ISSN 2278-3091

Volume 10, No.3, May - June 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse791032021.pdf

https://doi.org/10.30534/ijatcse/2021/801032021

Lung Cancer Detection Using Chi-Square Feature Selection and Support Vector Machine Algorithm

Vikas[1], Dr. Prabhpreet Kaur[2]

MTech, Student, Department of Computer Engineering and Technology, Guru Nanak Dev University Amritsar, India, vksverka94@gmail.com

[2] Assistant Professor, Department of Computer Engineering and Technology, Guru Nanak Dev University Amritsar, India, prabhpreet.cst@gndu.ac.in

ABSTRACT

Lung Cancer is the most general type of disease in the world of cancer. It affects the lungs of the human body. So, the prediction of lung cancer at its earlier stage is difficult. It is the deadliest cause of death in both men and women. Its symptoms are harder to recognize in the initial stages. Machine learning algorithms have made the prediction and detection of lung cancer easier. Chi-square is used for feature selection to select the relevant features in the lung cancer dataset. Different Machine Learning algorithms are used to predict Lung Cancer. The algorithms utilized in the proposed work are SVM and Random Forest. We have compared these algorithms with and without feature selection (Chi-square). SVM is identified as the best algorithm in the proposed work due to its accuracy and less execution time for detecting the model. The key objective of this paper is to enhance the accuracy and reduce the execution time of the classifier.

Keywords: Lung cancer detection, Chi-Square, Support Vector Machine, Random Forest.

1. INTRODUCTION

Lung cancer is a harmful disease in the whole world. It is caused when the uncontrollable cell increase in size out of control in the lungs. It starts through the bronchi, which is considered the main path into the lungs and spread to the various organs of the body.

19 different types of cancer have been reported in a study that can affect the human body. It has been estimated that 1.7 million people die due to lung cancer disease per year [16]. Firstly Doctors analyze the patient and its stage then prescribe different types of surgeries, therapies, and radiography which will eventually destroy or stop the expansion of cells in the particular part of the body. This process is highly expensive and time-consuming [14].

The major factor of Lung Cancer in the human body is smoking and air pollution which affects the Lungs. Lung cancer is hard to find in its initial stage. Thus, is it necessary to detect lung cancer at an earlier stage so that many lives could be saved through Computed Tomography (CT) scans, x-rays, and Magnetic Resonance Imaging (MRI). Since the dataset consists of noise it turns out to be hard to detect. Hence, preprocessing is applied to the dataset to remove irrelevant data and Machine Learning Algorithms are used to predict Lung Cancer [2].

There are two types of Lung Cancer:

1. Small Cell Lung Cancer (SCLC): Small Cell Lung Cancer is fast-growing cancer usually caused by smoking and tobacco. SCLC is often revealed when it is spread to the various parts of the body. It spreads faster than NSCLC. Since it is very aggressive it needs immediate treatment.

2. Non-Small Cell Lung Cancer (NSCLC): Most cases of lung cancer are of NSCLC. It does not spread to related organs. This type of Cancer does not need an immediate cure as it is not aggressive.

2. RELATED LITERATURE

In this section, we have proposed a Literature review of Lung cancer Prediction which involves various machine learning algorithms such as Naïve Bayes, SVM, K-NN, Decision Tree, Gradient-Boosting, etc. Some of the Literature reviews are as follows:

P. Mohamed et al [5] detects lung cancer by an improved deep neural network (IDNN). The main objective of this paper is to Increase the accuracy with optimized feature selection. Two steps used in this paper are the multilevel brightness preserving approach by CT image preprocessing and Improved deep neural network learning (IDNN) using Lung Image segmentation. The data is taken from 3500 images, out of which 2543 are chosen for the testing. And, the model analyzes lung cancer with 97.6% accuracy in Ensemble Classifier.

Radhika P R et al [13] show the comparative study of various classification algorithms such as logistic regression, Naïve Bayes, Decision Tree, and SVM using Lung Cancer Dataset, and the algorithms are compared. They evaluate the performance of each classification algorithms. Support Vector Machine has achieved the highest accuracy of 99.2%.

Gur Amrit Pal et al [16] have proposed a model in which they applied multilayer perceptron and neural network to detect lung cancer and achieved better accuracy. They have used four parameters which are accuracy, F1 score, precision, and recall. They have utilized 15750 clinical images which included 6910 benign and 8840 malignant lung cancer images for training and testing. And at last, they founded the Multilayer perceptron as a better classifier with 88.55% accuracy.

Negar Maleki et al [12] applied K-Nearest Neighbor (KNN) with feature selection (Genetic algorithm) to predict lung cancer with three stages is Low, Medium, and high. They have found that by using feature selection (Genetic algorithm) accuracy is increasing efficiency. 6 attributes are selected through the Genetic Algorithm and the best value of k was 6. This model provides 99.80% accuracy

Nikita Banerjee et al [2] proposed the extracted features have been trained through Machine Learning Algorithms such as Support Vector Machine (SVM), Random Forest, and Artificial Neural Network (ANN). Then the various parameters that are accuracy, Precision, and recall are evaluated to check whose accuracy is the best. And the artificial Neural Network (ANN) obtained the highest accuracy i.e. 96%.

Table 1: Comparison with the past researchers

Author	**Year**	**Dataset**	**Techniques**	**Advantages**	**Disadvantages**	**Accuracy**
KesavKancherla [9]	2013	CT Scan images	Nucleus Segmentation Based Features	Improve the accuracy using modified Features	Computational cost is not minimized.	87.8% (Bagging RBF Network)
Maciej Zieba[17]	2013	44 imbalanced dataset	Boosted SVM	Boosted SVM is used to solve imbalanced data	Boosted SVM does not efficiently reduce the false rate	87.9% (Boosted SVM)
Samuel Hawkins [7]	2014	CT Scan images	Relief- F	Improve the accuracy and Precision rate	The error rate is not minimized	77.5% (DT)
Jinsa Kuruvilla [10]	2014	CT Scan images	Back Propagation Network	Provide Good Accuracy and Provide low mean square error	Reduce the sensitivity parameter	93.3% (Neural Network)
Kui Liu [11]	2017	CT Scan Images	Multi-View Convolutional Neural Networks	Reduce the Classification time	Failed to achieve accurate prediction with attributes	_
Gur Amrit Pal [16]	2018	15750 Clinical images	Multi-Layer perceptron	Improve the recall parameter	Hard to recognize the number of neurons and layers	88.5% (MLP)
MuhammdImran[5]	2018	Lung Cancer Dataset	Majority Voting	Increase the efficiency and precision of lung cancer detection	Reduce the Recall Parameter	88.57% (MLP+ GBT+ SVM)
OzgeGunaydin [6]	2019	Standard digital image	Principal Component Analysis (PCA)	Minimize the classification time	Feature selection is not efficiently applied due to noise	82.43% (ANN)
Siddharth Bhatia [4]	2019	CT Scans	Deep Residual Networks	Improve the accuracy with feature extraction	The error rate is not effectively minimized	84% (EsembleMethods)
Jafar A. Alzubi [1]	2019	Thoracic Surgery Dataset	Weight optimize Neural Network with maximum likelihood boosting technique (WONN-MLB)	Improve the classification accuracy and achieve the classification time, minimum false rate.	This technique is not implemented on a small dataset	93% (WONN-MLB)
Radhika P R [13]	2019	Lung Cancer Dataset	SVM	Improve the accuracy on a small dataset	Difficult task to choose optimal Kernal by SVM	99.2% (SVM)
Nikita Banerjee [2]	2020	CT Scan Imagery data	ANN	ANN technique provides good accuracy and precision rate	By using the ANN technique, the Recall rate is very low	96% (ANN)
Kyamelia Roy [14]	2020	CT Scan Images	Speeded Up Robust Features (SURF)	Achieve better accuracy with feature extraction	Recall and Sensitivity rate is low	94.5% (SVM)
SubratoBharati[3]	2020	Lung Cancer Dataset	Principal Component Analysis (PCA)	Provide maximum Kappa statistics value, Roc value	Minimize the F1 parameter	57.04 % (Naïve Bayes)
P. Mohamed [15]	2020	Lung Cancer Images	Improved deep neural network (IDNN)	Increase the accuracy with optimized feature selection	Computational cost is not reduced	97.6% (Ensemble classifier)
Negar Maleki[12]	2020	Lung Cancer Dataset	Genetic Algorithm	Provide better accuracy and minimize the computational time	Complexity is not minimized	99.80% (K-NN)
Our Proposed Methodology	2021	Lung Cancer Dataset	Chi-Square	Provide better accuracy and precision, reduce computational time	_	98% (SVM)

3. PROBLEM DESCRIPTION

Many issues have been founded in the literature for lung cancer detection. Most of the researchers have used the meta-heuristic (Genetic algorithm). Since the genetic algorithm uses fitness value and due to reptation of it, genetic algorithm is time-consuming. As it gradually takes more time to provide the outcome.

To overcome this problem, we have issues with the statistical test (chi-square). Results show that chi-square is more efficient and effective to predict cancer detection.

4. OBJECTIVES

The objective of the given research work is mentioned below:

- Our proposed work is based on the Support Vector Machine and Random Forest techniques for Lung Cancer Detection.
- To identify the relevant attributes by using the chi-square feature selection technique.
- Compare the Support Vector Machine and Random Forest model by using with and without Chi-Square Feature Selection.
- Increase the Accuracy, Precision, and Recall Rate.
- Reduce the execution time of the model.

5. The Proposed Approaches

5.1 Feature Selection

The feature Selection Method is used to reduce the number of irrelevant features. The main aim of feature selection is to recognize the best attributes that are useful in increasing the accuracy of the model. In this proposed work, we use the statistical test (chi-Square) for the Feature selection.

5.1.1 Chi-square Feature selection

A Chi-square test is commonly used in testing the connection among two categorical outcome attributes in feature selection. It allows you to test whether the attributes are independent or not. We have a minimum chi-square value when two features are independent and the outcome count is approximate to the expected value. The hypothesis of independence is false when the chi-square value is high. More chi-square value indicates that the feature is more dependent on response and also it can be utilized for model training.

$$X^2 = \sum (O_i - E_i)^2 / Ei$$

$\sum$ = the sum of observed and expected value

O = Observed Value

E = Expected Value

5.2 Support Vector Machine

SVM is the most successful method to predict lung cancer. SVM is a machine learning algorithm that examines the classification and regression data, but mostly it is used for classification. SVM is more appropriate where the minimum dataset is used.

SVM creates a decision boundary between the two classes. Various decision boundaries separate the two classes but we have to find the best decision boundary is known as the hyperplane of the SVM. The main aim is to find the hyperplane which has the maximum distance between the data points of the two classes. The following figure 1 shows the margin and hyperplane of two classes:

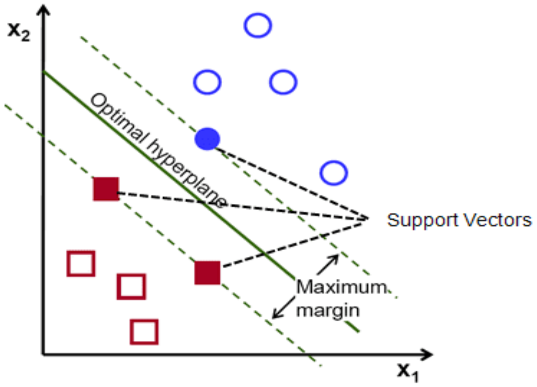


Figure 1: Support vector machine

The two types of Support Vector Machine are:

- **Linear SVM:** At the point when we can easily divide the data by using one or more hyperplanes is Linear SVM.
- **Non-Linear:** At the point when we can't divide the data with a straight line then we utilize the Non-Linear SVM. By using the Kernel function, we can change the Non-Linear data into linear data.

5.3 Random Forest

Random Forest is a classifier that is a collection of decision trees in a given dataset. If the dataset includes very inconsistent features, out of which very few attributes are useful for the model [8]. Random Forest takes a prediction from every decision tree and combines them based on majority voting and provides the outcomes. Below Figure 2 is the explanation of the Random Forest algorithm:

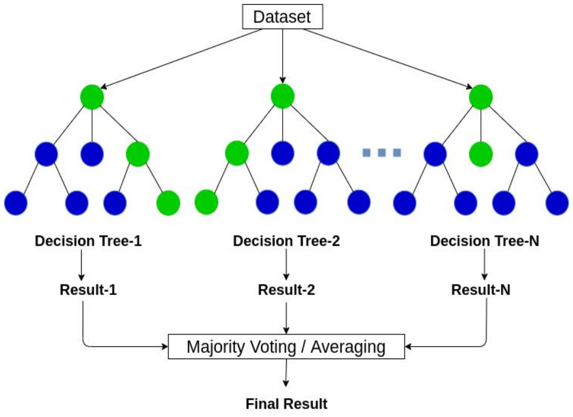
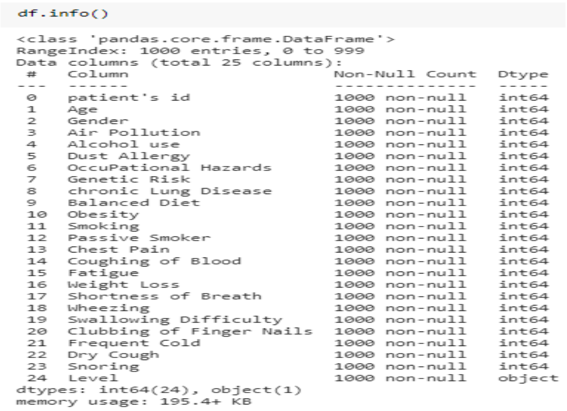


Figure 2: Random Forest

5.4 Data Description

The dataset used in this research paper is taken from the site https://data.world/cancerdatahp/lung-cancer-data.that consists of 1000 samples and 25 attributes.[18] The Following figure 3 is the information on the proposed dataset.



```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 25 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   patient's id               1000 non-null   int64
 1   Age                        1000 non-null   int64
 2   Gender                     1000 non-null   int64
 3   Air Pollution              1000 non-null   int64
 4   Alcohol use                1000 non-null   int64
 5   Dust Allergy               1000 non-null   int64
 6   OccuPational Hazards       1000 non-null   int64
 7   Genetic Risk               1000 non-null   int64
 8   chronic Lung Disease       1000 non-null   int64
 9   Balanced Diet              1000 non-null   int64
 10  Obesity                    1000 non-null   int64
 11  Smoking                    1000 non-null   int64
 12  Passive Smoker             1000 non-null   int64
 13  Chest Pain                 1000 non-null   int64
 14  Coughing of Blood          1000 non-null   int64
 15  Fatigue                    1000 non-null   int64
 16  Weight Loss                1000 non-null   int64
 17  Shortness of Breath        1000 non-null   int64
 18  Wheezing                   1000 non-null   int64
 19  Swallowing Difficulty      1000 non-null   int64
 20  Clubbing of Finger Nails   1000 non-null   int64
 21  Frequent Cold              1000 non-null   int64
 22  Dry Cough                  1000 non-null   int64
 23  Snoring                    1000 non-null   int64
 24  Level                      1000 non-null   object
dtypes: int64(24), object(1)
memory usage: 195.4+ KB
```

Figure 3: Dataset Information

The purpose of the dataset is the target attribute which is labeled into three stages i.e. Low, Medium, and High. Figure 4 shows the purpose of the data set as the target attribute.

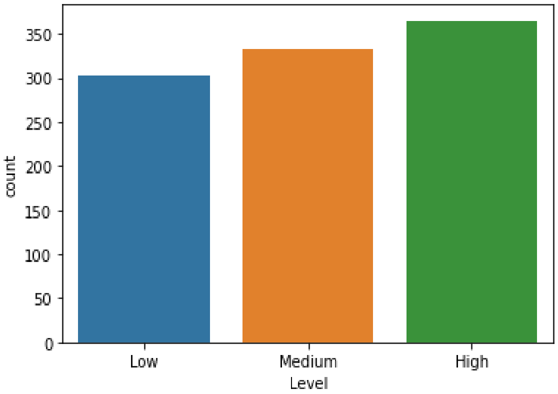
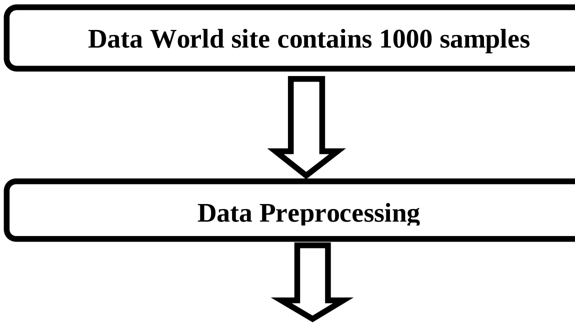
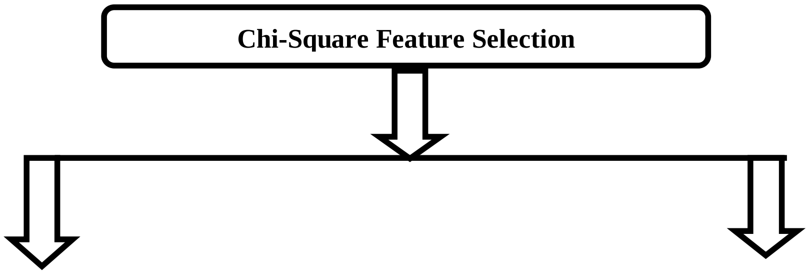


Figure 4: Target Attribute

6. PROCEDURE OF THE PROPOSED METHODOLOGY

In this section, we have listed the process comprises of a few stages from gathering the dataset to handling it and finishing different stages to accomplish the required outcomes. The illustration of the proposed paper is given below through flowchart in figure 5:





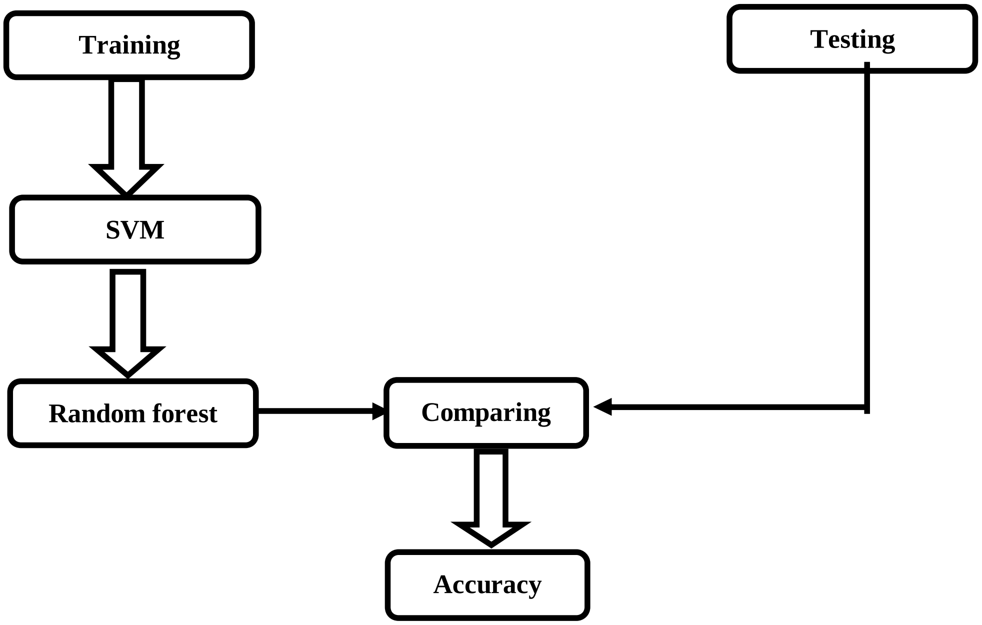
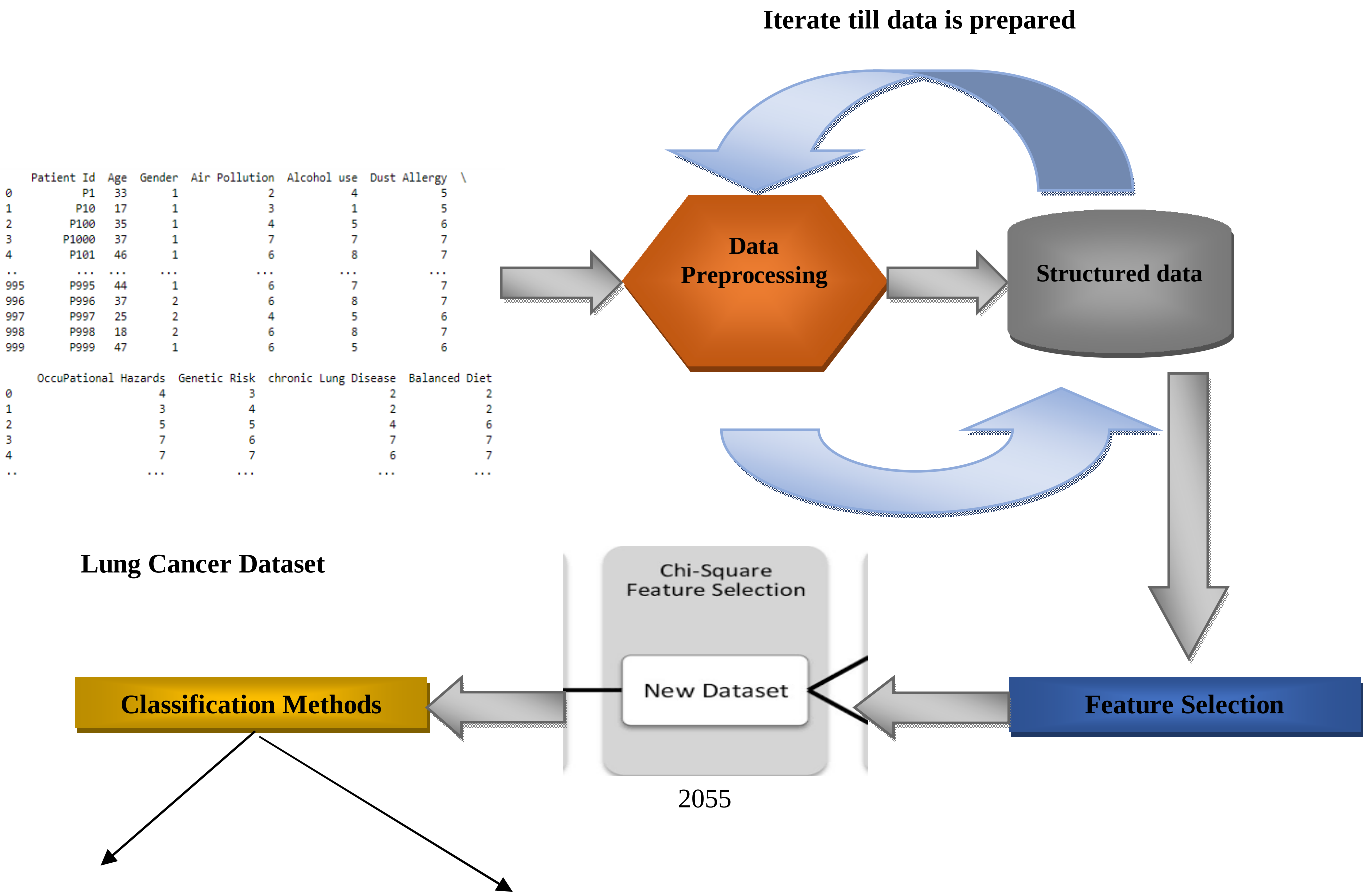


Figure 5: Flow chart of Proposed Approach

7. DIAGRAM OF THE PROPOSED METHODOLOGY

The diagram of the proposed methodology is shown in figure 6. The figure indicates how the model is built for the evaluation and the working of machine learning algorithms applied.



```
    Patient Id  Age  Gender  Air Pollution  Alcohol use  Dust Allergy  \
0           P1   33       1              2            4             5
1          P10   17       1              3            1             5
2         P100   35       1              4            5             6
3        P1000   37       1              7            7             7
4         P101   46       1              6            8             7
..         ...  ...     ...            ...          ...           ...
995       P995   44       1              6            7             7
996       P996   37       2              6            8             7
997       P997   25       2              4            5             6
998       P998   18       2              6            8             7
999       P999   47       1              6            5             6

     OccuPational Hazards  Genetic Risk  chronic Lung Disease  Balanced Diet
0                       4             3                     2              2
1                       3             4                     2              2
2                       5             5                     4              6
3                       7             6                     7              7
4                       7             7                     6              7
..                    ...           ...                   ...            ...
```

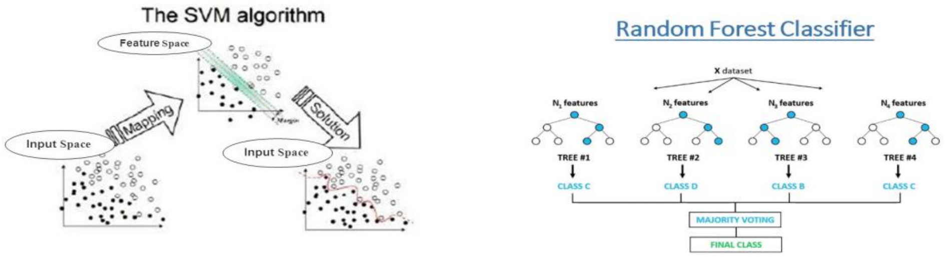
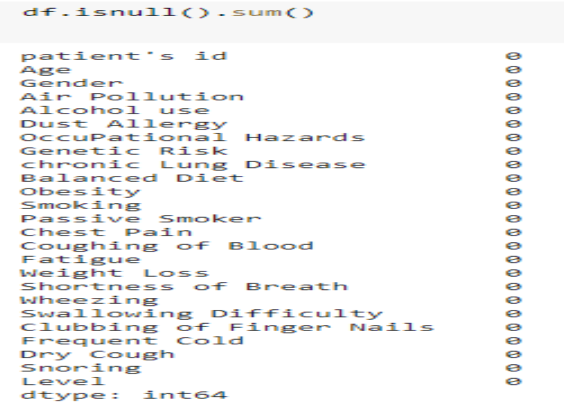


Figure 6: Diagram of the proposed methodology

In the Lung Cancer dataset, the SVM gives better accuracy and less time to classify the model as compared to SVM with Chi-square feature selection. Firstly, the dataset will be gathered and examined. After examining the dataset, preprocessing step is required to remove the null and duplicate values. Data Preprocessing is the initial step of making the model. It is a method in which the preparation of raw data takes place to make it available for machine learning models. In simple words, it is a process that helps data in transforming in such a way that the machine can analyze it. Data Preprocessing is an important step in machine learning because the quality of data is based on this process.

The main aim of the data preprocessing is to remove the irrelevant features, reduce the training time, and improves the performance of the classifier. In the lung cancer dataset, no null values exist and show only one duplicate value which is dropped in the data cleaning step.

Following Figure 7 shows the dataset without any null values.

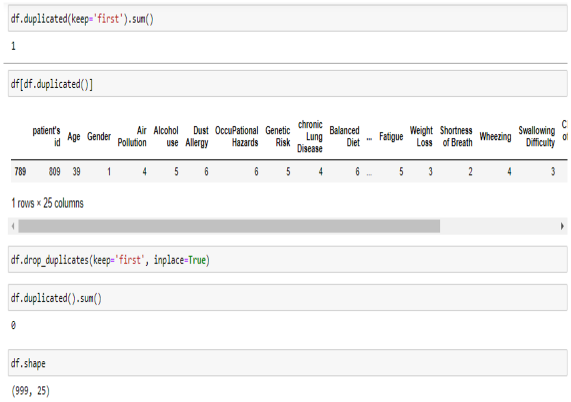


```
df.isnull().sum()

patient's id                    0
Age                             0
Gender                          0
Air Pollution                   0
Alcohol use                     0
Dust Allergy                    0
OccuPational Hazards            0
Genetic Risk                    0
chronic Lung Disease            0
Balanced Diet                   0
Obesity                         0
Smoking                         0
Passive Smoker                  0
Chest Pain                      0
Coughing of Blood               0
Fatigue                         0
Weight Loss                     0
Shortness of Breath             0
Wheezing                        0
Swallowing Difficulty           0
Clubbing of Finger Nails        0
Frequent Cold                   0
Dry Cough                       0
Snoring                         0
Level                           0
dtype: int64
```

Figure 7: Screenshot without any null values in the dataset

Figure 8 represents after dropping the duplicate values.



```
df.duplicated(keep='first').sum()
```

```
1
```

```
df[df.duplicated()]
```

```
     patient's  Age  Gender      Air  Alcohol    Dust  OccuPational  Genetic  chronic  Balanced  ...  Fatigue  Weight  Shortness  Wheezing  Swallowing  C
            id                Pollution    use  Allergy      Hazards     Risk     Lung      Diet            Loss  of Breath             Difficulty  o
                                                                                Disease
789        809   39       1        4      5       6             6        5        4         6  ...        5       3          2         4           3

1 rows × 25 columns
```

```
df.drop_duplicates(keep='first', inplace=True)
```

```
df.duplicated().sum()
```

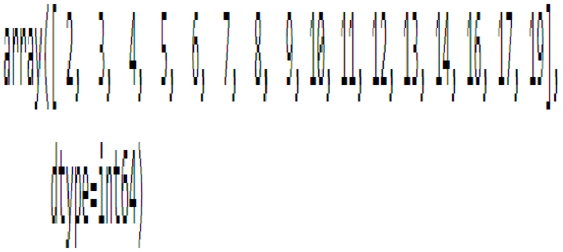
```
0
```

```
df.shape
```

```
(999, 25)
```

Figure 8: After dropping the duplicate values in the dataset

Then the dataset is divided into training and test datasets in the ratio of 65% and 35%. Now, the feature selection (Chi-square) is applied. Feature selection (Chi-square) provides the best attributes which are highly correlated between the attribute and the target attribute. In the Chi-square, when the K parameter is set to 15, it extracts the best highly correlated attributes from the lung cancer dataset. The Following figure 9 represents the selected attributes in the proposed dataset:



```
array([ 2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 16, 17, 19],
      dtype=int64)
```

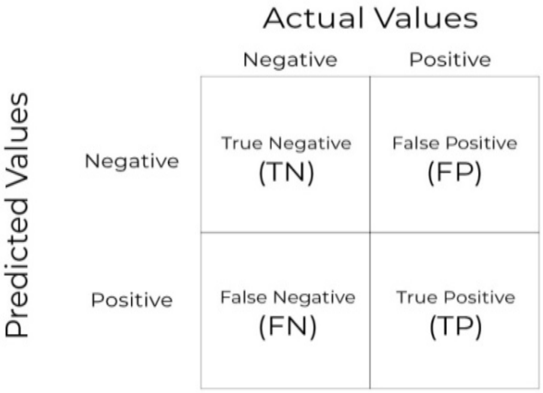
Figure 9: Selected attributes

At last, we have applied three classification algorithms, Random Forest and Support Vector Machine to classify the dataset. Then the two classification algorithms are compared with or without Chi-square based on Accuracy, Precision, Recall, and Time.

8. Performance Evaluation

Confusion Matrix is a complete justification of the classification or misclassification model. Confusion Matrix has True Positive (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

- **True Positives (TP):** When they predict yes, it means they have the disease.
- **True Negatives (TN):** When they predict no, then they don't have a disease.
- **False Positives (FP):** When they predict yes, they don't have disease in reality.
- **False Negatives (FN):** When they predict no, they have a disease.



		Actual Values	
		Negative	Positive
Predicted Values	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Figure 10: Confusion matrix

In this study, the following parameters are used to calculate the result of the algorithm by using with or without selected features:

1. **Accuracy:** It is used to test how many times a classifier is correct. Overall, it helps in calculating the performance of our correctly predicted model.

$$\mathbf{Accuracy} = \frac{\mathbf{T_{pos} + T_{neg}}}{\mathbf{T_{pos} + F_{pos} + T_{neg} + F_{neg}}}$$

2. **Recall:** It calculates the ratio of the true positive rate.

$$\mathbf{Recall} = \frac{\mathbf{T_{pos}}}{\mathbf{T_{pos} + F_{neg}}}$$

3. **Precision:** It tells us how many times correctly predicted outcomes are true.

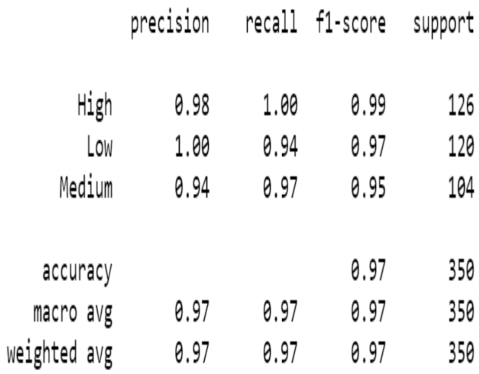
$$\mathbf{Precision} = \frac{\mathbf{T_{pos}}}{\mathbf{T_{pos} + F_{pos}}}$$

4. **F1 score:** One product that equates Precision, as well as Recall, is the F1 score. It is the average of Precision and Recall.

$$\mathbf{F1\ score} = \frac{\mathbf{2T_{pos}}}{\mathbf{2T_{pos} + F_{pos} + F_{neg}}}$$

5. **Roc Curve:** ROC Curve is a Receiver Operating Characteristic Curve. ROC Curve, generally, tells us the connection between True Positive (TP) and False Positive (FP) rate in a graphical way.

6. **Precision-Recall Curve:** Precision-Recall Curve is a graph that describes the relation between Precision and Recall.
7. **Time**: time is a parameter to evaluate the performance of different models. The Time parameter does not give accurate time, however, it is the approximate time achieved by the model.

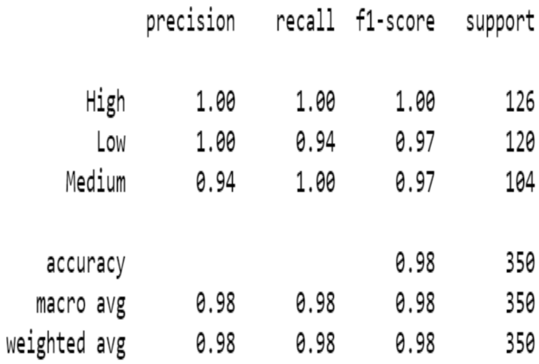


```
              precision    recall  f1-score   support

        High       0.98      1.00      0.99       126
         Low       1.00      0.94      0.97       120
      Medium       0.94      0.97      0.95       104

    accuracy                           0.97       350
   macro avg       0.97      0.97      0.97       350
weighted avg       0.97      0.97      0.97       350
```

Figure 11: Classification Report of the SVM without Chi-Square

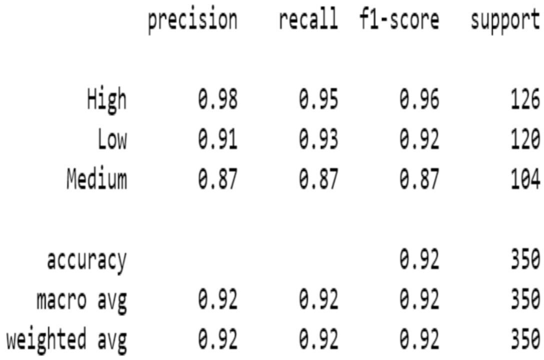


```
              precision    recall  f1-score   support

        High       1.00      1.00      1.00       126
         Low       1.00      0.94      0.97       120
      Medium       0.94      1.00      0.97       104

    accuracy                           0.98       350
   macro avg       0.98      0.98      0.98       350
weighted avg       0.98      0.98      0.98       350
```

Figure 12: Classification Report of the SVM with Chi-Square (“K is set to 15”)

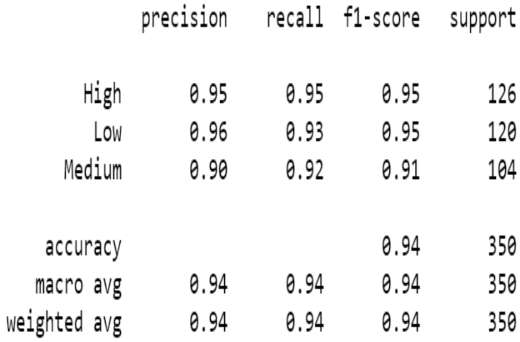


```
              precision    recall  f1-score   support

        High       0.98      0.95      0.96       126
         Low       0.91      0.93      0.92       120
      Medium       0.87      0.87      0.87       104

    accuracy                           0.92       350
   macro avg       0.92      0.92      0.92       350
weighted avg       0.92      0.92      0.92       350
```

Figure 13: Classification Report of the Random-forest without Chi-Square



```
              precision    recall  f1-score   support

        High       0.95      0.95      0.95       126
         Low       0.96      0.93      0.95       120
      Medium       0.90      0.92      0.91       104

    accuracy                           0.94       350
   macro avg       0.94      0.94      0.94       350
weighted avg       0.94      0.94      0.94       350
```

Figure 14: Classification Report of the Random-forest with Chi-Square (“K is set to 15”)

Roc Curve: ROC Curve is a Receiver Operating Characteristic Curve. ROC Curve, generally, tells us the connection between True Positive (TP) and False Positive (FP) rate in a graphical way.

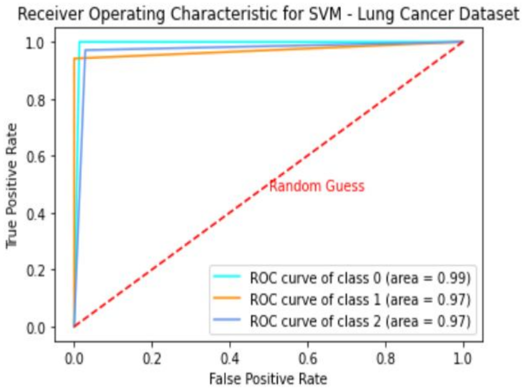


Figure 15: Roc Curve of the SVM without Chi-Square

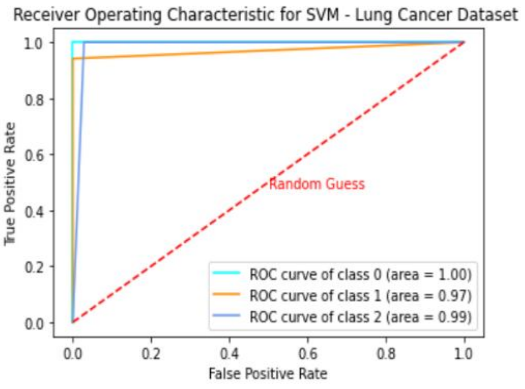


Figure 16: ROC Curve of the SVM with Chi-Square ("K is set to 15")

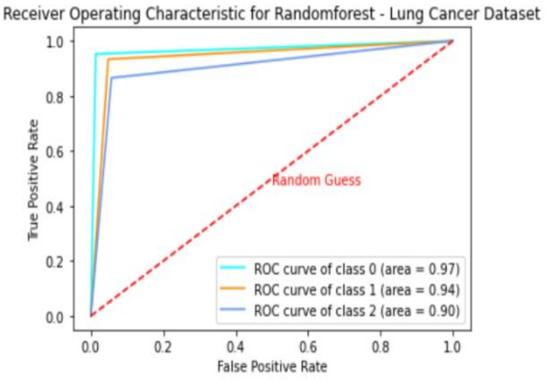
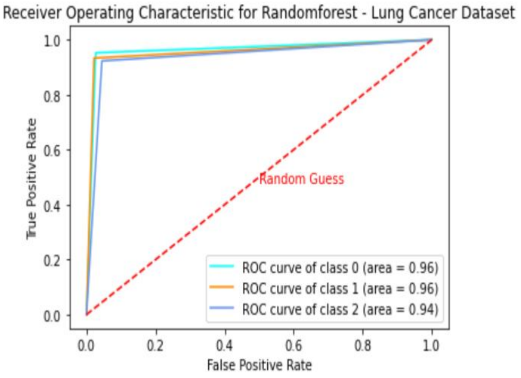


Figure 17: Roc Curve of the Random-forest without Chi-Square



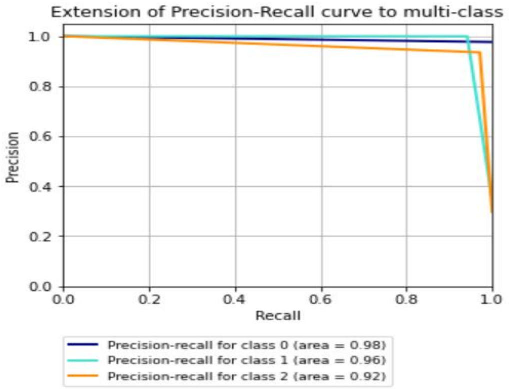


Figure 18: ROC Curve of the Random-forest with Chi-Square ("K is set to 15")

Precision-Recall Curve: Precision-Recall Curve is a graph that describes the relation between Precision and Recall.

Figure 19: Precision-Recall Curve of the SVM without Chi-Square

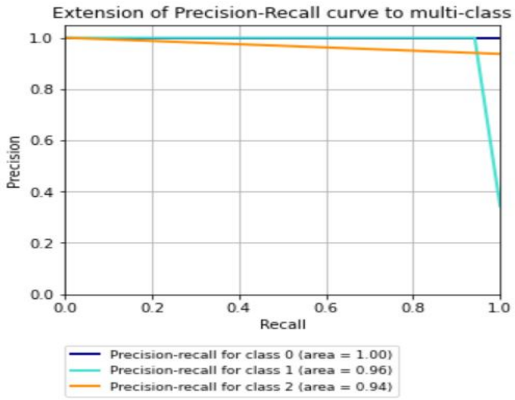


Figure 20: Precision-Recall Curve of the SVM with Chi-Square ("K is set to 15")

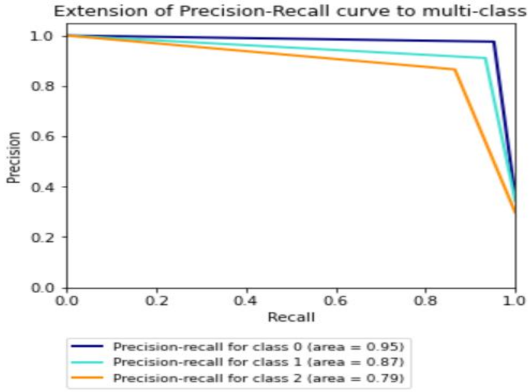


Figure 21: Precision-Recall Curve of the Random-forest without Chi-Square

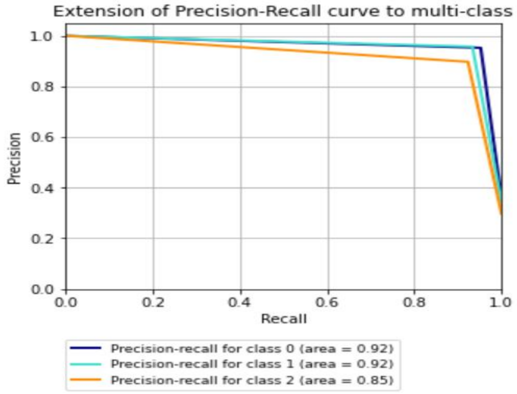


Figure 22: Precision-Recall Curve of the Random-forest with Chi-Square (“K is set to 15”)

9.COMPARISON ANALYSIS

At Present, this section provides the comparison of the SVM and Random Forest Algorithms using with and without Chi-square which is illustrated in Table 2. The following Figure 22, 23 shows that the Random Fore stand SVM have attained better results by using Chi-Square but their execution Time is different. SVM has also obtained good performance in precision, recall, and F1- score.

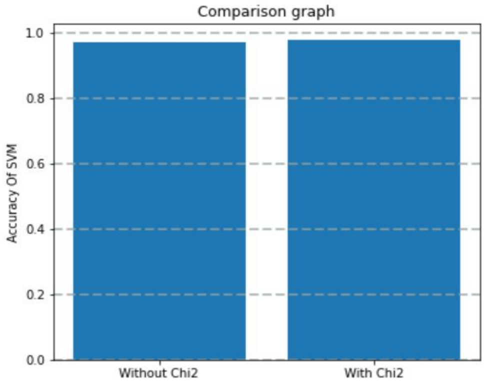


Figure 23: Comparison of SVM with and without Chi-Square

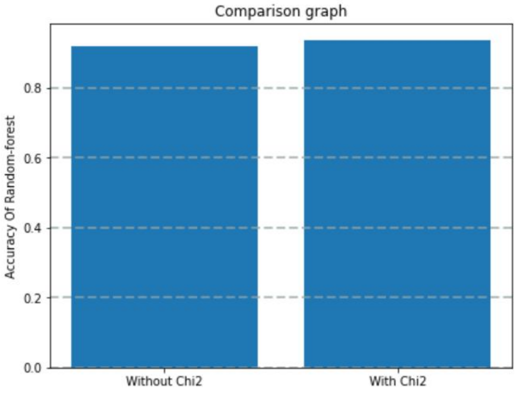


Figure 24: Comparison of Random-forest with and without Chi-Square

Table 2: Results and performance analysis

Algorithm	Accuracy	Precision	Recall	Execution Time (seconds)
SVM	97%	98%	100%	0.015 (sec)
Chi-square (K=15) first, Then SVM	**98%**	**100%**	**100%**	**0.010 (sec)**
Random-forest	92%	98%	95%	0.18(sec)
Chi-square (K=15) first, Then Random-forest	93%	95%	95%	0.14(sec)

SVM with Chi-square gives better results as compared to Random forest because its Execution time is less. The obtained accuracy of SVM is 98%, Precision (100%), Recall (100%), F1-score (100%) and Execution Time (0.010 sec).

10.CONCLUSION

While doing this study, mainly two Machine learning algorithms have been used namely Random Forest and Support Vector Machine, and then applied the Chi-Square feature selection method to reduce the irrelevant attributes for predicting Lung Cancer more efficiently. The proposed work aims to reduce the classifier execution time. The comparison of SVM and Random Forest with or without Chi-Square including K-parameter is based on Accuracy, Precision, Recall, F1 Score, and Time metrics. If the K-Parameter is set to 15, it selects the best attributes from the model.

From the results, it can be concluded that the SVM provides us better results for Lung Cancer prediction and is the appropriate algorithm that can be used for evaluating the performance of the model. The SVM has achieved the highest accuracy of 98% using Chi-Square and reduces the execution time of the model as compared to the Random Forest.

Future work may include the use of additional machine learning algorithms and can utilize another statistical test for feature selection to predict lung cancer.

REFERENCES

1. Alzubi, j. A., Bharathikannan, B., Tanwar, S., Manikandan, R., khanna, A., & Thaventhiran, C. (2019). **Boost neural network ensemble classification for lung cancer disease diagnosis.** *Elsevier*, 579-591.
2. Banerjee, N., & Das, S. (2020, July 04). **prediction lung cancer - In machine learning perspective.** *IEEE*.
3. Bharati, S., Podder, p., Mondal, R., Mahmood, A., & Al-masud, M. R. (2018). **comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer.** *Springer nature Switzerland*, 447-457.
4. Bhatia, S., Sinha, Y., & Goel, L. (2019). **Lung cancer detection: A deep learning Approch.** *springer nature singapore*.
5. Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018). **An Evalutation of Machine Learning Classifiers and ensembles for Early Stage Prediction of Lung Cancer.**
6. Gunaydin, O., Gunay, M., & Sengel, O. (2019). **Comparison of lung cancer detection Algothim.** *IEEE*.
7. Hawjins, S., Koreci, J. N., Balagurunathan, Y., GU, Y., Kumar, V., Basu, S., . . . Gillies, R. J. (2014). **Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features.** *IEEE*.
8. Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2020, November 04). **AI-based smart predidiction of clinical disease using random forest classifier and Navie Byes.** *Springer*.
9. Kancherla, K., & Mukkamala, S. (2013). **Early Lung cancer Detection using nucleus segementation based features.** *IEEE*.
10. Kuruvilla, J., & Gunavathi, K. (2014). **Lung cancer classification using neural networks for CT images.** *Elsevier Ireland*, 202-209.
11. Liu, K., & Kang, G. (2017). **Multiview convolutional nerural networks for lung nodule classification.** *IEEE*.
12. Maleki, N., Zeinali, Y., & Niaki, S. T. (2020, September 11). **A k-NN method for lung cnacer prognosis with the use of a genetic algorithm for feature seledtion.** *Elsevier*.
13. R, R. P., Nair, R. A., & G, V. (2018). **A comparative study of lung cancer detection using machine learning algorithms.** *IEEE*.
14. Roy, K., Chaudhury, S. S., Burman, M., Ganguly, A., Dutta, C., Banik, R., & Banik, S. (2019). **A Comparative study of Lung Cancer detection using supervised neural network.** *IEEE*.
15. Shakeel, P. M., Burhanuddin, M. A., & Desa, M. I. (2020, April 08). **Automatic lug cancer detection from CT image using Imprived deep neural network ensembale classififier.** *Springer*.
16. Singh, G. P., & Gupta, P. K. (2018, may 05). **Performance analysis of various machine lerning-based approaches for detection and classification of lung cancer in humans.** *Springer*.
17. Zieba, M., Tomczak, J. M., Lubicz, M., & Swiatek, J. (2013). **Boosted SVM for exteacting rules from imblanced data in application to prediction of the post-operative life expectancy in the lung cancer patients.** *Elsevier*.
18. https://data.world/cancerdatahp/lung-cancer-data.
19. Miah, M. B. A., & Yousuf, M. A. (2015, May). **Detection of lung cancer from CT image using image processing and neural network.** In *2015 International conference on electrical engineering and information communication technology (ICEEICT)* (pp. 1-6). ieee.
20. Alam, J., Alam, S., &Hossan, A. (2018, February). **Multi-stage lung cancer detection and prediction using multi-class svm classifier.** In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.
21. Kancherla, K., &Mukkamala, S. (2012, April). **Feature selection for lung cancer detection using SVM based recursive feature elimination method.** In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* (pp. 168-176). Springer, Berlin, Heidelberg.
22. Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020, March). **Lung Cancer Prediction using Machine Learning: A Comprehensive Approach.** In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 108-115). IEEE.
23. Vikul J. Pawar, D.Kharat, Suraj R.Pardeshi, Prashant D. Pathak.(2020 August). **Lung Cancer Detection Using Image Processing and Machine Learning Techniques.** IJATCSE.