	5
29	P122	19	1	3	2	4	2	3	2	3	3	2	2	3	3	4	5	6	5
30	P123	33	1	6	7	7	7	7	6	7	7	4	8	7	7	4	4	5	6
31	P124	28	2	1	6	7	5	3	2	6	2	3	3	2	2	3	3	7	7
32	P125	35	2	2	6	2	3	6	6	6	4	6	8	7	6	5	5	4	6
33	P126	42	1	2	4	5	6	5	5	4	6	7	7	2	3	8	7	7	3
34	P127	32	2	1	6	7	8	7	6	7	7	3	4	8	7	3	2	6	4
35	P128	33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2
36	P129	25	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4
37	P13	14	1	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1

Figure-1: RAW dataset created using Questionnaire collected from NSCLC patients

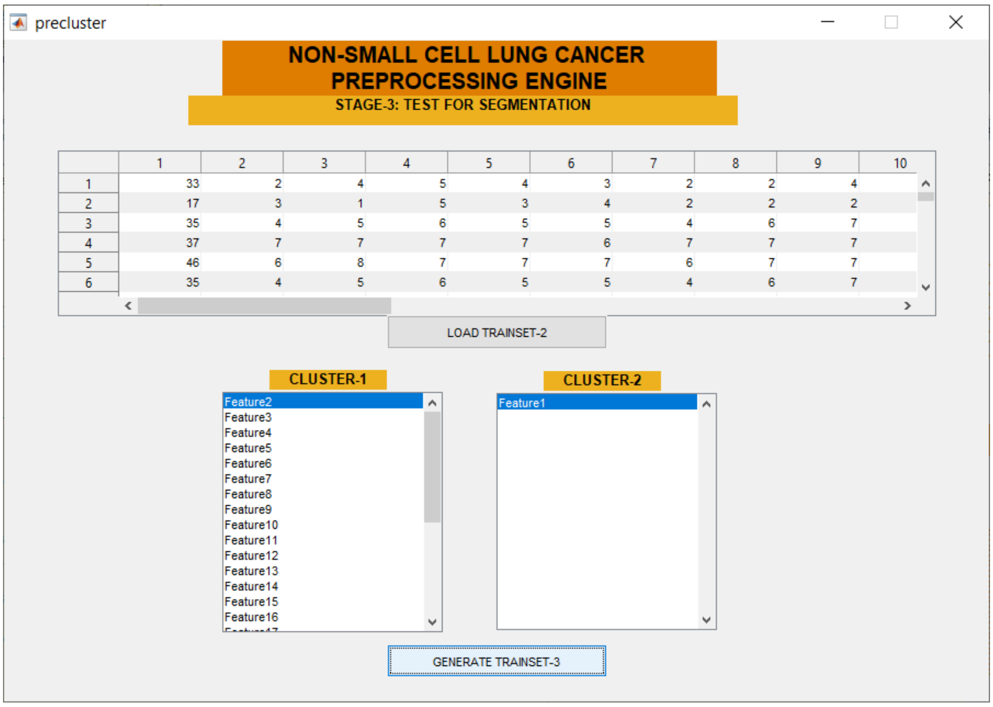
Based on the responses collected from various patients, the dataset was prepared and loaded into the MATLAB interface [22]. The initial pre-processing to determine the irrelevant datasets and remove them from the dataset was planned in the initial stage. The initial pre-processing [23] was further divided into three stages followed by the feature elimination method.

- Pre-Processing with multistage feature removal
  Stage-I : Test for Relevancy
  Stage-II : Test for Regression
  Stage-III: Test for Segmentation and Clustering
- Reverse Feature Elimination method

The worst features attained at the end of the Reverse Feature Elimination process was removed from the original dataset and trimmed as the testing set for further evaluation process. The initial pre-processing was carried out in an experimentative environment with GUI design in MATLAB. The initial pre-processing was completed in three stages.

3.1 Test for Relevance and Numeracy values

The Raw dataset was examined as independent features and tested for relevancy of data in comparison with the remaining features. The test was completed for finding missing values [24], non-numeric or text values [25], empty values, Not-A-Number (NAN) [26] and NULL [27] values. The missing and empty values represent the non-availability of values in the feature set. The clinical dataset when evaluated has more chance for evaluation of performance if only it contains numeric values. But in many situations, non-numeric values including alphabets, special characters and alpha numeric values were formed. The implementation was completed with the identification of irrelevant feature as shown in Figure-2.



	1	2	3	4	5	6	7	8	9	10
1	33	2	4	5	4	3	2	2	4	
2	17	3	1	5	3	4	2	2	2	
3	35	4	5	6	5	5	4	6	7	
4	37	7	7	7	7	6	7	7	7	
5	46	6	8	7	7	7	6	7	7	
6	35	4	5	6	5	5	4	6	7	

CLUSTER-1

Feature2
Feature3
Feature4
Feature5
Feature6
Feature7
Feature8
Feature9
Feature10
Feature11
Feature12
Feature13
Feature14
Feature15
Feature16

CLUSTER-2

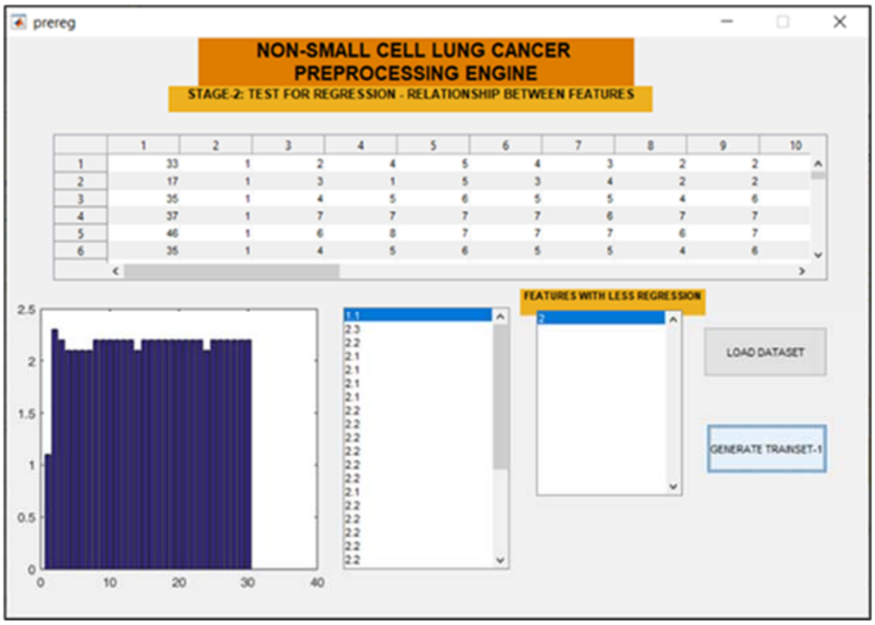
Feature1

Figure-2: Removal of irrelevant non-numeric feature in Test for Relevancy

The examination was conducted based on the novel algorithm for testing the relevancy. At the end of the examination, "Admission Number" was found to be highly irrelevant as it contained non-numeric values. Hence it was eliminated from the dataset and the remaining 31 features were formed as excel sheet for next stage of pre-processing.

3.2 Test for Relationship between features

In the second stage, the features are tested to identify the relationship or correlation between them using regression techniques [28]. The distance between values of features is computed and identified the irrelevant features. The threshold value was fixed based on the average of the feature values. The threshold value [29] was tested with the individual value of the feature to find the relationship. If the distance between the individual value and the threshold value was identified less than the threshold value, the feature was identified as the irrelevant feature. After the end of the test, the second feature 'Gender' was identified irrelevant and was removed from the dataset as indicated in Figure-3.


prereg
NON-SMALL CELL LUNG CANCER
PREPROCESSING ENGINE
STAGE-2: TEST FOR REGRESSION - RELATIONSHIP BETWEEN FEATURES
2.5
2
1.5
1
0.5
0
0
10
20
30
40
FEATURES WITH LESS REGRESSION
LOAD DATASET
GENERATE TRAINSET-1

	1	2	3	4	5	6	7	8	9	10
1	33	1	2	4	5	4	3	2	2	
2	17	1	3	1	5	3	4	2	2	
3	35	1	4	5	6	5	5	4	6	
4	37	1	7	7	7	7	6	7	7	
5	46	1	6	8	7	7	7	6	7	
6	35	1	4	5	6	5	5	4	6	

1.1
2.3
2.2
2.1
2.1
2.1
2.1
2.2
2.2
2.2
2.2
2.2
2.2
2.1
2.2
2.2
2.2
2.2
2.2

FEATURES WITH LESS REGRESSION

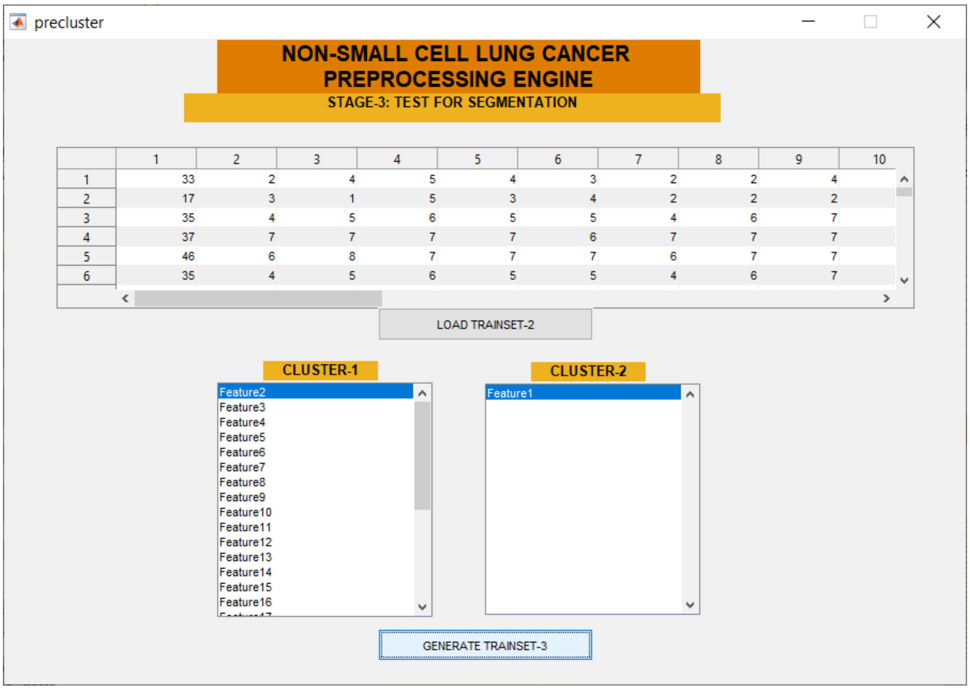
2

Figure-3: Pre-Processing test for relationship between features

The selected features after removing the "Gender" feature are formed as subset and developed as excel sheet for next stage pre-processing.

3.3 Test for Clustering of features

The final stage of pre-processing involved identification of segmented data using centroid values [30] to classify the related and unrelated data. The unrelated data are formed as clusters and identified as features with no relation to the other data. The 'age' feature was identified as unrelated and removed from dataset as shown in Figure-4.



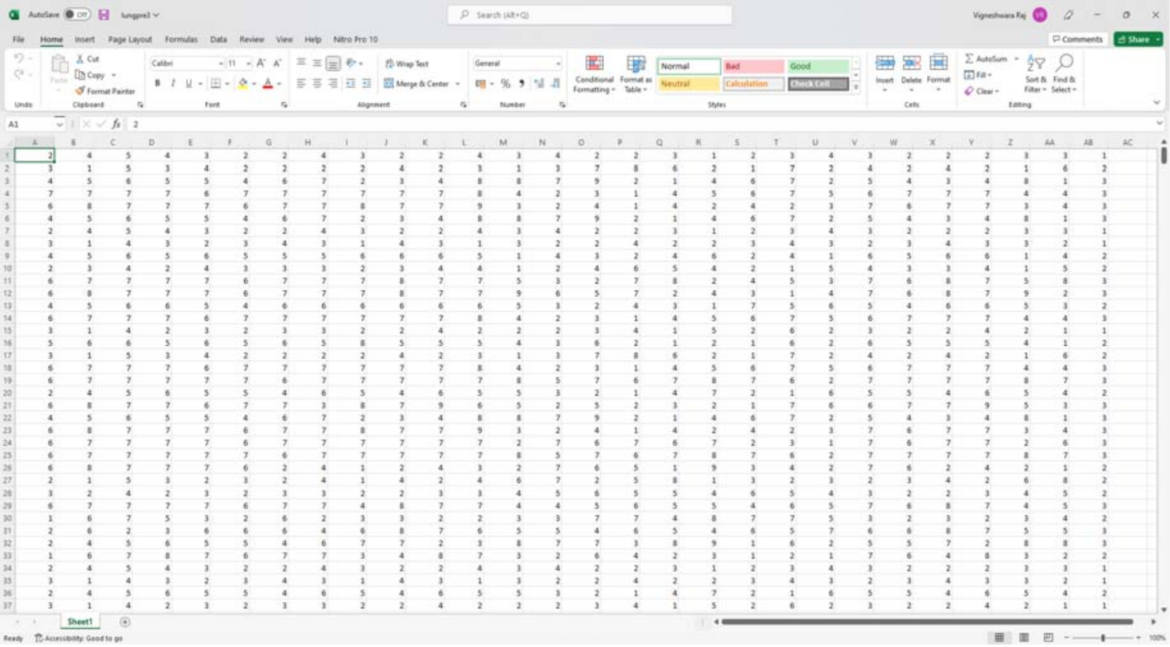
	1	2	3	4	5	6	7	8	9	10
1	33	2	4	5	4	3	2	2	4	
2	17	3	1	5	3	4	2	2	2	
3	35	4	5	6	5	5	4	6	7	
4	37	7	7	7	7	6	7	7	7	
5	46	6	8	7	7	7	6	7	7	
6	35	4	5	6	5	5	4	6	7	

CLUSTER-1: Feature2, Feature3, Feature4, Feature5, Feature6, Feature7, Feature8, Feature9, Feature10, Feature11, Feature12, Feature13, Feature14, Feature15, Feature16, Feature17

CLUSTER-2: Feature1

Figure-4: Pre-processing involving formation of clusters to identify irrelevant feature.

As shown in Figure-4, the irrelevant feature 'age' is removed and remaining 28 features were formed as training set for further feature elimination and dimensionality reduction [31] process in excel form as shown in Figure-5.



A1 | 2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	3	2	2	2	3	3	1
2	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	4	2	4	2	1	6	2
3	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	5	4	3	4	8	1	3
4	7	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	6	7	7	7	4	4	3
5	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	7	6	7	7	3	4	3
6	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	5	4	3	4	8	1	3
7	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	3	2	2	2	3	3	1
8	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	2	3	4	3	3	2	1
9	4	5	6	5	6	5	5	5	6	6	6	5	1	4	3	2	4	6	2	4	1	6	5	6	6	1	4	2
10	2	3	4	2	4	3	3	3	2	3	4	4	1	2	4	6	5	4	2	1	5	4	3	3	4	1	5	2
11	6	7	7	7	7	6	7	7	7	8	7	7	5	3	2	7	8	2	4	5	3	7	6	8	7	5	8	3
12	6	8	7	7	7	6	7	7	7	8	7	7	9	6	5	7	2	4	3	1	4	7	6	8	7	9	2	3
13	4	5	6	6	5	4	6	6	6	6	6	6	5	3	2	4	3	1	7	5	6	5	4	6	6	5	3	2
14	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	6	7	7	7	4	4	3
15	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	3	2	2	4	2	1	1
16	5	6	6	5	6	5	6	5	8	5	5	5	4	3	6	2	1	2	1	6	2	6	5	5	5	4	1	2
17	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	4	2	4	2	1	6	2
18	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	6	7	7	7	4	4	3
19	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6	7	8	7	6	2	7	7	7	7	8	7	3
20	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1	4	7	2	1	6	5	5	4	6	5	4	2
21	6	8	7	7	6	7	7	3	8	7	9	6	5	2	5	2	3	2	1	7	6	6	7	7	9	5	3	3
22	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	5	4	3	4	8	1	3
23	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	7	6	7	7	3	4	3
24	6	7	7	7	7	6	7	7	7	7	7	7	2	7	6	7	6	7	2	3	1	7	6	7	7	2	6	3
25	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6	7	8	7	6	2	7	7	7	7	8	7	3
26	6	8	7	7	7	6	2	4	1	2	4	3	2	7	6	5	1	9	3	4	2	7	6	2	4	2	1	2
27	2	1	5	3	2	3	2	4	1	4	2	4	6	7	2	5	8	1	3	2	3	2	3	4	2	6	8	2
28	3	2	4	2	3	2	3	3	2	2	3	3	4	5	6	5	5	4	6	5	4	3	2	2	3	4	5	2
29	6	7	7	7	7	6	7	7	4	8	7	7	4	4	5	6	5	5	4	6	5	7	6	8	7	4	5	3
30	1	6	7	5	3	2	6	2	3	3	2	2	3	3	7	7	4	8	7	7	5	3	2	3	2	3	4	2
31	2	6	2	3	6	6	6	4	6	8	7	6	5	5	4	6	5	4	6	5	7	6	6	8	7	5	5	3
32	2	4	5	6	5	5	4	6	7	7	2	3	8	7	7	3	8	9	1	6	2	5	5	7	2	8	8	3
33	1	6	7	8	7	6	7	7	3	4	8	7	3	2	6	4	2	3	1	2	1	7	6	4	8	3	2	2
34	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	3	2	2	2	3	3	1
35	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	2	3	4	3	3	2	1
36	2	4	5	6	5	5	4	6	5	4	6	5	5	3	2	1	4	7	2	1	6	5	5	4	6	5	4	2
37	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	3	2	2	4	2	1	1

Figure-5: Refined Dataset after Pre-Processing in three different levels

As shown in Figure 5, the RAW dataset was pre-processed in three levels and the refined dataset with 28 predictive features and 1 class feature are selected for next stage of elimination process. It is important to understand that one feature in each level of pre-processing accounting to only three features being totally removed from the dataset. This resulted in 9% of the features being removed. Since the removed features are not sufficient for evaluation, the next phase of irrelevant features removal process was initiated.

4. Reverse Feature Elimination Process

The primary focus of the research work is to design an algorithm that is capable of ranking the features based on its ability to predict the results with enhanced accuracy. After pre-processing, the formed training set is experimented into a series of iterations that continuously ranked the features based on its performance and then arranges the list of features based on its rank in ascending order determining the low ranked worst features. This process is formulated into a novel algorithmic model called Reverse Feature Elimination Dimensionality Reduction (RFEDR) algorithm. The algorithm comprised of model to perform 40 iterations in series order to determine the rank of features based on its performance. The best cost is calculated based on the summation of relevance to its subordinate features as shown in Eq. (1).

$$\boldsymbol{S(x) = (Diff(x) + \sum_{n=1}^{40}(\mathbf{x0} * \mathbf{exp}(\mathbf{x0}) + \mathbf{x1} * \mathbf{exp}(\mathbf{x1})))/n}\text{------- Eq. (1)}$$

Where x0 is the first feature and x1 is the next feature and the summated value is added with difference value of x and finally divided with total number of values for finding the best cost expression. The best cost is identified for 40 iterations and the features are ordered in ascending order. The best cost value would fall in consistent order at a certain point where maximum iteration carry the same best cost value and the same order of features. This series of iterations are found to be the line of convergence. This line of convergence is determined after completing all iterations as the best cost and the relevant feature order is the final order based on the rank and performance of features. During the initial phase of the algorithm, the dataset for evaluation was loaded with $\boldsymbol{\phi_V}$ as the features and C as the class feature to train and test the outcomes. The similarity (S) and relevancy (R) are initialised with total number samples (n). The initial values were set and the algorithm was loaded with random generated values and original samples from dataset. The loaded data are then processed using the Fitness function.

4.1 Fitness Function

The Fitness function is used to compute the relevance of the features with each other and their ability to predict the output at ease. This fitness function for Reverse Feature Elimination Dimensionality Reduction (RFEDR) algorithm manipulates the dataset features based on its values to rank the features and order them from highest relevance to lowest relevance as a set of features.

Based on the Pseudocode given, the total values 28000 from the dataset comprising of 28 features and 1000 samples are loaded and initialised.

```
For  ∀i∈N do
        For  ∀j∈N_Att do
                ∂i←randi([0, 3],ϕ_V,1)
                S_i←Sim_F  (ϕ_Vj)
        End For
End For
```

The above code indicates that random number has been generated with range from 0 to 3 with test of empty or non-numeric values. Later, all the values were selected and moved to the initial set ‘S’ to test for relevancy of the data based on the elimination process. The fitness function $(\boldsymbol{Fit_i})$ is assigned the value of the individual feature. The iteration begins with index value from N to 1 in reverse order. The independent sample $(\boldsymbol{\tau})$ is compared with other samples from the dataset $(\boldsymbol{\alpha})$ to find the relevance of feature with each other. If the condition is true, the fitness value is calculated by multiplying the weights of individual feature with the individual value of feature and summation of other features $(\boldsymbol{\delta_{1i}, \delta_{2i}, \delta_{1i}})$ and the difference between the feature and adjacent features respectively. If the condition is false, the difference is calculated and assigned to next iteration for further evaluation. This process is continued until the features are identified under fitness function to achieve highest prediction level to its lowest level.

4.2 Ranking Phase

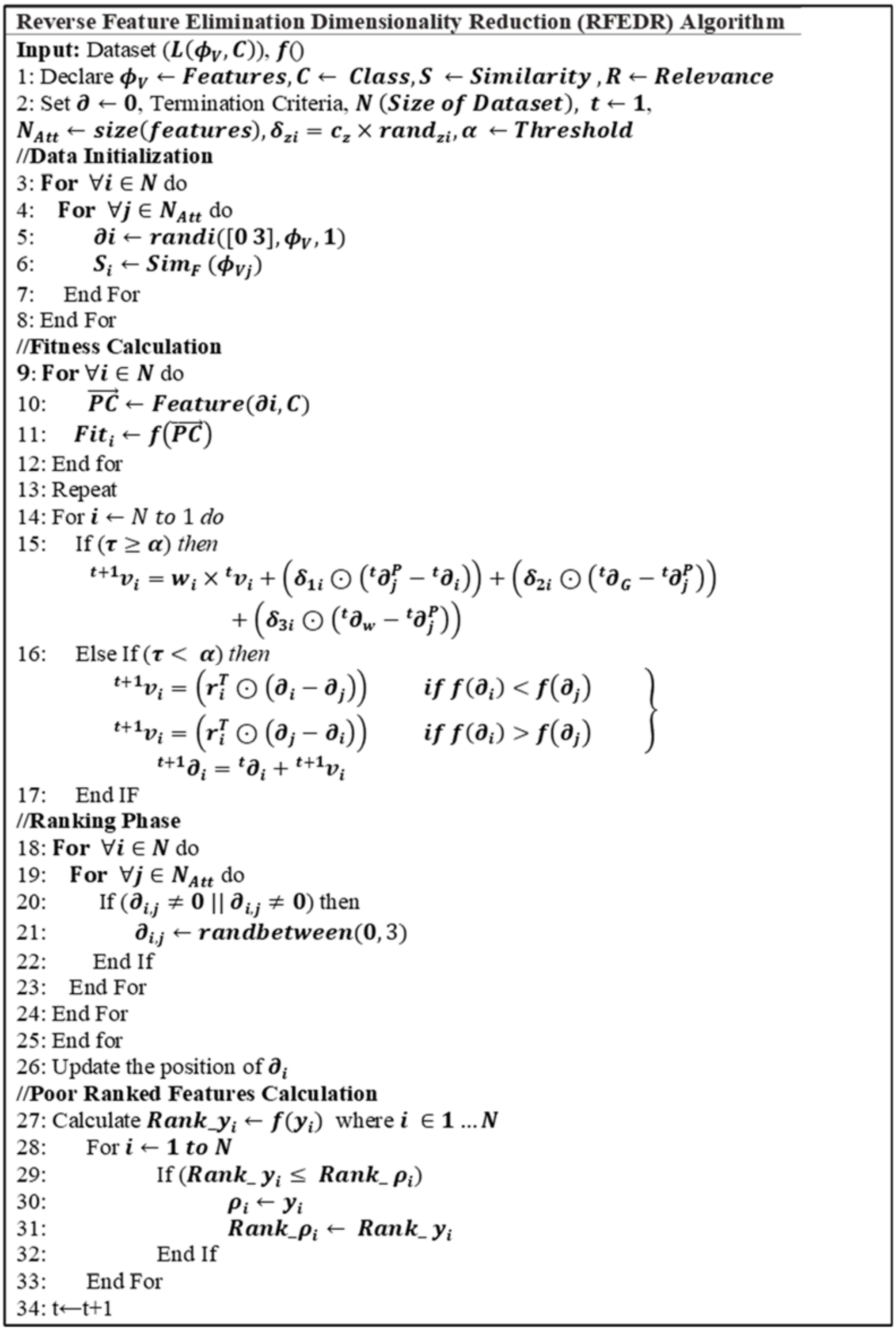
The Ranking phase follows after calculating the fitness function. The ranking phase collects the individual feature and its adjacent feature from every sample (N) and attribute feature (N_{Att}). The random number is generated between the values 0 to 3 to encompass the expected outcome of NSCLC prediction at beginner, moderate and advanced level. The random number is used to update the position of the feature at the end of each iteration.

```
For $\forall i \in N$ do
  For $\forall j \in N_{Att}$ do
    If ($\partial_{i,j} \neq 0$ || $\partial_{i,j} \neq 0$) then
      $\partial_{i,j} \leftarrow randbetween(0,3)$
    End If
  End For
End For
Update the position of $\partial_i$
```

The loop tests for non-zero values were performed and generated the random number between 0 to 3 to test the relationship with the individual feature. If the rank of the individual value was less than the existing value, the rank gets updated. After updating the position, the features with poor rank have to be segregated from high and moderate ranked features. To counteract this modification, the rank calculation is made by assigning best predictive feature from top order to the worst predictive feature in the lowermost order. After sorting and ordering of features, the iteration is incremented to the next level.

```
Calculate $Rank\_y_i \leftarrow f(y_i)$ where $i \in 1 \ldots N$
For $i \leftarrow 1\ to\ N$
      If ($Rank\_\ y_i \leq Rank\_\ \rho_i$)
            $\rho_i \leftarrow y_i$
            $Rank\_\rho_i \leftarrow Rank\_\ y_i$
      End If
End For
t←t+1
```

Thus, the algorithm performs utmost comparison of features with 40 iterations to find the line of convergence and thereby indicating the best cost. The Best cost [32] in this algorithm is not used to find the best or worst ranked features. Instead, it is used to identify the line of convergence where the ranked features can be finalised for selection or elimination process. The Reverse Feature Elimination Dimensionality Reduction (RFEDR) algorithm is shown in Figure-6.



```
Reverse Feature Elimination Dimensionality Reduction (RFEDR) Algorithm
Input: Dataset ($L(\phi_V, C)$), $f()$
Declare $\phi_V \leftarrow Features, C \leftarrow Class, S \leftarrow Similarity, R \leftarrow Relevance$
Set $\partial \leftarrow 0$, Termination Criteria, $N$ ($Size\ of\ Dataset$), $t \leftarrow 1$,
$N_{Att} \leftarrow size(features), \delta_{zi} = c_z \times rand_{zi}, \alpha \leftarrow Threshold$
//Data Initialization
For $\forall i \in N$ do
    For $\forall j \in N_{Att}$ do
        $\partial i \leftarrow randi([0\ 3], \phi_V, 1)$
        $S_i \leftarrow Sim_F(\phi_{Vj})$
    End For
End For
//Fitness Calculation
For $\forall i \in N$ do
    $\overrightarrow{PC} \leftarrow Feature(\partial i, C)$
    $Fit_i \leftarrow f(\overrightarrow{PC})$
End for
Repeat
For $i \leftarrow N$ to 1 do
    If ($\tau \geq \alpha$) then
        $^{t+1}v_i = w_i \times {}^{t}v_i + \left(\delta_{1i} \odot \left({}^{t}\partial_j^P - {}^{t}\partial_i\right)\right) + \left(\delta_{2i} \odot \left({}^{t}\partial_G - {}^{t}\partial_j^P\right)\right)$
                $+ \left(\delta_{3i} \odot \left({}^{t}\partial_w - {}^{t}\partial_j^P\right)\right)$
    Else If ($\tau < \alpha$) then
        $\left.\begin{array}{ll} {}^{t+1}v_i = \left(r_i^T \odot (\partial_i - \partial_j)\right) & if\ f(\partial_i) < f(\partial_j) \\ {}^{t+1}v_i = \left(r_i^T \odot (\partial_j - \partial_i)\right) & if\ f(\partial_i) > f(\partial_j) \end{array}\right\}$
          $^{t+1}\partial_i = {}^{t}\partial_i + {}^{t+1}v_i$
    End IF
//Ranking Phase
For $\forall i \in N$ do
    For $\forall j \in N_{Att}$ do
        If ($\partial_{i,j} \neq 0\ ||\ \partial_{i,j} \neq 0$) then
            $\partial_{i,j} \leftarrow randbetween(0, 3)$
        End If
    End For
End For
End for
Update the position of $\partial_i$
//Poor Ranked Features Calculation
Calculate $Rank\_y_i \leftarrow f(y_i)$ where $i \in 1 \ldots N$
    For $i \leftarrow 1$ to $N$
        If ($Rank\_\,y_i \leq Rank\_\,\rho_i$)
            $\rho_i \leftarrow y_i$
            $Rank\_\rho_i \leftarrow Rank\_\,y_i$
        End If
    End For
t←t+1
```

Figure-6: Reverse Feature Elimination Dimensionality Reduction (RFEDR) algorithm

5. Implementation and Evaluation

The model is implemented using MATLAB GUI tool and the algorithm is initialised with the NSCLC dataset with 28 features. The selected features were evaluated, ranked and arrived at the best cost values as shown in Table-4.

Iteration-1	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-2	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-3	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-4	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-5	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-6	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-7	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-8	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-9	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-10	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-11	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-12	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-13	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-14	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-15	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-16	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-17	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-18	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24

Best Cost	5.6105e-25
Iteration-19	
Selected Features	16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26, 10, 14, 18, 25, 27, 7, 19, 21, 4, 15, 2
Irrelevant Features	5, 1, 23, 6, 24
Best Cost	5.6105e-25
Iteration-20	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-21	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-22	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-23	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-24	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-25	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-26	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-27	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-28	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-29	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-30	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-31	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-32	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-33	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-34	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-35	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26
Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26
Iteration-36	
Selected Features	19, 21, 4, 15, 2, 10, 1, 23, 6, 24, 16, 3, 22, 9, 8, 12, 11, 20, 13, 17, 26

Irrelevant Features	5, 14, 18, 25, 27, 7
Best Cost	1.8919e-26

Table-4: Entire 40 iterations of the Reverse Feature Elimination Dimensionality Reduction Process.

Based on the iteration process information in Table-4, the iterations, the features selected with less relevant features for elimination process and their respective best cost was presented in Table-5.

Iteration No	Less Relevant Features	Best Cost
1	13 5 16 12 27	8.7621e-24
2	18 25 16 12 14	1.6415e-25
3	21 18 25 16 12 14	1.6415e-25
4	21 18 25 16 12 14	1.6415e-25
5	14 7 1 23 6 25	3.6874e-26
6	14 7 1 23 6 25	3.6874e-26
7	20 9 2 22 4 27	1.0042e-26
8	20 9 2 22 4 27	1.0042e-26
9	20 9 2 22 4 27	1.0042e-26
10	20 9 2 22 4 27	1.0042e-26
11	20 9 2 22 4 27	1.0042e-26
12	20 9 2 22 4 27	1.0042e-26
13	20 9 2 22 4 27	1.0042e-26
14	20 9 2 22 4 27	1.0042e-26
15	20 9 2 22 4 27	1.0042e-26
16	20 9 2 22 4 27	1.0042e-26
17	20 9 2 22 4 27	1.0042e-26
18	20 9 2 22 4 27	1.0042e-26
19	20 9 2 22 4 27	1.0042e-26
20	20 9 2 22 4 27	1.0042e-26
21	20 9 2 22 4 27	1.0042e-26
22	20 9 2 22 4 27	1.0042e-26
23	20 9 2 22 4 27	1.0042e-26
24	20 9 2 22 4 27	1.0042e-26
25	20 9 2 22 4 27	1.0042e-26
26	20 9 2 22 4 27	1.0042e-26
27	20 9 2 22 4 27	1.0042e-26
28	15 19 21 18 10 3	2.5434e-28
29	15 19 21 18 10 3	2.5434e-28
30	15 19 21 18 10 3	2.5434e-28
31	15 19 21 18 10 3	2.5434e-28
32	15 19 21 18 10 3	2.5434e-28
33	15 19 21 18 10 3	2.5434e-28
34	15 19 21 18 10 3	2.5434e-28
35	15 19 21 18 10 3	2.5434e-28
36	15 19 21 18 10 3	2.5434e-28
37	15 19 21 18 10 3	2.5434e-28
38	15 19 21 18 10 3	2.5434e-28
39	15 19 21 18 10 3	2.5434e-28
40	15 19 21 18 10 3	2.5434e-28

Table-5: Implementation of RFEDR to determine low ranked features for elimination

In the above Table-4, the low ranked features are identified after each iteration. Based on the similarity of features selected in each iteration and the best cost identified, the following observations are grouped and summarized as shown in Table-6.

Iteration Range	Total Iterations	Less Relevant Features	Best Cost
1	1	13 5 16 12 27	8.7621e-24
2-4	3	18 25 16 12 14	1.6415e-25
5-6	2	14 7 1 23 6 25	3.6874e-26
7-27	21	20 9 2 22 4 27	1.0042e-26
28-40	13	15 19 21 18 10 3	2.5434e-28

Table-6: Ordering of Low Ranked features and their respective iteration range after completing RFEDR

As Shown in Table-5, the best cost has to be identified and finalized based on the number of iterations of the reverse feature elimination process. Accordingly, it was found that the iteration range from Iteration-7 to Iteration-27 with a total frequency of 21 iterations has achieved the line of convergence as shown in the MATLAB graph portrayed in Figure-7.

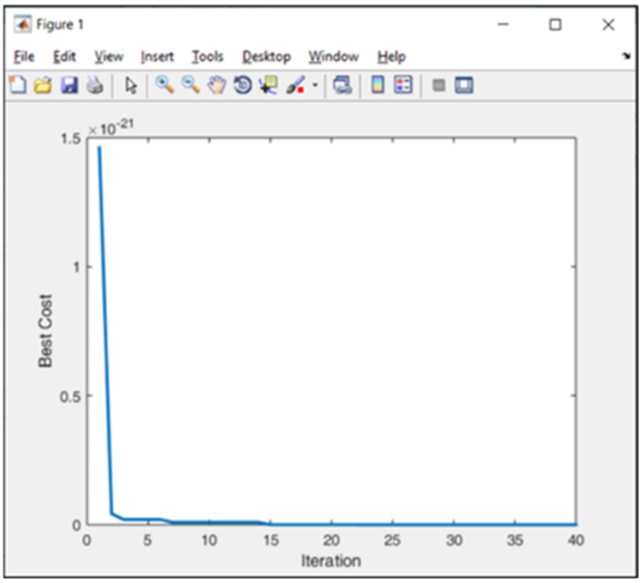


Figure 7 – Graph obtained in MATLAB after Elimination Process.

It is evident from Table-6 and Figure 7 that the iteration range 7-27 has maximum iterations with similar ordering of elements {2, 4, 9, 20, 22, 27} and also the best cost 1.0042e-26 remains the same. Thus, these features were selected as subset from the original dataset ready for elimination process. The second highest found in range of iterations between 28 and 40 with values {15, 19, 21, 18, 10, 3} respectively. But since it is less than the earlier sequence, the range of Iteration-7 to Iteration-27 achieving a continuous line of convergence as shown in Figure-8 has been selected as the worst features for elimination.

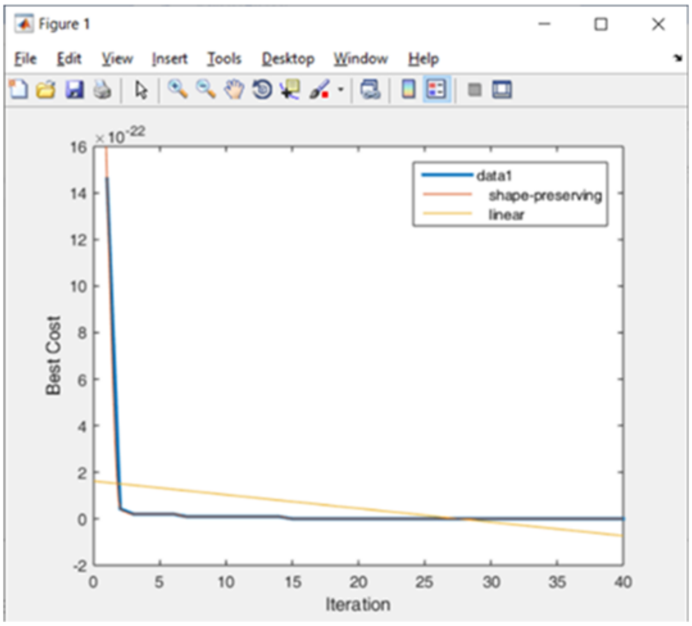
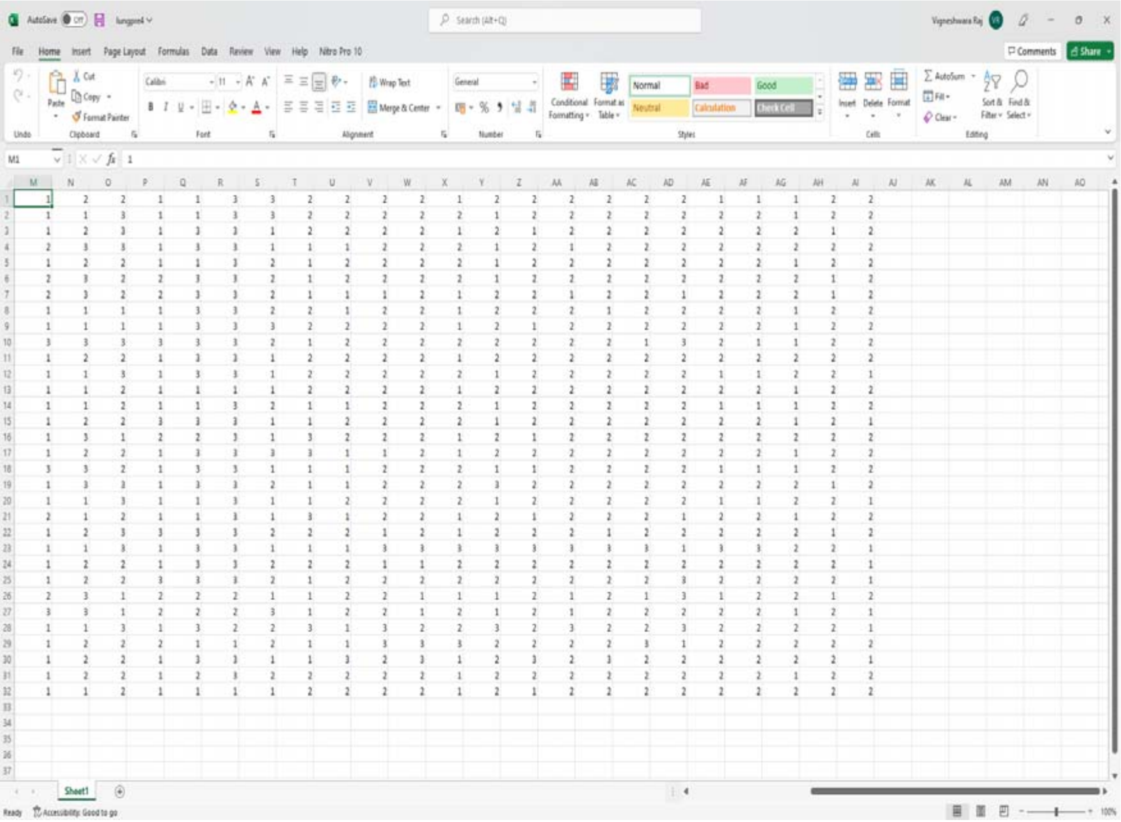


Figure 8 – Line of Convergence obtained after Iterations.

However, the total of 21 iterations is in no match to other line of convergence ranges with 13 or less iterations as shown in Figure-4. Hence 7-27 is found to have attained the maximum convergence and the subsequent features {2, 4, 9, 20, 22, 27} are selected. Based on the analysis with the available features in dataset, it is found to be Alcohol Intake (Feature 2), Work Threats (Feature-4), Smoking Habit (Feature-9), Cold (Feature-20), Snoring (Feature-22) and Headache (Feature-27) respectively. After removing these 6 features, the remaining 22 features are created as excel sheet for further processing and prediction of NSCLC as shown in Figure 9.



M1 fx 1

	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	1	2	2	1	1	3	3	2	2	2	2	1	2	2	2	2	2	2	1	1	1	2	2
2	1	1	3	1	1	3	3	2	2	2	2	2	1	2	2	2	2	2	2	2	1	2	2
3	1	2	3	1	3	3	1	2	2	2	2	1	2	1	2	2	2	2	2	2	2	1	2
4	2	3	3	1	3	3	1	1	1	2	2	2	1	2	1	2	2	2	2	2	2	2	2
5	1	2	2	1	1	3	2	1	2	2	2	2	1	2	2	2	2	2	2	2	1	2	2
6	2	3	2	2	3	3	2	1	2	2	2	2	1	2	2	2	2	2	2	2	2	1	2
7	2	3	2	2	3	3	2	1	1	1	2	1	2	2	1	2	2	1	2	2	2	1	2
8	1	1	1	1	3	3	2	2	1	2	2	1	2	2	2	1	2	2	2	2	1	2	2
9	1	1	1	1	3	3	3	2	2	2	2	1	2	1	2	2	2	2	2	2	1	2	2
10	3	3	3	3	3	3	2	1	2	2	2	2	2	2	2	2	1	3	2	1	1	2	2
11	1	2	2	1	3	3	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2
12	1	1	3	1	3	3	1	2	2	2	2	2	1	2	2	2	2	2	1	1	2	2	1
13	1	1	2	1	1	1	1	2	2	2	2	1	2	2	2	2	2	2	2	2	1	2	2
14	1	1	2	1	1	3	2	1	1	2	2	2	1	2	2	2	2	2	1	1	1	2	2
15	1	2	2	3	3	3	1	1	2	2	2	2	1	2	2	2	2	2	2	2	1	2	1
16	1	3	1	2	2	3	1	3	2	2	2	1	2	1	2	2	2	2	2	2	2	2	2
17	1	2	2	1	3	3	3	3	1	1	2	1	2	2	2	2	2	2	2	2	1	2	2
18	3	3	2	1	3	3	1	1	1	2	2	2	1	1	2	2	2	2	1	1	1	2	2
19	1	3	3	1	3	3	2	1	1	2	2	2	3	2	2	2	2	2	2	2	2	1	2
20	1	1	3	1	1	3	1	1	2	2	2	2	1	2	2	2	2	2	1	1	2	2	1
21	2	1	2	1	1	3	1	3	1	2	2	1	2	1	2	2	2	1	2	2	1	2	2
22	1	2	3	3	3	3	2	2	2	1	2	1	2	2	2	1	2	2	2	2	2	1	2
23	1	1	3	1	3	3	1	1	1	3	3	3	3	3	3	3	3	1	3	3	2	2	1
24	1	2	2	1	3	3	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	1
25	1	2	2	3	3	3	2	1	2	2	2	2	2	2	2	2	2	3	2	2	2	2	1
26	2	3	1	2	2	2	1	1	2	2	1	1	1	2	1	2	1	3	1	2	2	1	2
27	3	3	1	2	2	2	3	1	2	2	1	2	1	2	1	2	2	2	2	2	1	2	1
28	1	1	3	1	3	2	2	3	1	3	2	2	3	2	3	2	2	3	2	2	2	2	1
29	1	2	2	2	1	1	2	1	1	3	3	3	2	2	2	2	3	1	2	2	2	2	2
30	1	2	2	1	3	3	1	1	3	2	3	1	2	3	2	3	2	2	2	2	2	2	1
31	1	2	2	1	2	3	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1	2	2
32	1	1	2	1	1	1	1	2	2	2	2	1	2	1	2	2	2	2	2	2	2	2	2

Figure 9 – The training dataset after feature elimination process

The training dataset after removing further 6 features with 21% feature elimination contained 22 features at the end of the process. The features contained in the dataset at the end of Feature Pre-Processing and Reverse Feature elimination process were given in Table-7.

General Information	**Physical Exposure**	**Genetic Problems**	
NIL	1. Pollution in Air 2. Allergic to Dust	3. Genetic Risk 4. Chronic Lung Disease 5. Balanced Diet 6. Obesity	
Common Habits	**General Symptoms**	**Frequent Symptoms**	**NSCLC specific Symptoms**
7. Exposure to Smoking 8. Occasional Chest Pain 9. Coughing with Blood	10. Tiredness 11. Sudden Weight Loss 12. Reduced SPO2 13. Breathing Difficulty	14. Wheezing Nature 15. Difficult to Swallow 16. Finger Nails Clubbing 17. Dry Cough	18. Sudden Weight Loss 19. Appetite 20. Hoarseness 21. Haemoptysis 22. Bone Pain
23. Stage of Cancer diagnosed clinically (Class Feature)			
Beginner	Moderate	Advanced	

Table-7: Features finalised after Pre-Processing and Feature Elimination Process

It was evident from the Table-7 that the irrelevant features in each category were removed based on its redundancy nature and unwanted values in the dataset. The features including “Admission Number”, “Gender” and “Age” were removed from the dataset in the pre-processing stage. These three values represent the general information in the questionnaire. Hence it is by default that these three features couldn’t be a major factor for predicting the NSCLC cancer thereby justifying the reduction of features from the RAW dataset. During the second phase of Feature Elimination Dimensionality Reduction method, the six features Alcohol Intake (Feature 2), Work Threats (Feature-4), Smoking Habit (Feature-9), Cold (Feature-20), Snoring (Feature-22) and Headache (Feature-27) were removed. They are also removed from different categories of Questionnaire as shown in Table-7 and finally formed as training set as shown in Figure-9 respectively.

6. Results and Discussions

The major merit in designing this Feature elimination model is to reduce the incompetent and irrelevant feature from the collected features in the dataset to enhance prediction.

The core knowledge behind this study and implementation is that by reducing the irrelevant features,

- The size of the dataset is reduced from 31000 to 22000 values,
- The computational time can be improved as size is reduced because the number of iterations and columns are reduced in the matrix,
- The prediction accuracy can be augmented through testing with various data mining classifiers.
- The research work created a model algorithm that is capable of identifying the work relevant features that are eligible for elimination from a dataset,

The dataset had 31 predictive and 1 class feature in the beginning of the evaluation. In the initial stage, through pre-processing 3 features were removed from three levels accounting to 29 features selected for evaluation. Henceforth, the pre-processing reduced 10% of features that are found totally irrelevant to the prediction of NSCLC. After RFEDR, further 6 more features were removed and the total was found to be 9 features with further loss of 29.03% of features. Overall, the total predictive features of 28 was reduced to 22 with elimination of low ranked 6 features. Therefore, the selected 22 features showed 82% of selection of relevant features and 18% of elimination of irrelevant features. Finally, the result shows that the Elimination process has removed more features in comparison with the Pre-Processing stage as shown in Figure-9 and Figure-10.

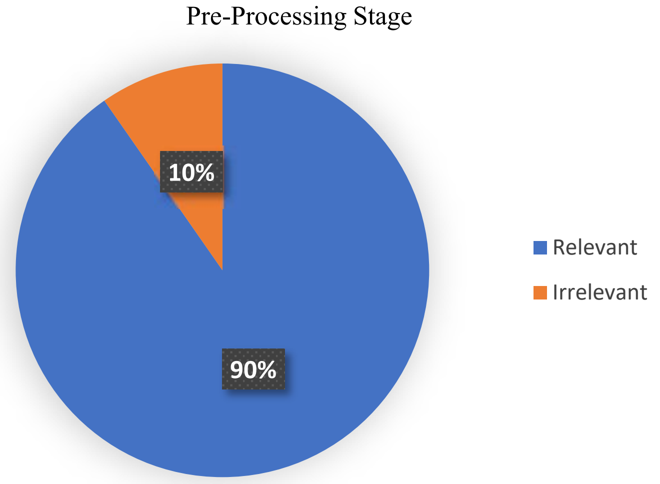


Figure 9 – Percentage Feature reduction after Pre-Processing

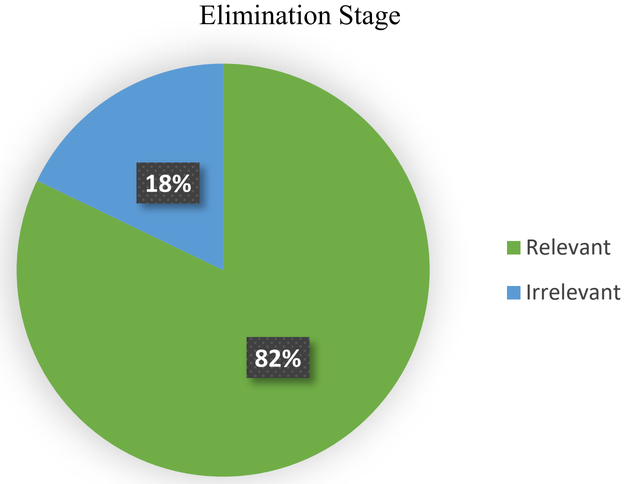


Figure 10 – Percentage Feature reduction after Elimination Process

It is evident from Figure 5 and Figure 6 that the elimination stage has 82% of selected features with 18% removal which outperforms 90% selection and 10% elimination of the features in pre-processing level. The model algorithm Reverse Feature Elimination Dimensionality Reduction (RFEDR) has achieved the objective of enhanced performance with reduced features. The testing dataset can be evaluated and predicted using classifiers in data mining. These kind of ranking algorithms in reverse order to determine worst features assists in medical complex datasets to filter and remove irrelevant features that do not ponder any successful prediction but remains in the dataset to reduce the performance. Thereby removing such features can surely improve the performance in terms of size, computational performance and efficiency.

7. Conclusion

The research work focused on the importance of removing irrelevant, inconsistent and non-predictable features from the set of features thereby improving the performance of prediction and evaluation of clinical NSCLC dataset. The dataset was formed based on various expected symptoms, habitual nature of patients. However, it contained irrelevant features that are not capable of prediction with enhanced accuracy levels. Hence the proposed model algorithm Reverse Feature Elimination Dimensionality Reduction (RFEDR) was implemented with fitness and ranking function to determine worst ranked features and removed to form a better dataset. This model can further be propounded to feature extraction and Expert Engine creation to predict NSCLC cancer in a successful manner.

Conflict Of Interest:

The Authors have no conflict of interest to declare

References

[1] Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature, 553(7689), 446-454.

[2] Wang, M., Herbst, R. S., & Boshoff, C. (2021). Toward personalized treatment approaches for non-small-cell lung cancer. Nature medicine, 27(8), 1345-1356.

[3] Wang, X., Kerrigan, K., Puri, S., Shen, J., Akerley, W., & Haaland, B. (2022). Dynamic Prediction of Near-Term Overall Survival in Patients with Advanced NSCLC Based on Real-World Data. Cancers, 14(3), 690.

[4] Patel, A. J., Tan, T. M., Richter, A. G., Naidu, B., Blackburn, J. M., & Middleton, G. W. (2022). A highly predictive autoantibody-based biomarker panel for prognosis in early-stage NSCLC with potential therapeutic implications. British journal of cancer, 126(2), 238-246.

[5] Duma, N., Santana-Davila, R., & Molina, J. R. (2019, August). Non–small cell lung cancer: epidemiology, screening, diagnosis, and treatment. In Mayo Clinic Proceedings (Vol. 94, No. 8, pp. 1623-1640). Elsevier.

[6] Hinestrosa, J. P., Lewis, J. M., Schroder, G., Balcer, H. I., Kurzrock, R., & Krishnan, R. (2022). Detection of early-stage lung cancer using a liquid biopsy test based on extracellular vesicle proteins. Cancer Research, 82(12_Supplement), 727-727.

[7] Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems, 92, 374-382.

[8] Khorrami, M., Prasanna, P., Gupta, A., Patil, P., Velu, P. D., Thawani, R., ... & Madabhushi, A. (2020). Changes in CT radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non–small cell lung cancer. Cancer immunology research, 8(1), 108-119.

[9] Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., ... & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine, 24(10), 1559-1567.

[10] Bracht, J. W. P., Mayo-de-Las-Casas, C., Berenguer, J., Karachaliou, N., & Rosell, R. (2018). The present and future of liquid biopsies in non-small cell lung cancer: combining four biosources for diagnosis, prognosis, prediction, and disease monitoring. Current Oncology Reports, 20(9), 1-10.

[11] Liang, H., Huang, J., Wang, B., Liu, Z., He, J., & Liang, W. (2018). The role of liquid biopsy in predicting post-operative recurrence of non-small cell lung cancer. Journal of thoracic disease, 10(Suppl 7), S838.

[12] Haragan, A., Field, J. K., Davies, M. P., Escriu, C., Gruver, A., & Gosney, J. R. (2019). Heterogeneity of PD-L1 expression in non-small cell lung cancer: Implications for specimen sampling in predicting treatment response. Lung Cancer, 134, 79-84.

[13] Khorrami, M., Prasanna, P., Gupta, A., Patil, P., Velu, P. D., Thawani, R., ... & Madabhushi, A. (2020). Changes in CT radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non–small cell lung cancer. Cancer immunology research, 8(1), 108-119.

[14] Li, Y., Lu, L., Xiao, M., Dercle, L., Huang, Y., Zhang, Z., ... & Zhao, B. (2018). CT slice thickness and convolution kernel affect performance of a radiomic model for predicting EGFR status in non-small cell lung cancer: a preliminary study. Scientific reports, 8(1), 1-10.

[15] Zhang, J., Zhao, X., Zhao, Y., Zhang, J., Zhang, Z., Wang, J., ... & Han, J. (2020). Value of pre-therapy 18F-FDG PET/CT radiomics in predicting EGFR mutation status in patients with non-small cell lung cancer. European journal of nuclear medicine and molecular imaging, 47(5), 1137-1146.

[16] Karlsson, A., Cirenajwis, H., Ericson-Lindquist, K., Brunnström, H., Reuterswärd, C., Jönsson, M., ... & Staaf, J. (2019). A combined gene expression tool for parallel histological prediction and gene fusion detection in non-small cell lung cancer. Scientific reports, 9(1), 1-13.

[17] Guibert, N., Jones, G., Beeler, J. F., Plagnol, V., Morris, C., Mourlanette, J., ... & Mazieres, J. (2019). Targeted sequencing of plasma cell-free DNA to predict response to PD1 inhibitors in advanced non-small cell lung cancer. Lung Cancer, 137, 1-6.

[18] Rinaldi, L., De Angelis, S. P., Raimondi, S., Rizzo, S., Fanciullo, C., Rampinelli, C., ... & Botta, F. (2022). Reproducibility of radiomic features in CT images of NSCLC patients: an integrative analysis on the impact of acquisition and reconstruction parameters. European radiology experimental, 6(1), 1-13.

[19] Hinestrosa, J. P., Lewis, J. M., Schroder, G., Balcer, H. I., Kurzrock, R., & Krishnan, R. (2022). Detection of early-stage lung cancer using a liquid biopsy test based on extracellular vesicle proteins. Cancer Research, 82(12_Supplement), 727-727.

[20] Mouritzen, M. T., Junker, K. F., Carus, A., Ladekarl, M., Meldgaard, P., Nielsen, A. W., ... & Bjørnhart, B. (2022). Clinical features affecting efficacy of immune checkpoint inhibitors in pretreated patients with advanced NSCLC: a Danish nationwide real-world study. Acta Oncologica, 61(4), 409-416.

[21] Paul, T., Rana, M. K. Z., Tautam, P. A., Kotapati, T. V. P., Jampani, Y., Singh, N., ... & Mosa, A. S. M. (2022). Investigation of the Utility of Features in a Clinical De-identification Model: A Demonstration Using EHR Pathology Reports for Advanced NSCLC Patients. Frontiers in Digital Health, 14.

[22] Zhang, H. (2022). A Comparative Study of Radiomics and Deep-Learning Approaches for Predicting Surgery Outcomes in Early-Stage Non-Small Cell Lung Cancer (NSCLC).

[23] Patel, A. J., Tan, T. M., Richter, A. G., Naidu, B., Blackburn, J. M., & Middleton, G. W. (2022). A highly predictive autoantibody-based biomarker panel for prognosis in early-stage NSCLC with potential therapeutic implications. British journal of cancer, 126(2), 238-246.

[24] Johnson, B. E., Baik, C. S., Mazieres, J., Groen, H. J., Melosky, B., Wolf, J., ... & Planchard, D. (2022). Clinical Outcomes with Dabrafenib Plus Trametinib in a Clinical Trial Versus Real-World Standard of Care in Patients With BRAF-Mutated Advanced NSCLC. JTO clinical and research reports, 3(5), 100324.

[25] Ye, W., Zhang, L., Zhang, W., Wu, X., Yi, D., & Wu, Y. (2022). A comparison of single imputation and multiple imputation methods for missing data in different oncogene expression profiles. Biostatistics & Epidemiology, 6(1), 113-127.

[26] Osinski, B. L., BenTaieb, A., Ho, I., Jones, R. D., Joshi, R. P., Westley, A., ... & Stumpe, M. C. (2022). AI-augmented histopathologic review using image analysis to optimize DNA yield and tumor purity from FFPE slides. arXiv preprint arXiv:2203.13948.

[27] Abushukair, H. M., Al-Kraimeen, L. M., & Saeed, A. (2022). Predictors of response to immune checkpoint inhibitors (ICI) rechallenge post-disease progression in solid tumors: A systematic review and meta-analyses. Journal of Clinical Oncology, 40(16_suppl), 2612-2612.

[28] Laeseke, P., Ng, C., Ferko, N., Naghi, A., Wright, G., Zhang, Y., ... & Pritchett, M. (2022). Overall Survival Associated with Image-Guided Thermal Ablation (IGTA) and Stereotactic Body Radiation Therapy (SBRT) for Patients with Non-Small Cell Lung Cancer: A Systematic Review and Meta-Regression Analysis. A109. THE ODYSSEY, NO LONGER A TRAGEDY: THE CONTINUUM OF LUNG CANCER, A2354-A2354.

[29] Iaconangelo, C., McManus, S., Serrano, D., Podger, L., MA, Y., Zhan, L., ... & Barnes, G. (2022). P5 Anchor-Based Thresholds for Meaningful within-Patient Change in the Phase 3 Trial to Evaluate Tislelizumab for the Treatment of 2/3L NSCLC. Value in Health, 25(7), S288.

[30] Mirhadi, S., Tam, S., Li, Q., Moghal, N., Pham, N. A., Tong, J., ... & Tsao, M. S. (2022). Integrative analysis of non-small cell lung cancer patient-derived xenografts identifies distinct proteotypes associated with patient outcomes. Nature communications, 13(1), 1-17.

[31] Kaur Bijral, R., Singh, I., Manhas, J., & Sharma, V. (2022). Discovery of EGFR kinase's T790M variant inhibitors through molecular dynamics simulations, PCA-based dimension reduction, and hierarchical clustering. Structural Chemistry, 1-8.

[32] Englmeier, F., Bleckmann, A., Brückl, W., Griesinger, F., Fleitz, A., & Nagels, K. (2022). Clinical benefit and cost-effectiveness analysis of liquid biopsy application in patients with advanced non-small cell lung cancer (NSCLC): a modelling approach. Journal of Cancer Research and Clinical Oncology, 1-17.

AUTHORS PROFILE

Mrs. Sumalatha Mani is a Phd Research Scholar in Computer science Department at Periyar University, Salem. Currently she is working as an Assistant professor in Computer science department at C. Kandaswami Naidu College for Women, Cuddalore. Her Research interest includes DataMining, Machine Learning, Deep learning and Computer-aided diagnosis in biomedical science

Dr. Latha Parthiban, is working as an Assistant professor in Department of computer Science at Community college, Pondicherry University, India. She received Bachelors of Engineering in Electronics from Madras University in the year 1994. M. E from Anna University in the year 2008 and Ph D from Pondicherry University in the year 2010.Her research activities mainly focus on Datamining, image processing, computer-aided diagnosis in biomedical science, in her research credited. She published more than 100 journal in Peer reviewed Scopus indexed/SCI journals