EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

Augmentation of Predictive Competence of Non-Small Cell Lung Cancer Datasets through Feature Pre-Processing Techniques

M. Sumalatha [1,*], Latha Parthiban [2]

[1] Research Scholar, Periyar University, Salem, Tamil Nadu, India
[2] Head In-charge, Department of Computer Science, Pondicherry University, Community College, Pondicherry, India

Abstract

The major Objective of the Study is to augment the predictive analytics of Non-Small Cell Lung Cancer (NSCLC) datasets with Feature Pre-Processing (FPP) technique in three stages viz. Remove base errors with common analytics on emptiness or non-numerical or missing values in the dataset, remove repeated features through regression analysis and eliminate irrelevant features through clustering methods. The FPP Model is validated using classifiers like simple and complex Tree, Linear and Gaussian SVM, Weighted KNN and Boosted Trees in terms of accuracy, sensitivity, specificity, kappa, positive and negative likelihood. The result showed that the NSCLC dataset formed after FPP outperformed the raw NSCLC dataset in all performance levels and showed good augmentation in predictive analytics of NSCLC datasets. The research proved that pre-processing is essential for better prediction of complex medical datasets

Keywords: Non-small Cell Lung Cancer, Competency of Prediction, Relevancy Analysis, Regression Analysis, Cluster Analysis, Feature Pre-Processing (FPP Model, Competency Analytics.

Received on 23 June 2021, accepted on 10 October 2022, published on 02 November 2022

doi: 10.4108/eetpht.v8i5.3169

1. Introduction

Non-Small Cell Lung Cancer (NSCLC) is one of the escalating cancers found in many parts of the world. The study aimed to provide a solution to build effective predictive system for NSCLC in complex high dimensional datasets. Medical datasets are generally complex and highly susceptible for utilization in prediction of lung cancer. The NSCLC datasets [1] are complex in nature due to the presence of biomarkers features. The complexity of the features depends on the quantity and quality of the features present in a dataset. Numerous NSCLC datasets were recognized to have high quantity of over 50 features which makes the complexity more hard-hitting scenario. Also, the quality of recorded samples was not appealing as it contained missing values, irrelevant values [2], redundant features that makes the prediction more complicated. Hence the major problem to be addressed in this paper is to minimize the irrelevant features to augment the predictive competency of NSCLC datasets through a series of feature reduction methods collectively represented as Feature Pre-Processing (FPP) using data mining techniques.

The major objective of the research is to increase the predictive competency of Non-small Cell Lung Cancer dataset through Feature Pre-Processing (FPP) techniques and test the performance of prediction of datasets before and after the FPP process. The specific objectives are to identify the class and predictive features of the NSCLC dataset and perform the analytics based on the efficiency, performance and likelihood nature of the dataset after identifying the flaws in the features presented before the FPP stage. The scope of the research work is applied to Non-Small Cell Lung Cancer (NSCLC) datasets collected from primary or secondary sources and applied with the pre-processing techniques modeled within the limit of data mining techniques. The research work focused on the importance of choosing the right features for better prediction. Hence the study is recommended for prediction of complex biological datasets and predict complicated diseases like Non-Small Cell Lung Cancer (NSCLC) that seldom shows symptoms at its earliest stage of infection.

[*] Corresponding author. Email: latha7sumaphd@gmail.com

2. Literature reviews on pre-processing techniques for lung cancer

The research work concentrating on the competency of predictive performance after preprocessing requires study of relevant works completed in different scenarios. Chief pre-processing methods were applied in different medical datasets to counteract the major problems like redundancy, missing values, irrelevancy of data, high dimensionality etc. A deep learning approach to identify the lung cancer using chest x-rays and CT scan images was proposed by Yu. Gordienko et.al (2018) [3] where pre-processing techniques like segmentation, bone shadow exclusion techniques were applied on the BSE-JSRT dataset. The removal of unrelated bone data in the dataset enhanced accuracy of prediction. Choon Sen Seah et.al (2018) [4] developed a pre-processing model called Significant Directed Random Walk (SDRW) in three stages. During the first stage, unwanted attributes were removed along with missing values and arrangement of data. Secondly, the normalization techniques were applied followed by the filtering methods at the third stage. Shigang Liu et.al (2020) [5] identified that Feature Selection with SVM Classifier pre-processing technique has overwhelmed the existing KNN model of Pre-processing methods applied in Biomedical datasets. Biological datasets were highly complex and hence pre-processing methods were expected to be highly reliable.

Various other pre-processing methods were also proposed and tested with biological datasets like lung cancer datasets. Gur Amrit Pal Singh & P. K. Gupta (2018) [6] analyzed the lung cancer CT images using various classifiers like KNN, SVM, decision tree, RFT, and Multi-Layer Perceptron based on 15750 lung images where class variable of dataset classified 6910 as early-stage lung cancer and 8840 as advanced malignant lung cancer respectively. The accuracy after pre-processing methods were found to be highest as 88.55% with MLP model among all other tested models. Anna Meldo et.al (2019) [7] applied Computer Aided Diagnostic (CAD) pre-processing method for Lung cancer on the intellectual dataset called LIRA. A similar automated lung cancer prediction on Kaggle datasets was proposed by Gustavo Perez & Pablo Arbelaez (2020) [8] with an accuracy of 99.6% based on Malignation Prediction Pre-processing test. NegarMaleki et.al (2020) [9] proved that hybrid usage of pre-processing with KNN and Genetic Algorithm could enhance the prediction accuracy of complex lung cancer datasets using classifiers. Chip M. Lynch et.al (2017) [10] showed that statistical methods like Root Mean Squared Error (RMSE) and unsupervised learners like k-means could enhance pre-processing methods and enhance predictions. M. S. Kavitha et.al (2019)[11] utilized pre-processing techniques like Gabor Filter for Lung image enhancement, gaussian filter for smoothing of lung images for effective prediction using classifiers like SVM and Fuzzy C-Mean Clustering. Thus, pre-processing techniques serves a significant role in enhancing the prediction of the lung cancer datasets like Luna16 datasets as shown by Nasibeh and Mortezapour (2019) [12]. They were also able to classify the images into different categories based on the effectiveness of pre-processing methods. Even in the recent analysis by Ankush Kumar Gulia, et. al. (2021)[13], the prediction of lung cancer datasets was proved to be effective in prediction after proper pre-processing techniques applied to the existing models. Some of the Pre-Processing methods and the classifiers used in Lung Cancer in the existing scenario were presented in Table.1

.

Table.1. Existing Pre-Processing methods and classifiers for lung cancer datasets

Ref. No	Author and Year	Pre-Processing Methods	Purpose	Classifiers used for Prediction
[14]	Sindhu Priya & Ramamurthy (2018)	Image enhancement, image segmentation, Feature extraction methods	To enhance quality of images	Classifiers of Image Processing Tool in Matlab
[15]	Kavitha & Prabakaran (2019)	Top-hat transform, median and adaptive bilateral filter	To improve the CT images of Lung to predict cancer cells	Marker-Watershed method based on PSO and Fuzzy C-mean Clustering method
[16]	SurenMakaju et.al (2018)	Segmentation Feature Extraction Classification	To process images thereby removing outliers	Computer aided techniques, Machine Learning methods
[17]	El-Regaily et.al (2018)	Image Pre-Processing, Acquisition, Segmentation process	To improve the quality of lung CT images and to classify them	Nodule detection and false positive reduction Classifiers

[18]	Mohamed Shakeel et.al (2021)	Image denoising methods like feature extraction, segmentation, surface examination	To test the images and remove noisy data.	Improved profuse clustering technique (IPCT)
[19]	Chakravarthy & Rajaguru et.al (2019)	Computed Tomography (CT) Probabilistic Neural Network (PNN)	CT for examining image modality and PNN for classifying the tasks.	Gray-Level Co-Occurrence Matrix (GLCM) and chaotic crow search algorithm (CCSA) based feature selection
[20]	Talha Meraj et.al (2020)	Adaptive Thresholding Technique (ATT) and the semantic segmentation	To remove noise and promote filtering process	Principal components analysis (PCA) and Computerized Tomography (CT)
[21]	Bhalerao et.al (2019)	Conversion of RGB image to gray-scale image. Gray-scale image is further converted to Binary image	To handle complex RGB and Gray scale images in Lung Cancer	Convolution Neural Network. Convolution Filtering, Max Pooling filtering
[22]	YuxinZhou et.al. (2021)	Computed Tomography (CT) & Image pre-processing techniques	Used for Commotion Exclusion and Picture division	Convolutional Neural Network (CNN)

The above pre-processing models were used in complex Lung Cancer datasets to predict the benign and malignant tumors using the position of lymph in cells of Lung.

2.1 Research Gaps of the Study

Based on the analysis on various pre-processing models in the existing scenario, few of the flaws were identified in the existing frameworks.

Some of them are presented as follows:

- The existing models mostly were based on the image pre-processing techniques applied on lung cancer images. The application of pre-processing techniques in numerical analysis were found to be missing among the models.
- A pre-processing framework with sequence of stages were found earlier in Significant Directed Random Walk (SDRW) Choon Sen Seah et.al (2018) [4]. However, the stages were generally made with no specific algorithm generated in novel form.
- The models tested with the classifiers were not benchmarked with the existing methods before and after pre-processing to know the importance of pre-processing in lung cancer datasets
- The datasets utilized in the research comprised of not more than 50 features at a time. The high dimensionality could be addressed better with features at least fifty or above for better scope and reliability of pre-processing techniques
- Supervised and unsupervised models were not tested at the same time during the pre-processing stage. It is essential for effective identification of relevant and irrelevant features in a biological dataset.

.

To counteract all the above disadvantages of the existing pre-processing methods, the novel framework model specifically is designed for pre-processing technique in combination of data mining, regression and clustering methods.

3. Materials and Methods

After analyzing the disadvantages of the existing systems and their flaws, it's important to identify better techniques in combination that would assist in refining all the problems in the given complex NSCLC dataset to encourage better predictions using classifiers. The pre-processing is an important stage of data mining process, where irrelevant, redundant and unprecedented data [23] containing features has to be removed. Hence, various pre-processing techniques has to be performed on the numerical dataset to be used. The dataset would be the high dimensional of minimum fifty features or more with a class feature. The pre-processing architecture proposed in this research work "Feature Pre-Processing (FPP)" is a novel method comprising of three major phases as shown in Figure1

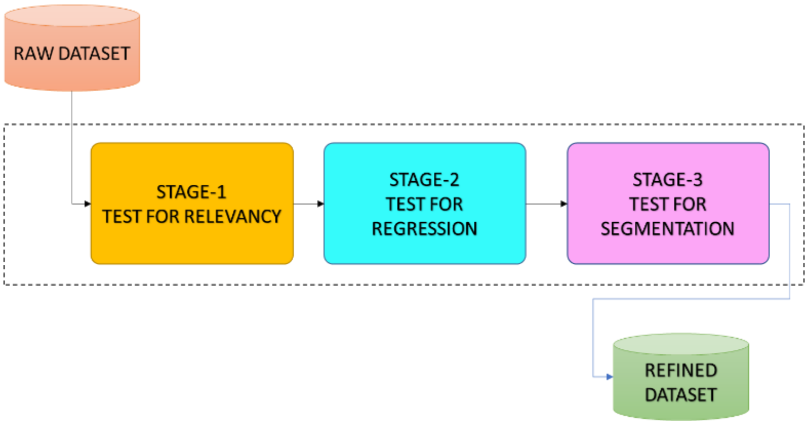


Figure.1. Phases of Feature Pre-Processing (FPP) for NSCLC dataset

As shown in Figure.1, there are three phases in this proposed architecture FPP. The first phase tests the relevancy of the data to the NSCLC dataset based on numeric nature [24], Null data [25] or non-medical data [26] out of range. During the second phase, it tests the regression analysis where the correlation between the existing data and the overall mean data can be measured. Later, suing the third phase, based on the centroid of the dataset, the relevant and irrelevant data could be measured. Thus, after finding the relevance in three different phases, the irrelevant data in each phase corresponding feature can be removed and remaining features can be developed into a best feature set for further processing. The architecture applies to numerical dataset of NSCLC types to the maximum rather than other type of datasets. The dataset should also be of high dimension with huge features above fifty with a class feature to measure the competency of the prediction after reducing the irrelevant features respectively.

3.1 NSCLC Dataset Analysis

The Multivariate Dataset was collected from UCI repository based on the T, N and M values of the predictions on Non-Small Cell Lung Cancer (NSCLC) cells. The dataset was donated by Aeberhard et.al where initially the accuracies were achieved with 62.5% for RDA, 53.1% for KNN, 59.4% for Opt. Disc. Plane [27] respectively. The dataset comprised of 57 features that described three major types of pathological lung cancers. The donor of the dataset did not furnish the name of the features; hence, it is identified as hidden feature. The dataset was found to have unknown values in the dataset indicated as '?' due to non-reliability of unknown data. The feature 1 is found to be the class feature remaining being predictive in type as shown in Table.2.

Table.2. Feature names, types and range of values in NSCLC dataset

S. No	Feature Name	Feature Type	Value Range
1	VarName1	Class Feature	1,2,3
2	VarName2	Predictive	0,1
3	VarName3	Predictive	1,2,3
4	VarName4	Predictive	0,1,2,3
5	VarName5	Predictive	0,1,2
6	VarName6	Predictive	0,1
7	VarName7	Predictive	1,2,3
8	VarName8	Predictive	1,2,3
9	VarName9	Predictive	1,2,3
10	VarName10	Predictive	1,2,3
11	VarName11	Predictive	1,2,3
12	VarName12	Predictive	1,2,3
13	VarName13	Predictive	0,1,2,3
14	VarName14	Predictive	1,2,3
15	VarName15	Predictive	1,2,3
16	VarName16	Predictive	1,2,3
17	VarName17	Predictive	1,2,3
18	VarName18	Predictive	1,2
19	VarName19	Predictive	1,2
20	VarName20	Predictive	0,1,2
21	VarName21	Predictive	0,1,2
22	VarName22	Predictive	1,2
23	VarName23	Predictive	1,2
24	VarName24	Predictive	1,2
25	VarName25	Predictive	1,2,3
26	VarName26	Predictive	1,2,3
27	VarName27	Predictive	1,2,3
28	VarName28	Predictive	2,3
29	VarName29	Predictive	1,2,3
30	VarName30	Predictive	1,2,3
31	VarName31	Predictive	1,2,3
32	VarName32	Predictive	1,2,3
33	VarName33	Predictive	1,2,3
34	VarName34	Predictive	1,2,3
35	VarName35	Predictive	1,2,3
36	VarName36	Predictive	1,2,3
37	VarName37	Predictive	1,2,3

38	VarName38	Predictive	1,2,3
39	VarName39	Predictive	0,1,2
40	VarName40	Predictive	1,2,3
41	VarName41	Predictive	1,2,3
42	VarName42	Predictive	1,2,3
43	VarName43	Predictive	1,2,3
44	VarName44	Predictive	1,2,3
45	VarName45	Predictive	1,2,3
46	VarName46	Predictive	1,2,3
47	VarName47	Predictive	1,2,3
48	VarName48	Predictive	2,3
49	VarName49	Predictive	2,3
50	VarName50	Predictive	1,2,3
51	VarName51	Predictive	1,2,3
52	VarName52	Predictive	1,2,3
53	VarName53	Predictive	1,2,3
54	VarName54	Predictive	1,2,3
55	VarName55	Predictive	1,2
56	VarName56	Predictive	1,2
57	VarName57	Predictive	1,2

As shown in Table.2, the VarName1 is the class feature that represents the outcome received through medical analysis. The remaining 56 features were considered as predictive in nature. The values of the features frequently range from 1 to 3 indicating values of T, N and M values respectively. The predictive features have to be examined for the proposed pre-processing methods to identify the eligible features and remove the irrelevant features to form a best feature set.

3.2 Phases of Feature Pre-Processing

The pre-processing of NSCLC dataset has been carried out with the novel architecture Feature Pre-Processing (FPP) under three different phases as explained with its schematic diagrams, formulations and algorithms.

3.2.1 Phase-I: Test for Relevancy of Features

The first Phase tests the Raw NSCLC dataset to test for relevancy of data through testing the behavioral data like presence of '?' instead of negative data [27], null data and empty set. It also tests the numerical analysis of data as the major research work is to test the numerical data rather than image analysis. The major processes in the Phase-I are represented in the Equation 1.

$$minimize \sum_{k=1}^{n}(fti \neq \{\}||NULL) \qquad (1)$$

Where minimize indicates the minimization of features to form the subset. The data in the dataset ranges from k=1 to n. features where every individual data ft_i is tested for empty, Null or NAN. The Raw dataset initially loaded and normalized before testing for relevancy of data Conditional constructs were applied to test for null values emptiness of the data without any values and also non-numeric data represented as NAN (Not A Number) [28] values in the dataset. If the selected data results true for any one of the above behavioral relevancy problems, the corresponding feature is removed from the dataset and the remaining features are added to the refined or processed NSCLC dataset for further testing process as indicated in Figure.2.

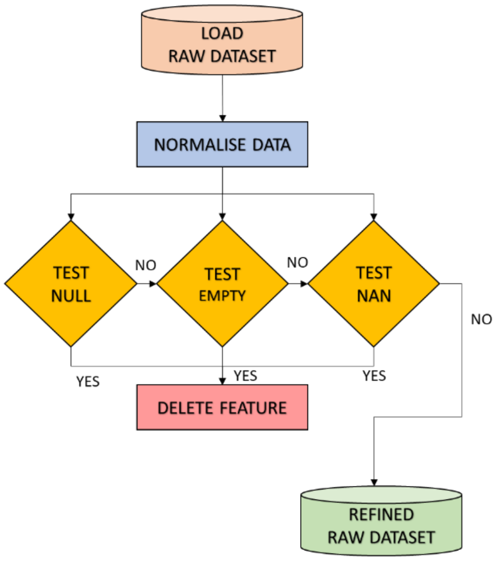


Figure.2. Phase-I: Test for Relevancy of data in NSCLC dataset

The explained process has been given as algorithm in Table.3.

Table.3. Algorithm for Test for Relevancy in Feature Pre-Processing (FPP) Stage

```
Algorithm TFRPP
    Input: NSCLC RAW Dataset A (), A1()
    Define fi ← $features$
    Initialize $n,\ i \leftarrow 1, fn \leftarrow 1,$
             $s \leftarrow size(Features)$
    //Pre-Processing
    For $\forall i \in n$ do
        Normalize (fi)
        If(f(i)==NULL)
        If(f(i)== {})
        If(NAN(fi))
          delete f(i)
        else
          A1 ←A1 + 1
    End For
    //Subset formation
    A ()← A1 ()
End
```

The algorithm in Table.3. shows two datasets A() as source Raw Dataset, A1() as destination Refined Dataset. Other variables include s for size of features, fn as the individual element and n being the total elements in dataset respectively. After initial normalization [29] of the data, every individual element f(i) is tested individually for null, emptiness and

NAN ruleset. If the data falls in any one of the categories, it is identified as corrupted data and the corresponding feature is deleted. After consequent iterations, all the feature elements are tested and the subset A1() is formed from remaining features in source dataset A(). Thus, this stage separates externally identifiable corrupted features and removes them.

3.2.2 Phase-II: Test for Regression Analysis

During the Second Phase, the Raw dataset from first stage is applied with regression analysis to test the relationship that exists between the data in the NSCLC dataset. The regression analysis process is indicated in Equation.2.

$$Minimize\ F = \sum_{f=0}^{n} \mathrm{F} = 0 \sum_{f=0}^{n} (\mathrm{fi} - \bar{\mathrm{X}}) \qquad (2)$$

Where minimize represents the reduction of non-regressive features that shows similarities, F representing the feature set, $\bar{\mathrm{X}}$ representing the mean of data in the dataset, 'fi' the individual data and 'n' the total values in the tested dataset. The overall function of the regression analysis is shown in Figure.4.

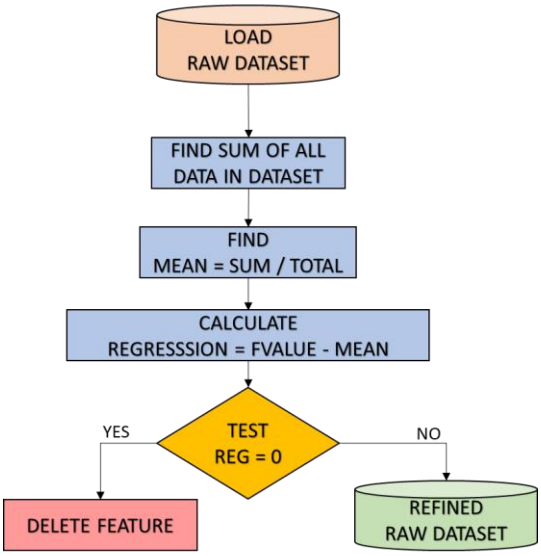


Figure.4. Phase-II: Test for Regression Analysis in Feature Pre-Processing (FPP)

The above Fig.4. indicates that the Raw dataset is loaded in the training platform to test for regression of data. The regression indicates the relationship that exists between the individual data representing a feature. The total sum of all data in the NSCLC data is calculated in the initial stage. Later the mean for the dataset is found by dividing the sum by the number of samples in the dataset [30]. Later every feature is tested to find the correlation by finding difference of every element from the mean value. After finding the regression co-efficient, the value is tested to find if it equals zero. If the condition is true, it is identified as redundant feature and deleted. If the condition is false, they are considered as different features and hence added to new refined dataset. The overall process is presented in the algorithmic form as Table.4.

Table.4: Algorithm for Test for Regression Analysis (TFRGPP)

```
Algorithm TFRGPP
Input: NSCLC Refined Dataset A1 (), A2()
Define fi ← $features$
Initialize $n,\ i \leftarrow 1, fn \leftarrow 1,$
$f1 \leftarrow 0, f2 \leftarrow 0$, reg()
For $\forall i \in n$ do
   sm ← sum(f(i))
End For
mean ← sm / n;
//Fitness Function
For $\forall i \in n$ do
   reg(i) ← f(i) - mean
End For
For $\forall i \in n$ do
If (reg(i) == 0)
        delete f(i)
else
        add (A2(f(i))
End For
```

The Algorithm in Table.4 shows the regression analysis in FPP where the dataset F1() is the raw dataset after first level pre-processing and A2() is the dataset to be created after regression analysis. The fitness function is created to find sum and mean of the entire data in the dataset. Later, the regression co-efficient is calculated by subtracting the individual f(i) value from mean. After testing regression co-efficient value for zero, the selected features are stored in new dataset. This phase separates the redundant features from independent features.

3.2.3 Phase-III: Test for Clustering Analysis

The third and final stage of Pre-Processing applies the segmentation and clustering methods to find the cluster with irrelevant features and another cluster with relevant features respectively. The clustering model applies k-means clustering technique [31] where the major intention is to find the centroid of the dataset and compare it with individual dataset. The overall process of the third phase is indicated in Eq.3.

$$minimize\ (x) = \sum_{i=0}^{n} \mathrm{f(i)} > fc \sum_{i=0}^{n} \frac{\sum f}{n} \qquad (3)$$

where again the minimize is used to indicate the reduction of irrelevant features of total NSCLC dataset (x). The centroid is calculated by identifying the median of the features in statistical analysis and dividing it by the total of values (n). Initially, the Raw dataset from the first phase is loaded for test of segmentation and clustering in the third phase. The median is calculated by find the middle feature among the existing features. The median is the centroid of the entire dataset. After computing the centroid as the threshold value, the fitness function is calculated by subtracting individual feature from the centroid value. Finally, the segmentation fitness co-efficient is tested to find if it is greater than centroid [31]. If the condition is true, it is added to the irrelevant feature set whereas if it is false, it is added to the relevant features refined dataset. The overall process of the third phase is shown in Figure.5 and algorithm represented in Table.5.

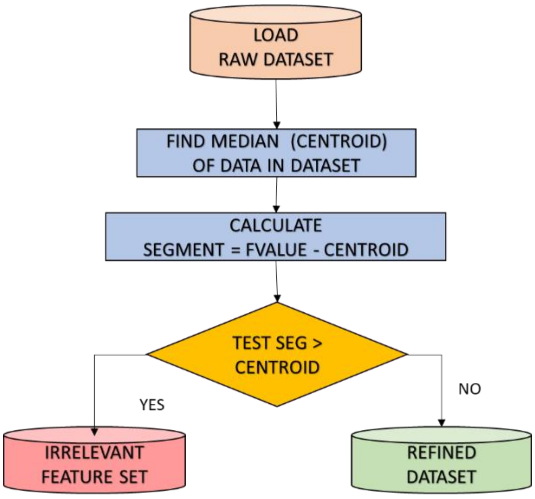


Figure.5. Phase-III: Test for Segmentation and Clustering in FPP

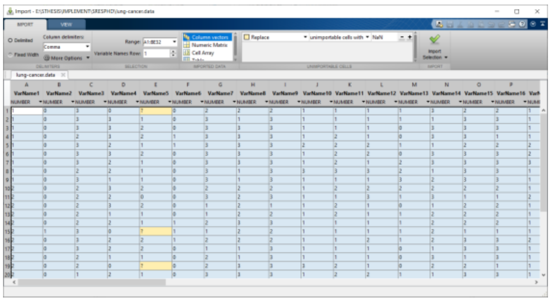
Table.5 Algorithm for Test for Segmentation and Clustering in FPP

```
Algorithm TFSCPP
    Input: NSCLC Refined Dataset A1 (), A3(), A4()
    Define fi ← $features$
    Initialize $n,\ i \leftarrow 1, fn \leftarrow 1, fc \leftarrow 1$
    $f1 \leftarrow 0, f2 \leftarrow 0$, seg()
    For $\forall i \in n$ do
      fc ← median(f(i))
    End For
    //Fitness Function
    For $\forall i \in n$ do
      seg(i) ← f(i) - fc
    End For
    For $\forall i \in n$ do
    If (seg(i) > fc)
          add(A3(f(i))
    else
          add (A4(f(i))
    End For
```

The above algorithm in Table.5. indicates that segmentation and clustering assists in identifying the irrelevant features through differentiating the features using centroid values. It is also applied to hidden data in the existing dataset to identify whether it is required for further processing or can be removed due to non-relevant or non-classifiable nature of the data. Thus, the entire Feature Pre-Processing stage represents a way to perform the best possible method to identify the highly irrelevant features thereby enhancing the predictive analytics of the NSCLC dataset.

4. Experimentation and Evaluation

The first Phase of Implementation was performed in MATLAB tool by loading the initial Raw dataset obtained from the UCI repository sources. The dataset contained errors like missing values '?' as it was clearly witnessed in Figure.6.



○ Delimited
Column delimiters: Comma
Range: A1:BE32
Variable Names Row: 1
Column vectors
☐ Replace unimportable cells with NaN

lung-cancer.data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
	VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11	VarName12	VarName13	VarName14	VarName15	VarName16	VarN
	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMB
1	1	0	3	0	?	0	2	2	2	1	1	1	1	3	2	2	1
2	1	0	3	3	1	0	3	1	3	1	1	1	1	1	3	3	1
3	1	0	3	3	2	0	3	3	3	1	1	1	0	3	3	3	1
4	1	0	2	3	2	1	3	3	3	1	2	1	0	3	3	1	1
5	1	0	3	2	1	1	3	3	3	2	2	2	1	1	2	2	2
6	1	0	3	3	2	0	3	3	3	1	2	2	0	3	3	3	2
7	1	0	3	2	1	0	3	3	3	1	2	1	2	3	3	3	3
8	1	0	2	2	1	0	3	1	3	3	3	3	2	1	3	3	1
9	1	0	3	1	1	0	3	1	3	1	1	1	3	2	3	3	1
10	2	0	2	3	2	0	2	2	2	1	2	2	2	2	2	2	1
11	2	0	2	2	0	0	3	2	3	1	1	3	1	3	1	1	2
12	2	0	2	3	2	0	1	2	1	1	2	1	0	1	2	2	1
13	2	0	2	1	1	0	1	2	2	1	2	1	1	2	2	2	1
14	2	0	2	2	1	1	2	3	3	1	1	1	1	2	2	2	1
15	2	1	3	0	?	1	1	2	2	1	1	1	1	2	1	1	1
16	2	0	3	2	2	1	2	2	2	1	1	2	1	2	3	3	2
17	2	0	3	2	2	0	1	1	3	1	1	1	0	1	3	3	1
18	2	0	2	1	1	0	2	1	3	1	1	1	0	3	1	3	1
19	2	0	2	0	?	0	2	3	3	3	2	1	0	2	2	1	1
20	2	0	1	2	1	0	3	3	3	1	2	2	1	1	3	3	1

Fig.6: Loading the initial Raw NSCLC dataset with errors.

It is evident from Figure.6. that the regions marked with yellow color with '?' are not readable form and hence couldn't be involved in the prediction of NSCLC without proper modification.

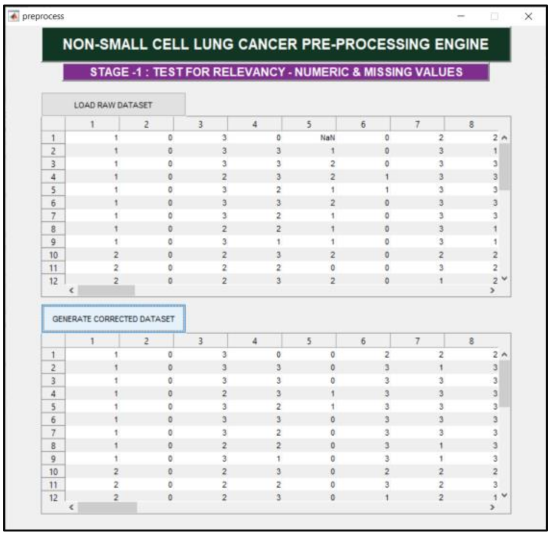
To conduct the test for relevancy of data present in the dataset and the values associated with it. Feature Pre-Processing (FPP) technique is designed in MATLAB and applied with the current dataset. The Raw dataset is loaded in the MATLAB interface and tested for the various pre-processing criterion as mentioned in the Phase-I of FPP known as Test for Relevancy of Data.

To normalize the data based on rows and columns. It was identified that there are 57 features showing columns and 32 records showing rows. The data is stored in the form of an array. Then each feature data is individually tested for numeric or non-numeric nature. Then it is also checked for empty data or negative values in the dataset. Also, the Null condition is checked in the Raw dataset. It was found with the following outcomes as shown in Table.7.

Table.7. Outcomes achieved from NSCLC dataset

Non-Numeric Values	Feature-5 and Feature-39
Empty Values	Nil
Null values	Nil
Negative Values	Nil

It is identified that Features with VarName5 and VarName39 are found to have errors in the form of missing values or non-numeric values. This may affect the quality of the prediction. Hence both the features are removed and new dataset is formed thorough 'Generate Corrected Dataset' as shown in Figure.7.

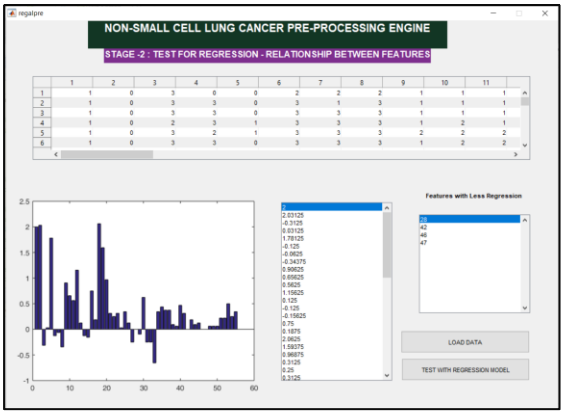

preprocess
NON-SMALL CELL LUNG CANCER PRE-PROCESSING ENGINE
STAGE -1 : TEST FOR RELEVANCY - NUMERIC & MISSING VALUES
LOAD RAW DATASET
GENERATE CORRECTED DATASET

	1	2	3	4	5	6	7	8	
1	1	0	3	0	NaN	0	2	2
2	1	0	3	3	1	0	3	1
3	1	0	3	3	2	0	3	3
4	1	0	2	3	2	1	3	3
5	1	0	3	2	1	1	3	3
6	1	0	3	3	2	0	3	3
7	1	0	3	2	1	0	3	3
8	1	0	2	2	1	0	3	1
9	1	0	3	1	1	0	3	1
10	2	0	2	3	2	0	2	2
11	2	0	2	2	0	0	3	2
12	2	0	2	3	2	0	1	2

	1	2	3	4	5	6	7	8
1	1	0	3	0	0	2	2	2
2	1	0	3	3	0	3	1	3
3	1	0	3	3	0	3	3	3
4	1	0	2	3	1	3	3	3
5	1	0	3	2	1	3	3	3
6	1	0	3	3	0	3	3	3
7	1	0	3	2	0	3	3	3
8	1	0	2	2	0	3	1	3
9	1	0	3	1	0	3	1	3
10	2	0	2	3	0	2	2	2
11	2	0	2	2	0	3	2	3
12	2	0	2	3	0	1	2	1

Figure.7. Refined dataset after loading Raw dataset and removing the features 5 & 39 to form the refined dataset

After the first stage, the refined dataset is formed and this forms the basis for second and third phase evaluation respectively. During the Second Phase, the newly formed excel file was loaded into the MATLAB interface and tested with regression analysis as described in the methodological part of the paper. After testing for regression, the irrelevant features were separately loaded in a list box. In this experiment, it is found to be features 28, 42, 46 and 47 respectively. The regression co-efficient values are listed in another list box with graphical representations as shown in Figure.8.


regalpre
NON-SMALL CELL LUNG CANCER PRE-PROCESSING ENGINE
STAGE -2 : TEST FOR REGRESSION - RELATIONSHIP BETWEEN FEATURES
2.5
2
1.5
1
0.5
0
-0.5
-1
0
10
20
30
40
50
60
Features with Less Regression
LOAD DATA
TEST WITH REGRESSION MODEL

	1	2	3	4	5	6	7	8	9	10	11
1	1	0	3	0	0	2	2	2	1	1	1
2	1	0	3	3	0	3	1	3	1	1	1
3	1	0	3	3	0	3	3	3	1	1	1
4	1	0	2	3	1	3	3	3	1	2	1
5	1	0	3	2	1	3	3	3	2	2	2
6	1	0	3	3	0	3	3	3	1	2	2

2
2.03125
-0.3125
0.03125
1.78125
-0.125
-0.0625
-0.34375
0.90625
0.65625
0.5625
1.15625
0.125
-0.125
-0.15625
0.75
0.1875
2.0625
1.59375
0.96875
0.3125
0.25
0.3125

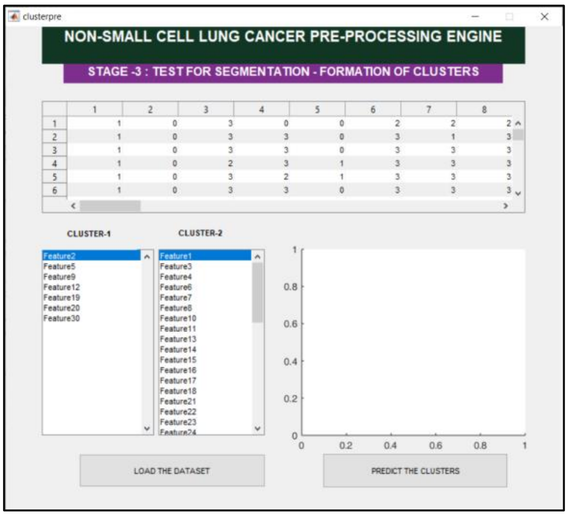
Features with Less Regression

28
42
46
47

Figure.8. Phase-2: Test for Regression in FPP implemented using MATLAB.

The Figure.8 shows that both positive and negative values are not redundant values whereas the regression factor being ‘ZERO’ indicates that the features are similar and will not be useful for prediction. Hence features, VarName28, VarName42, VarName46 and VarName47 can be removed from the dataset to form further refined dataset. However, the changes were not affected before the completion of third phase as both regression and clustering analysis are quite similar in analyzing the irrelevant features of the dataset.

After finding the regression based irrelevant features, it is important to segment and identify the irrelevant features. Hence Phase-III is initiated in MATLAB by loading the Raw dataset obtained after first phase. The dataset loaded is tested to find the centroid value. Later, the centroid value is compared with the individual values of the features in the dataset. the comparison is based on the individual values higher than the centroid values. If higher, they are irrelevant and are moved to a list box whereas lesser features are relevant and moved to another list box in the MATLAB interface. It was identified that the features 2, 5, 9, 12, 19, 20, 30 are found irrelevant and are placed separately as shown in Figure.9.


clusterpre
NON-SMALL CELL LUNG CANCER PRE-PROCESSING ENGINE
STAGE -3 : TEST FOR SEGMENTATION - FORMATION OF CLUSTERS
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
LOAD THE DATASET
PREDICT THE CLUSTERS

	1	2	3	4	5	6	7	8
1	1	0	3	0	0	2	2	2
2	1	0	3	3	0	3	1	3
3	1	0	3	3	0	3	3	3
4	1	0	2	3	1	3	3	3
5	1	0	3	2	1	3	3	3
6	1	0	3	3	0	3	3	3

CLUSTER-1

Feature2
Feature5
Feature9
Feature12
Feature19
Feature20
Feature30

CLUSTER-2

Feature1
Feature3
Feature4
Feature6
Feature7
Feature8
Feature10
Feature11
Feature13
Feature14
Feature15
Feature16
Feature17
Feature18
Feature21
Feature22
Feature23
Feature24

Figure.9: Phase-3: Test for Segmentation and Clustering in FPP implemented using MATLAB.

In the above Figure.9, it is shown that the relevant features are accepted as refined set whereas the irrelevant features VarName2, VarName5, VarName9, VarName12, VarName19, VarName20, VarName30 were removed from the existing dataset. After completing the overall implementation, the entire features identified as irrelevant and unnecessary were removed from the original dataset. The remaining features are considered effective for prediction and tested with classifiers in MATLAB.

5. Results and discussion

The performance comparison is carried out in two phases viz. 1) Before Feature Pre-Processing (BFPP) and 2) After Feature Pre-Processing (AFPP) based on the removal of irrelevant features and testing it with selected classifiers listed below:

- Simple Tree Classifier,
- Complex Tree Classifier,
- Linear SVM,

- Gaussian SVM,
- Weighted KNN,
- Boosted Tree.

In the above classifiers chosen for testing the performance, simple and complex tree [32] is based on non-supervised models, Linear SVM [33], Gaussian SVM [34] are based on supervised models, Weighted KNN [35] and Boosted Tree [36] are optimization models respectively. Hence, they can be better identified as the right classifiers to test the competency nature of complex NSCLC datasets than any other classifiers.

5.1 NSCLC Dataset Competency Analytics

The NSCLC dataset Competency analytics begins with the identification of classifier performance retrieval in the form of confusion matrix achieved after training the models. The total prediction possible in the dataset is 32 values. Hence the limit of the confusion matrix [37] is restricted to 32 values in four criterions viz., 1) True Positive, 2) True negative, 3) False positive and 4) False Negative respectively. According to the Predictive Analytics of NSCLC datasets, True Positive and False Negative [38] are correct predictions whereas the True Negatives and False Positives [39] are wrong predictions.

5.2 Before Feature Pre-Processing

Initially, the RAW dataset without FPP is trained and tested with the classifiers to obtain the accuracy as shown in Table.9.

Table.9. Confusion Matrix and accuracy values of NSCLC dataset before FPP

Classifier	True Positive	True negative	False positive	False Negative	Accuracy
Simple Tree	17	4	2	9	81.3%
Complex Tree	19	3	1	9	87.5%
Linear SVM	15	2	5	10	78.1%
Weighted KNN	17	2	0	13	93.8%
Gaussian SVM	18	3	1	10	87.5%
Boosted Trees	15	6	4	7	68.8%

Based on the confusion matrix values obtained in Table.9, the Competency analysis measures like accuracy, kappa, sensitivity, specificity, positive likelihood and negative likelihood were calculated and summarized in Table.10

Table.10. Competency Analysis of NSCLC dataset before FPP

Competency	Simple Tree	Complex Tree	Linear SVM	Gaussian SVM	Weighted KNN	Boosted Tree
Accuracy	81.3%	87.5%	78.1%	87.5%	93.8%	68.8%
Kappa	0.60	0.72	0.55	0.73	0.87	0.33
Sensitivity	0.82	0.90	0.66	0.91	1	0.64
Specificity	0.81	0.86	0.88	0.86	0.89	0.71
Positive Likelihood	0.008	0.009	0.006	0.009	0.010	0.006
Negative Likelihood	0.008	0.008	0.008	0.008	0.009	0.007

As shown in Table.10, the NSCLC dataset competency analysis before FPP showed results that are recorded as benchmark results to test with the competency analysis of NSCLC dataset after FPP.

5.3 After Feature Pre-Processing

After performing the FPP in three phases, the irrelevant features were removed and a new dataset is created in excel format. That dataset is loaded again in MATLAB to train and test it using the classifiers in the benchmark model. The results obtained in the form of confusion matrix are tabulated in Table.12.

Table.12. Confusion Matrix and accuracy values of NSCLC dataset after FPP

Classifier	True Positive	True negative	False positive	False Negative	Accuracy
Simple Tree	14	4	2	12	81.3%
Complex Tree	14	4	1	13	84.4%
Linear SVM	15	4	0	13	87.5%
Weighted KNN	18	1	0	13	96.9%
Gaussian SVM	16	3	0	13	90.6%

Boosted Trees	18	1	5	8	81.3%

As identified in Table.12, the confusion matrix values are used to find competency analysis measures for the classifiers and presented in Table.13

Table.13. Competency Analysis of NSCLC dataset after FPP

Competency	**Simple Tree**	**Complex Tree**	**Linear SVM**	**Gaussian SVM**	**Weighted KNN**	**Boosted Tree**
Accuracy	81.3%	84.4%	87.5%	90.6%	96.9%	81.3%
Kappa	0.63	0.69	0.75	0.81	0.94	0.95
Sensitivity	0.78	0.78	0.79	0.84	0.95	0.62
Specificity	0.86	0.93	1	1	1	0.64
Positive Likelihood	0.008	0.009	0.010	0.010	0.010	0.006
Negative Likelihood	0.007	0.007	0.008	0.008	0.009	0.009

The Table.13 showed various competency analysis measures as tested in the benchmark model. The obtained results are expected to show good results comparing to the benchmark model. The competency testing has to be performed in the overall comparison and discussions.

6. Discussion and Findings

The overall competency augmentation is measured based on minimization of features and maximization of performance. The overall comparative analysis of the NSCLC dataset based on the performance of the classifiers before and after FPP are summarized in Table.14.

Table.14. Overall comparative competency analysis of NSCLC dataset before and after FPP

Competency	**Simple Tree**		**Complex Tree**		**Linear SVM**		**Gaussian SVM**		**Weighted KNN**		**Boosted Tree**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Accuracy	81.3%	81.3%	87.5%	84.4%	78.1%	87.5%	87.5%	90.6%	93.8%	96.9%	68.8%	81.3%
Kappa	0.60	0.63	0.72	0.69	0.55	0.75	0.73	0.81	0.87	0.94	0.33	0.95
Sensitivity	0.82	0.86	0.90	0.93	0.66	1	0.91	1	1	1	0.64	0.64
Specificity	0.81	0.78	0.86	0.78	0.88	0.79	0.86	0.84	0.89	0.95	0.71	0.62
+ve Likelihood	0.008	0.008	0.009	0.009	0.006	0.010	0.009	0.010	0.010	0.010	0.006	0.006
-ve Likelihood	0.008	0.007	0.008	0.007	0.008	0.008	0.008	0.008	0.009	0.009	0.007	0.009

Based on the comparative analysis in Table.14, various parameters have been analyzed to measure the competency of FPP when applied in NSCLC datasets. The individual analysis on various performance aspects is discussed in three criteria viz., Accuracy and Kappa analysis, Sensitivity and Specificity Analysis, Positive and Negative Likelihood Analysis respectively.

6.1 Accuracy and Kappa Competency Analysis

The Accuracy and Kappa [40] is calculated based on the correct predictions of the classifiers on the dataset. Hence, both are found efficient only when the value increases from the benchmark model. Based on the values in Table.14, it is found that the accuracy of Simple tree remains the same, complex tree shows reduction in prediction whereas the remaining classifiers showed good improvement. The average accuracy enhancement in each of the algorithms are 7%. This showed that accuracy s highly competent after FPP Process. The Accuracy and Kappa Analysis is graphically represented in Figure.10(a) and Figure.10(b).

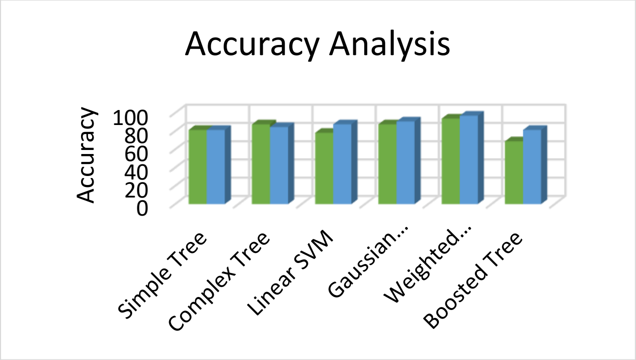


Figure.10(a) Accuracy Competency Analysis

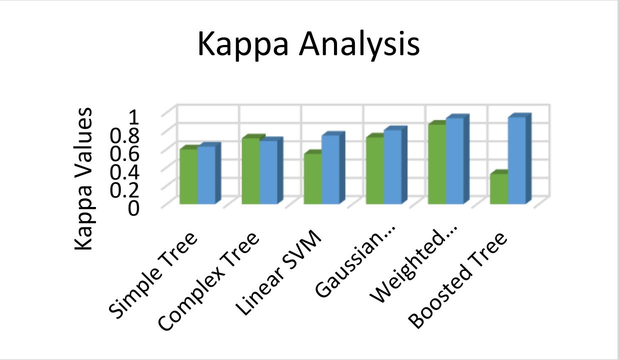


Figure.10(b) Kappa Competency Analysis

The Kappa is found less with complex tree classifier, whereas it showed high competency of improvement in other models. Even simple tree classifier which remained same accuracy showed improvement from 0.90 to 0.63 in kappa competency analysis.

6.2 Sensitivity and Specificity Competency Analysis

The Sensitivity and Specificity [41] of the NSCLC dataset is calculated based on the ability of the classifier to predict the Truth as True data and False as Negative data respectively. The sensitivity of the NSCLC dataset is expected to show improvement for high competency. The sensitivity of the NSCLC dataset has showed improvement in all supervised and unsupervised classifiers whereas it remained the same in optimization models as shown in Figure.11(a).

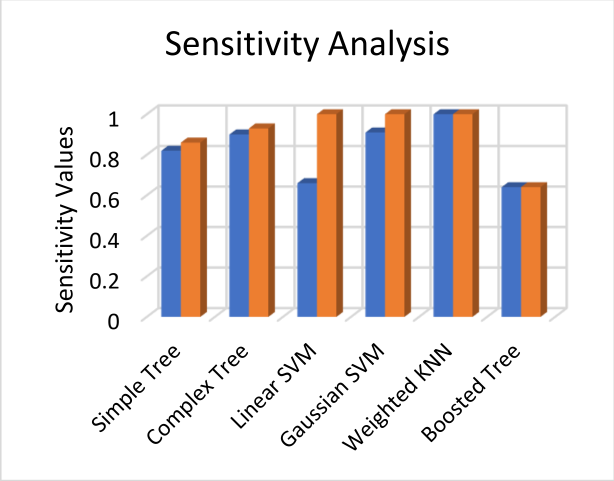


Figure.11(a) Sensitivity Competency Analysis

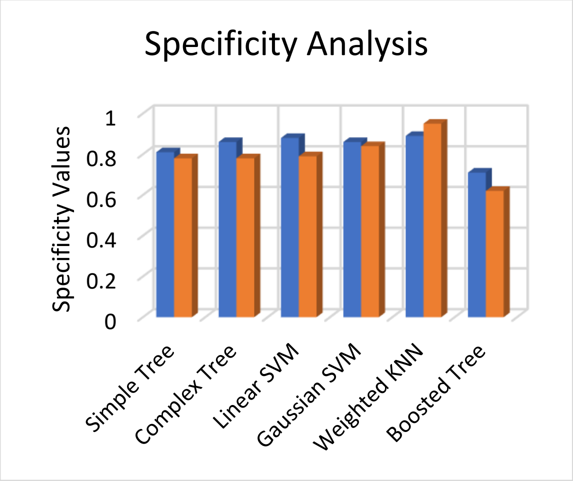


Figure.11(b) Specificity Competency Analysis

However, Specificity as shown in Figure.11(b) showed reduction in values thereby showing good competency in prediction. Since specificity is considered as error in prediction, reduction in value after FPP indicates good prediction. Thus, the test for sensitivity and specificity competency analysis was successful.

6.3 Positive and Negative Likelihood Competency Analysis

The likelihood competency analysis [42] indicates the possibility of the NSCLC dataset to be predicted correctly and wrongly in the future. The positive likelihood showed either improvement or remained the same in each of the classifiers whereas the negative likelihood remained less or remained same in all the classifiers except for boosted tree classifier as shown in Figure.12(a) and Figure.12(b).

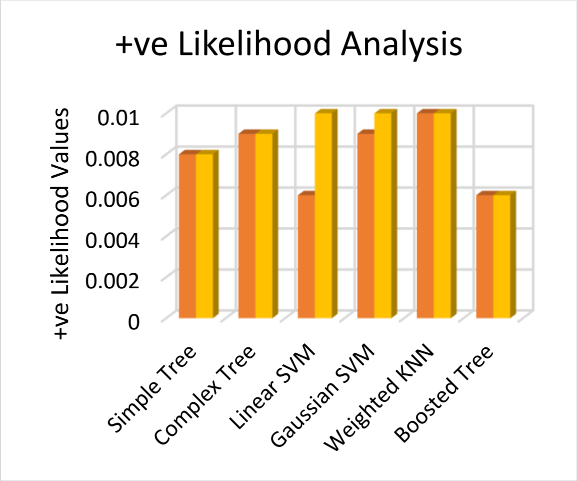


Figure.12(a) Positive Likelihood Competency Analysis

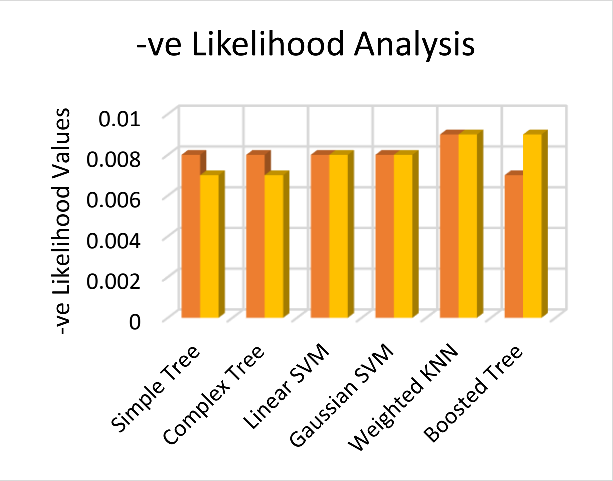


Figure.12(b) Negative Likelihood Competency

Analysis.

Based on the overall Competency Analysis, some of the following findings are identified at the end of the research. The overall competency of NSCLC dataset prediction has been augmented with good performance except for few flaws like reduction of accuracy in Complex Tree and increase of negative likelihood in boosted tree respectively,

The total features reduced in each phase is given in Table.15.

Table.15. Minimization of Features

Total Features before FPP	57
Features reduced after Phase-I	2
Features reduced after Phase-II	4
Features reduced after Phase-III	7
Total features reduced after 3 phases	13
Total Features after FPP	44
Percentage reduction (13/57 * 100)	22.81%

The findings shows that minimization of features is carried out with 22.81% reduction of features in the overall dataset as shown in Table.15. The competency analysis parameters like accuracy, kappa, sensitivity, specificity, positive and negative likelihood are found to be effective in testing with the classifiers. The classifiers included all three categories used in testing models like supervised, unsupervised and optimization models. Thus, it proves that the pre-processing methods are justified with all types of models. The benchmark model with the Raw dataset was outperformed by the proposed model thereby the alternate model is acceptable.

Based on the findings, the proposed model of this research work showed high level of competency in augmentation of prediction with NSCLC datasets.

7. Conclusion

The research work proposed a novel architecture for pre-processing of complicated datasets like NSCLC datasets. The dataset is a hidden type and hence anonymous data was handled as three phases presented in Feature Pre-Processing (FPP) model. The NSCLC dataset was also very complicated in terms of high number of features and the expected high prediction methods for better performance. The research study had three different phases to test the relevancy of data in behavioral, regression and segment-based categories. The overall proposed model showed high competency with the existing NSCLC prediction performance. This model can further be extended with high dimensionality reduction methods and feature extraction methods to provide more competency in the future.

References

[1] Lim, S. B., Tan, S. J., Lim, W. T., & Lim, C. T. (2018). A merged lung cancer transcriptome dataset for clinical predictive modeling. Scientific data, 5(1), 1-8.

[2] Tuor, T., Wang, S., Ko, B. J., Liu, C., & Leung, K. K. (2020). Data selection for federated learning with relevant and irrelevant data at clients. arXiv preprint arXiv:2001.08300.

[3] Gordienko, Y., Gang, P., Hui, J., Zeng, W., Kochura, Y., Alienin, O., ... & Stirenko, S. (2018, January). Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. In International Conference on Computer Science, Engineering and Education Applications (pp. 638-647). Springer, Cham.

[4] Seah, C. S., Kasim, S., Fudzee, M. F., Mohamad, M. S., Saedudin, R. R., Hassan, R., ... & Atan, R. (2018). An effective pre-processing phase for gene expression classification. Indonesian Journal of Electrical Engineering and Computer Science, 11(3), 1223.

[5] Liu, S., Zhang, J., Xiang, Y., Zhou, W., & Xiang, D. (2020). A study of data pre-processing techniques for imbalanced biomedical data classification. International Journal of Bioinformatics Research and Applications, 16(3), 290-318.

[6] Singh, G. A. P., & Gupta, P. K. (2019). Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Computing and Applications, 31(10), 6863-6877.

[7] Meldo, A., Utkin, L., Lukashin, A., Muliukha, V., & Zaborovsky, V. (2019, November). Database acquisition for the lung cancer computer aided diagnostic systems. In 2019 25th Conference of Open Innovations Association (FRUCT) (pp. 220-227). IEEE.

[8] Perez, G., & Arbelaez, P. (2020). Automated lung cancer diagnosis using three-dimensional convolutional neural networks. Medical & Biological Engineering & Computing, 58, 1803-1815.

[9] Maleki, N., Zeinali, Y., & Niaki, S. T. A. (2021). A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. Expert Systems with Applications, 164, 113981.

[10] Lynch, C. M., van Berkel, V. H., & Frieboes, H. B. (2017). Application of unsupervised analysis techniques to lung cancer patient data. PLoS One, 12(9), e0184370.

[11] Kavitha, M. S., Shanthini, J., & Sabitha, R. (2019). ECM-CSD: an efficient classification model for cancer stage diagnosis in CT lung images using FCM and SVM techniques. Journal of medical systems, 43(3), 73.

[12] Esmaeilishahmirzadi, N., & Mortezapour, H. (2018). A novel method for enhancing the classification of pulmonary data sets using generative adversarial networks.

[13] Gulia, A. K. (2021). Lung Cancer Prediction Using Machine Learning Classifiers. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(12), 1665-1672.

[14] Priya, S. S., & Ramamurthy, B. (2018). Lung cancer detection using image processing techniques. Research Journal of Pharmacy and Technology, 11(5), 2045-2049.

[15] Kavitha, P., & Prabakaran, S. (2019). A novel hybrid segmentation method with particle swarm optimization and fuzzy c-mean based on partitioning the image for detecting lung cancer.

[16] Makaju, S., Prasad, P. W. C., Alsadoon, A., Singh, A. K., & Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125, 107-114.

[17] El-Regaily, S. A., Salem, M. A., Abdel Aziz, M. H., & Roushdy, M. I. (2018). Survey of computer aided detection systems for lung cancer in computed tomography. Current Medical Imaging, 14(1), 3-18.

[18] Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2019). Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks. Measurement, 145, 702-712.

[19] Sannasi Chakravarthy, S. R., & Rajaguru, H. (2019). Lung cancer detection using probabilistic neural network with modified crow-search algorithm. Asian Pacific journal of cancer prevention: APJCP, 20(7), 2159.

[20] Meraj, T., Rauf, H. T., Zahoor, S., Hassan, A., Lali, M. I., Ali, L., ... & Shoaib, U. (2019). Lung nodules detection using semantic segmentation and classification with optimal features. Neural Computing and Applications, 1-14.

[21] Bhalerao, R. Y., Jani, H. P., Gaitonde, R. K., & Raut, V. (2019, March). A novel approach for detection of lung cancer using digital image processing and convolution neural networks. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 577-583). IEEE.

[22] Zhou, Y., Lu, Y., & Pei, Z. (2021). Accurate diagnosis of early lung cancer based on the convolutional neural network model of the embedded medical system. Microprocessors and Microsystems, 81, 103754.

[23] Guirgis, H. M. (2018). The impact of PD-L1 on survival and value of the immune check point inhibitors in non-small-cell lung cancer; proposal, policies and perspective. Journal for immunotherapy of cancer, 6(1), 1-6.

[24] Scungio, M., Stabile, L., Rizza, V., Pacitto, A., Russi, A., & Buonanno, G. (2018). Lung cancer risk assessment due to traffic-generated particles exposure in urban street canyons: A numerical modelling approach. Science of The Total Environment, 631, 1109-1116.

[25] Buizza, G., Toma-Dasu, I., Lazzeroni, M., Paganelli, C., Riboldi, M., Chang, Y., ... & Wang, C. (2018). Early tumor response prediction for lung cancer patients using novel longitudinal pattern features from sequential PET/CT image scans. Physica Medica, 54, 21-29.

[26] Wang, T., Gong, J., Duan, H. H., Wang, L. J., Ye, X. D., & Nie, S. D. (2019). Correlation between CT based radiomics features and gene expression data in non-small cell lung cancer. Journal of X-ray Science and Technology, 27(5), 773-803.

[27] Hong, Z. Q., & Yang, J. Y. (1991). Optimal discriminant plane for a small number of samples and design method of classifier on the plane. pattern recognition, 24(4), 317-324.

[28] Gainor, J. F., Curigliano, G., Kim, D. W., Lee, D. H., Besse, B., Baik, C. S., ... & Subbiah, V. (2020). Registrational dataset from the phase I/II ARROW trial of pralsetinib (BLU-667) in patients with advanced RET fusion+ non-small cell lung cancer (NSCLC). Chemotherapy, 71(100), 0.

[29] Reddy, V. K. (2020). Analysis of single cell RNA seq data to identify markers for subtyping of non-small cell lung cancer.

[30] Daoud, J. I. (2017, December). Multicollinearity and regression analysis. In Journal of Physics: Conference Series (Vol. 949, No. 1, p. 012009). IOP Publishing.

[31] Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. J—Multidisciplinary Scientific Journal, 2(2), 226-235.

[32] Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Systems with Applications, 89, 52-65.

[33] Kashef, R. (2021). A boosted SVM classifier trained by incremental learning and decremental unlearning approach. Expert Systems with Applications, 167, 114154.

[34] Xue, Y., Zhang, L., Wang, B., Zhang, Z., & Li, F. (2018). Nonlinear feature selection using Gaussian kernel SVM-RFE for fault diagnosis. Applied Intelligence, 48(10), 3306-3331.

[35] Ma, Y., Xie, Q., Liu, Y., & Xiong, S. (2019). A weighted KNN-based automatic image annotation method. Neural Computing and Applications, 1-12.

[36] Martinek, P., & Krammer, O. (2018). Optimising pin-in-paste technology using gradient boosted decision trees. Soldering & Surface Mount Technology.

[36] Bharati, S., Podder, P., Mondal, R., Mahmood, A., & Raihan-Al-Masud, M. (2018, December). Comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer. In International Conference on Intelligent Systems Design and Applications (pp. 447-457). Springer, Cham.

[37] Hunt, J. S., Cock, C., & Symonds, E. L. (2021). A True Positive and a False Negative? The Dilemma of Negative Colonoscopy After a Positive Fecal Occult Blood Test. Digestive Diseases and Sciences, 1-7.

[38] Panagiotidis, E., Paschali, A., Xourgia, X., & Chatzipavlidou, V. (2020). False-Positive 18F-PSMA-1007 and True-Negative 18F-Fluorocholine PET/CT Splenic Hemangioma. Clinical Nuclear Medicine.

[39] Sasikala, B. S., Biju, V. G., & Prashanth, C. M. (2017, May). Kappa and accuracy evaluations of machine learning classifiers. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 20-23). IEEE.

[40] Jaeger, J. (2018). Digit symbol substitution test: the case for sensitivity over specificity in neuropsychological testing. Journal of clinical psychopharmacology, 38(5), 513.

[41] Laporte, S., & Briers, B. (2019). Similarity as a Double-Edged Sword: The Positive and Negative Effects of Showcasing Similar Previous Winners on Perceived Likelihood of Winning in Sweepstakes. Journal of Consumer Research, 45(6), 1331-1349.