Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)
IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1

Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques

Chinmayi Thallam, Aarsha Peruboyina, Sagi Sai Tejasvi Raju, Nalini Sampath
Department of Computer Science and Engineering
Amrita School of Engineering, Bengaluru
Amrita Vishwa Vidyapeetham, India
Email: chinmayitallam@gmail.com, aarsha256@gmail.com, saitejasvi99@gmail.com, s_nalini@blr.amrita.edu

Abstract— **Lung cancer is one of the most common and serious diseases present around the world, which is observed in people of all age groups ranging from children to old people. Annually it costs a lot of money for the cure and diagnosis of people with lung cancer. The existing clinical techniques such as X-Ray and other imaging procedures require complex hardware and considerable expense. Thus the most important issue is the prediction to be accurate and to use a reliable method for that. This raises the need for (comparatively more effective and cheaper) machine learning models in medical diagnosis using medical data sets. Long-term tobacco smoking results in 85 percent of cases of lung cancer. About 10–15 percent of cases arise in people who never smoked. There are numerous methods and tools that are available now for data analysis and its computation. These technological advancements will be referred and used to develop prediction models in the project to predict the presence of lung cancer at an early stage in a patient.**
The study involves comparing various classification and ensemble models such as Support Vector Machine(SVM), K-Nearest Neighbour (KNN), Random Forest(RF), Artificial Neural Networks (ANN) and a hybrid model, Voting classifier. The performance of the various models is compared and evaluated in terms of their accuracy. Thus, it is easy to identify a patient with lung cancer at an early stage using various sophisticated technologies of today.

Keywords— Machine Learning, Support Vector Machine, Voting, Random Forest, Cancer, K-Nearest Neighbour, Neural Networks

I. INTRODUCTION

Lung cancer is mainly triggered by cigarette smoke. Smoke that penetrates into the lungs causes damage to the lung tissue. In nonsmokers, lung cancer may be induced by radon radiation, second hand smoking, air contamination or other causes. Heredity is another source of lung cancer, as well. While lung cancer (malignant growth) is hard to diagnose and cure, it may be avoided or treated in the early stages. Lung malignancy is one among the dangerous cancer forms and is commonly found. Currently an approximate of 2.09 million cases observed against lung cancer. Also depending on the stages of the cancer they assign grades to the cancer [1].

Lung cancer has deeply impacted almost every family and everyone's life and there is a rapid increase of population with lung malignancy. Also, it is caused due to major problems like jaundice, lymph node swelling and problems with the nervous system. A patient suffering with lung cancer has go through various complexities during his/her diagnosis of a disease. Thus an automation in this regard may speed up the process and can assist the pathologist [2]. According to the WHO, the number of people with lung cancer have risen from 1.8 million in 2012 to 2.09 million in 2018. From the historic records it has been observed that lung malignant caused 1.6 million deaths to 1.76 million deaths between the years 2012 and 2018. As of 2018, the World Health Organization has recorded cancer as the subsequent driving reason for the deaths across the world, of which lung malignancy is found to be the most analyzed and diagnosed disease. So, increasing the awareness and predicting the onset of lung cancer in early stages can help people to take necessary precautions and thus decreasing the death caused by lung cancer.

Lung Malignancy is of two forms. Small cell lung cancer (SCLC) generally occurs in the inner layer of the bronchial walls. It is less seen but is deadly serious and can result in loss of life. The development of this cancer spreads rapidly into surrounding tissues or other sections of the body. Symptoms continue not to be detected until the malignancy expands to various parts in the body. NonSmall cell lung cancer (NSCLC), progresses to other areas of the body less rapidly than SCLC. It is commonly seen in adults, mostly in elderly people. NSCLC is not as dangerous as SCLC in early stages. Apart from heredity, there are many other factors that add to the cause of lung cancer. Today's lifestyle is a major contributing factor to the increase of lung cancer patients.

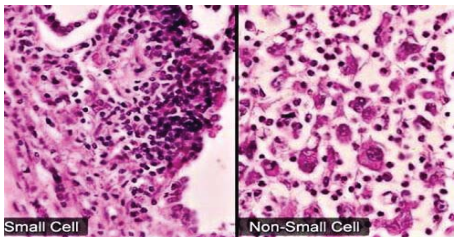

Fig. 1. Small cell and NonSmall cell lung cancer

The catalog classification is an essential part for operative electronic business applications and classical machine learning problems [3]. To classify whether a given set of features pertain to a person with lung cancer or not, the technology like machine learning is used . Machine learning is used in classifying, predicting and even in clustering of data. It is basically the training of a model which is then used to perform some operation. In technology like machine learning, the model learns something new when trained, and executes the learnt experience on a test data. Learning can be supervised, unsupervised or reinforcement. Supervised learning is the type where a data set with a specific class label/ dependent variable is used to train the model whereas unsupervised learning doesn't require any output variable in its data. Unsupervised learning discovers information by itself. Reinforcement learning is learning using experience and rewards and punishment. This type of learning doesn't require a dataset. Various ML algorithms are - SVM, Decision tree, Random Forest, K Means Clustering, Regression, K Nearest Neighbor and the rest.

The support vector machine model plays an important role in obtaining high accuracies for the small datasets like the one considered in the study and is effective when it comes to high dimensional spaces. Likewise a random forest algorithm also has a great contribution in the field of machine learning due to its diverse and simple nature. Because of the concept of ensemble study it easily solves a complex problem by combining many classifiers in short time thus giving high accuracies irrespective of the size of the dataset.

The other models like K-NN are simple and robust to train noisy data. As the size of the dataset increases the model becomes more and more effective giving best results. Similarly, voting algorithms give best results for classifications by training and evaluating the models in parallel. Hard voting considers votes of various models and chooses the class with high votes. Also, due to the parallel processing ability, artificial neural networks have their own significance. They are fault tolerant with distributed memory which make the machine learn in order to give accurate results with high accuracy.

Machine Learning deals with different statistical models that are used to implement a particular work, without the use of extra instructions depending on patterns and inferences. Machine Learning is the ability of a computer to learn from mined datasets. These techniques come under artificial intelligence. Classification is the process which is implemented to predict the label of the class from observed values. The output obtained will have the form such as high or low or yes or no.

II. LITERATURE SURVEY

Sangita Khare et al. [4] view says that an infection is an irregular condition that influences functioning of the body caused due to the outside and interior dysfunctions. The study investigates various data mining techniques that tells us about how a chronic disease differs from the other chronic diseases with the help of ICD9 demonstrative codes. They mainly focused on the people having heart disease and diabetes and drafted an ideal set of ICD9 demonstrative codes which are researched dependent on the human anatomic frameworks. However ICD classification is not suitable when there is no or very few data available about the patient.The conclusive diagnosis using ICD9 cannot be achieved in the patient's first visit and it takes several visits to examine the patient.

S.Sivakumar et al. [5] proposed that identification of the disease at a primitive level is the most mandatory prerequisite in administering the required care to the patients. The dataset used in this study is from The Lung Image Database Consortium (LIDC-IDRI) comprises demonstrative and lung disease screening thoracic CT filters with increased clarified sores. Proposed technique involves all the computed tomography pictures of the lungs that have a DICOM configuration of 512x512. Subsequent to upgrading the CT scan picture, the Weighted Fuzzy-Possibilistic C-Means (WFPCM) calculation is put in to fragment the picture. The image dataset is pre-processed using Fuzzy C-Means Clustering and Fuzzy-Possibilistic C-Means.The CT scan pictures are then segmented using the WFPCM algorithm from which the obtained result leads to feature extraction. The last stage of the proposed system is to use the obtained features from feature extraction as the inputs for the support vector machine classifier. The segmented data validated against the various components shows that WFPCM gives good partition results when compared with FPCM and FCM. The final outcome shows that the RBF kernel type has obtained 80.36% of accuracy, 82.05% of sensitivity, 76.47% of specificity which stands the best fromremaining.

Neesha Jothi et al. [6] discussed that data mining is the way towards perceiving and extracting designs in presence of huge amounts of data, using machine learning concepts, statistics and database systems. The discussed data mining techniques are machine learning, artificial intelligence, statistics, probability including various other models. Few of the models discussed are Anomaly Detection, Clustering, Classification, K-Nearest Neighbors, Swarm Intelligence (SI), Logistic Regression, Bayesian Classifier, SVM. Data mining has played a significant role in the healthcare sector, particularly in the prediction of various types of diseases. It comes to a conclusion that there is no one particular model in data mining to use for finding the best accuracy, which is the most important thing in medical diagnosis. So, to get better accuracy, designing a hybrid model is to be considered.

S. Senthil et al. [7] view says that lung malignancy is a condition that arises fromthe proliferation of abnormal tissues in the lung and it is necessary to predict and detect lung cancer before time by using optimal features of the neural network. Initially the lung database is collected and given to the system

as input. Data preprocessing is then added to the photos to improve the picture in order to produce the high contrast pictures. Particle Swarm Optimization is executed to get the attributes of the pictures given as input then the neural network classifier is used to classify certain characteristics of input images that are categorized as cancerous or noncancerous. The classification outcome of the checked data is weighed to verify the error rate or frequency error that happens during the classification process, and the error is fixed by adjusting the weights in the dataset. The extraction process of the app is carried out with the implementation of PSO. The extraction of the function is the part of the pattern recognition techniques that is conducted on the input data to obtain the appropriate features that are to be more descriptive, nonredundant and collect the cancer details to determine the patient conditions for interpretations.

Jennifer P. Cabrera [8] research aims to assess the quality articulation information from oligonucleotide microarrays to evaluate whether a patient has lung malignant growth, and use Support Vector Machines (SVM) to recognize the kind of lung malignancy present. A microarray dataset of oligonucleotides alluding to 12,600 sequences of copies in an aggregate of 203 examples of lung tumor gene information is utilized. The suggested framework takes microarray data referring to the rates of mRNA expression in the human lung tumor specimen as input oligonucleotides. The machine reads the input data collection file and gathers data for analysis from the microarray. The device picks 1000 genes and 100 marker genes by choice, respectively. In addition , the user can upload an optional gene definition file to provide more detail about the resulting genes. The tab dataset, the tab attributes, the tab chosen genes, the tab marker genes, the tab predictions, and the tab output emerge one by one as the device produces performance. The performance of this method can be used in the study and detection of genes which is most applicable to a particular disease class.

S. Sasikala et al. [9] view tells us that neural networks play a significant role in recognizing the cancer cells among normal human tissues in this paper. They used the convolutional neural network to forecast lung cancer, based on CT images of the cancered lungs. However it's computational power varies depending on the data size and network architecture mostly having complex architecture. The dataset is taken from the Lung Image Database Consortium Image Database Resource Initiative (LIDC-IDRI) which is available in Digital Imaging and Communications in Medicine (DICOM) format. The preprocessing is done using the median filter method which separates the tumor cells from the normal cells based on the sliced segments from the CT scans. The collective sum of the produced product of the training input weights are passed on to the activation function of the nodes. From the result obtained, every level is sent as the input to the next simultaneous film or the layer. The backpropagation algorithm divides the training into two phases. The first phase involves feature extraction and the second phase deals with the classifier containing different threshold layers, following a SoftMax layer which performs high-level reasoning for the neural network. An image sample has to be sent as an input to the model that is trained and it will detect the presence of cancer and its location in the sample input image. If a malignant cell is present it will be displayed with a message along with the input image. The accuracy using CNN for lung cancer detection is 96%.

Syed Saba Raoof et al. [10] described a comprehensive approach to deal with lung malignancy prediction with the help of different ML algorithms. They have discussed mainly the concept of deep learning(DL) in the medical field. The proposed paper summarizes the performance of various algorithms like Convolutional Neural Network (CNN), Deep Belief Network (DBN), Auto-Encoders (AE's), Fully Convolutional Networks (FCN) . These algorithms are bang on techniques to study and analyze imaging in health care like CT images, X-Ray and MRI images. The approaches of DL are quite complex with a high number of computations, opaqueness, technical challenges but stand useful for classification and detection of bruises and various images, segmentation of wounds and organs, generation of images and its enhancements. Thus the study says all the techniques of deep learning in the field of medicine are emphasised.

Janee Alam et al. [11] proposed a detailed view on detection of lung cancer, its diagnosis and also tells if a patient can get lung cancer or not using SVM. The work is processed using MATLAB software using various image processing methods and techniques implemented. The proposed technique distinguished 126 pictures as contaminated in a total of 130 and anticipated 87 pictures as dangerous out of 100 some time ago. The trial examination shows accuracy of 97% in identification and accuracy of 87%. The various image processing techniques like segmentation, detection, enhancement of various images along with feature extraction which is done using the Gray Level Co-occurrence method are implemented in detail. However the image processing techniques are time consuming, costly and it requires a qualified professional.

Shingo Kakeda et al. [12] have explained about detecting lung nodules using an automated method and CAD system that incorporates EpiSight/XR software and also an image server. They have considered lateral chest radiographs and conventional posteroanterior chest radiographs in their study. The entire work takes place in four basic steps. Firstly, the complex structures are reduced to various images which are then with the help of Multiple Gray Level thresholding techniques, the patients with nodules were found. Subsequently, to separate the genuine knob with bogus positive knob features are information and distinction pictures are utilized to extricate features. Using a rule based examination along with the help of ANN the extricated features are used for false positive knobs present in a patient.The framework takes a lot of time to train the model and to find the exact location of the object on image. As a result it was found that on the whole, out of 315 false positive identifications 75% of them were identified to have normal anatomic design and it is found that the 20% of 315 units do not have any connection with the anatomic design and they were all distributed in 6 different parts inside the lungs.

Wasudeo Rahane et al. [13] proposed a method which involves classification of lung malignancy in blood samples and CT scan images using ML and image processing. They have worked using java which is an open source programming language.The framework of the system involves image acquisition where the input given by the user is taken into consideration followed by gray scale conversion, that is, it converts RGB image into gray. Subsequently, using the median filter noise reduction in the images is carried out and then the process of binarization takes place which involves conversion of gray image into binary image with white and black pixels. Lastly with the help of SVM model lung disease presence is predicted.They have developed a web based application using JSP for framework development. They have used AWS to store the image datasets and used MySQL as their database to store the pictures of various blood samples along with CT scan images.

Rohit Y. Bhalerao et al. [14] proposed a fusion method of combining CNN and image processing techniques in order to identify lung disease in a person. The inputs used for the purpose of image processing are the Ct scan images taken from LIDC. Firstly the information is preprocessed, converts RGB pictures to grayscale pictures so as to avoid the complexity in further estimations and then it is transformed into a binary image in order to get refined and clear input pictures. These clear and efficient images are sent as the input for CNN which processes the images through the various layers present in it in order to train the data using the methods like Max Pooling filtering and Convolution Filtering.The work is carried out using MATLAB which is known for its smooth and good performance. The output of the proposed work tells if the lung image is malignant or benign.

Divya Chauhan et al. [15] discussed the role of data mining in identifying lung malignancy in a patient using linear discriminant and particle component analysis. The work is executed in MATLAB software to build a user-friendly version. Initially they collect all the information(images) for the training process and then they are detected to find if there are any edges in the uploaded category of images. Followed by applying PCA and performing feature selection to load a feature vector that is stored in the database for cancerous and noncancerous. The work then performs testing by stacking up the test data and then the classification of the images takes place with the help of LDA. They have also concluded that the proposed technique is way better than ICA and SURF, the traditional methods.

III. PROPOSED WORK

The procedure includes an itemized study of different machine learning algorithms and techniques. It gives us a thought regarding how to choose a specific ML model dependent on the performance estimates like accuracy so as to meet the important prerequisites.

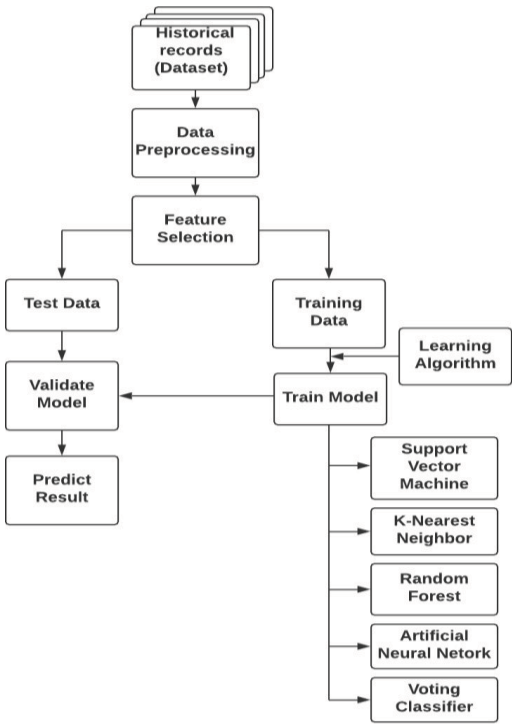

Fig. 2. Workflow of the proposed work

The system architecture for the whole experiment characterizes the different levels associated with the ML cycle. It includes the significant steps involved in transformation of raw data into training data which is capable of decision making in the system. When the information is isolated into test data & training data, a learning calculation is applied on to the training dataset. The model is now trained using various ML algorithms.The result acquired is utilized to approve the model with the assistance of the unused test information and thus the necessary outcome is anticipated.

Data pre-processing

The data is encoded or transformed into a particular form which is then made feasible for the further analysis of the machine. It allows the different algorithms to interpret the features of the data easily. The process involves the following steps:

A. Importing Required Libraries

Firstly, the fundamental and required libraries and modules are to be brought into the workspace.These libraries and modules like numpy, warnings, os, matplotlib, pandas and more improve the activity with an easy work process. It helps in plotting various pictorial representations, makes the activity simpler with complex numerical estimations, helps in creating data frames, gives a smooth interface between the client and the framework, and helps in taking care of the alerts as well.

B. Importing Dataset

Data collection plays an important role in the process. Thus the dataset collected should be audited and analysed through their features and records. The study involves the text data

collected from data.world which has got 25 features out of which the 'Level', an independent feature plays a significant role in deciding the malignant growth phase of a patient. Each instance is a combination of several features like obesity,chest pain,smoking,dry cough,genetic risk and many more which are taken on the scale of 10, along with other features like age and patient id.

C. Handling Missing Data

On the off chance if there are any invalid or missing quantities in the data it is essential to redress them for future estimations. This should be possible by either taking care of them with the mean, mode, median estimations of the data or by considering separate algorithms or by deleting the rows if they are less in number or by assigning them with an alternate class. The data gathered is seen as liberated from the missing qualities.

D. Encoding Categorical Data

The categorical values of the independent feature, 'Level' are converted from categorical to numerical. This is done using integer encoding strategy-mapping categorical values to integer values. This conversion helps to have smooth calculations throughout the process of the study.

E. Feature Selection

The process involves extraction of the required features and dismissing the rest so as to stay away from superfluous complexities for the future computations. The utilized method is the method of correlation matrix in order to select the features which contribute to the output feature to increase the accuracy and train faster. The obtained results in the correlation framework should extend from -1 to 1 however not 0. In this way it is easier to determine the necessary features by their relationship esteems with one another feature.

F. Splitting of Data

The gathered data is now important to be splitted into training and test data in the necessary proportions. The test data is used to train and prepare the model with different algorithms and then is sent for validation. The validation process takes place with the unused test data set to acquire better accuracies of the algorithms used.

Algorithms

A. Support Vector Machine

It is a supervised learning model that can be used for both classification and regression analysis. SVM helps in reducing the rate of misclassification and thus gives good results. It plays out a few emphasess to locate an ideal hyperplane that best isolates the various classes of a dataset in a multidimensional space. The hyperplane is framed by the support vectors that are nearest to the edge(margin). The one which is framed with the greatest edge is supposed to be the ideal hyperplane, that is, maximum marginal hyperplane(MMH). The output is then anticipated by plotting the test set data points on the obtained hyperplane. The workflow of the algorithm involves an SVC classifier to train the model and the model is fitted to the training set and prediction is done using the test set.

B. Random Forest

It is a model that follows a supervised learning mechanism, also an ensemble model that tends to be utilized for both classification and regression analysis. The model is also known for the calculations which involve the construction of decision trees. The forest is said to be more vigorous when it has many trees in it. It is also a strong learner that creates N number of decision trees. Every decision tree obtained is built from the subset of training set and features that are taken randomly. The output class is determined by the aggregating of the votes of N decision trees created. The workflow of the algorithm involves a Random Forest Classifier to train the model and also the creation of N decision trees whose votes are aggregated and the model is fitted to the training set and prediction is done using the test set.

C. K- Nearest Neighbor (KNN)

The model includes discovering k closest neighbors of a specific component vector out of the absolute N training vectors accessible. It calculates the difference between the observed and actual data points using a similarity measure. The similarity measure used is the euclidean distance, that is, the metric used is 'minkowski' with the estimation of p set to 2. After computing the similarity measure between the new case and the actual data, it classifies a new case by majority voting of the class labels of the first k closest neighbors, that is, the first k similarity values computed. The workflow of the algorithm involves a K Neighbors Classifier to train the model and the model is fitted to the training set and prediction is done using the test set. For a good practice the estimation of k should not be a multiple of the number of classes. It is likewise critical to take note of that the estimation of k ought to be odd while considering a 2 class issue to stay away from a tie between the classes.

D. Neural Networks

The working of the model includes bringing in fundamental libraries like keras (a library for neural systems) and tensorflow is utilized as the backend which makes neural systems simpler and quicker.. Artificial Neural Networks(ANN's) are created to behave as if they are interconnected brain cells, by programming regular computers. It has many layers arranged in series with each layer having many artificial neurons called units. The input unit is fed with various forms of data and propagates it to one hidden layer or more than one hidden layer where it learns about the result and produces an outcome to the output layer. The activation functions used for input and hidden layers are 'relu' and the activation function for the output layer is 'softmax'. The difference between the actual and desired is calculated and the error is adjusted using backpropagation until the error is possibly minimum. The workflow of the model involves choosing inputs and outputs that define the structure followed by initializing the weights using the 'adam' optimizer and biases randomly which leads to feeding the network to the training set. The subsequent step involves running the model and proceeding with the error correction. If the stop criteria results in a no then cascade down the error

rates to hidden nodes later, recalibrate the weights between hidden nodes and the input node. However if the stop criteria results is a yes then the focus shifts to the test data where the model is anticipated which then estimates the output and evaluates the result.

E. Voting Classifier

Rather than training separate models and predicting the accuracies, a voting classifier combines different models and predicts an output class using hard voting rather than training separate models and predicting the accuracies, a voting classifier. Hard voting takes the majority voting of the predicted class labels of each individual model used in the voting classifier. The models used in training the voting classifier are Support Vector Machine, K Nearest Neighbor and Random Forest for obtaining high accuracy using the combination of algorithms. This model is a hybrid model of the three different algorithms used.

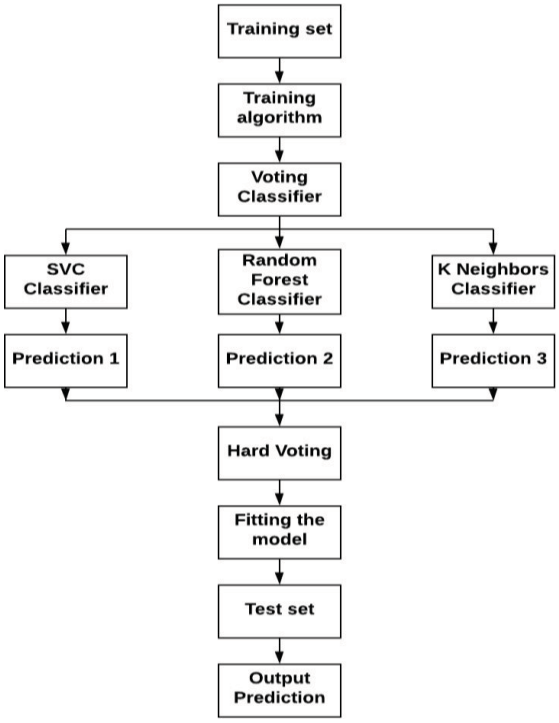

Fig. 3. Workflow of the voting classifier

IV. RESULTS AND DISCUSSION

A. Support Vector Machine

It gives an output of 95% accuracy with 0.8 training data and 0.2 test data. The plotted confusion matrix shows the model's performance in determining the final result. The outcome shows the presence of 10 wrong predictions out of 200 test records.

```
In [19]: y_pred = svm.predict(x_test)

In [20]: acc_svm = accuracy_score(y_true=y_test, y_pred= y_pred)
         print("Overall accuracy of SVM model using test-set is : %f" %(acc_svm*100))

         Overall accuracy of SVM model using test-set is : 95.000000
```

Fig. 4. Accuracy of the Support Vector Machine model

```
cm = confusion_matrix(y_true=y_test,y_pred=y_pred)
plot_confusion_matrix(cm, level,title='Confusion matrix
```

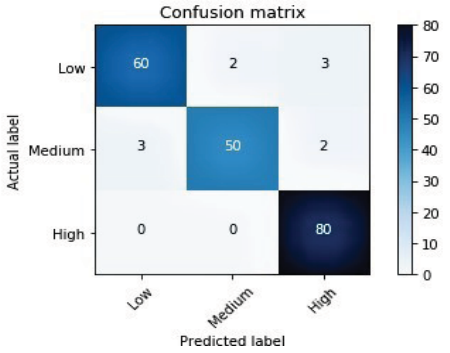

Fig. 5. Confusion matrix for the Support Vector Model

B. Random Forest

It gives an output of 97.5% accuracy with 0.8 training data and 0.2 test data. The plotted confusion matrix shows the model's performance in determining the final result. The outcome shows the presence of 4 wrong predictions out of 200 test records.

```
y_pred = forest.predict(x_test)

acc = accuracy_score(y_true=y_test, y_pred= y_pred)
print("Overall accuracy of random forest model using test-set is : %f" %(acc*100))

Overall accuracy of random forest model using test-set is : 97.500000
```

Fig. 6. Accuracy of the Random Forest model

```
cm = confusion_matrix(y_true=y_test,y_pred=y_pred)
plot_confusion_matrix(cm, level,title='Confusion matrix')
```

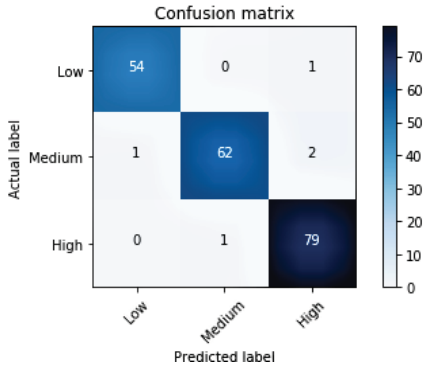

Fig. 7. Confusion matrix for Random Forest model

C. K-Nearest Neighbor

It gives an output of 97% accuracy with 0.8 training data and 0.2 test data. The plotted confusion matrix shows the model's

performance in determining the final result. The outcome shows the presence of 6 wrong predictions out of 200 test records.

```
y_pred = knn.predict(x_test)
```

```
acc = accuracy_score(y_true=y_test, y_pred= y_pred)
print("Overall accuracy of KNN model using test-set is : %f" %(acc*100))
```

```
Overall accuracy of KNN model using test-set is : 97.000000
```

Fig. 8. Accuracy of K-Nearest Neighbor model

```
cm = confusion_matrix(y_true=y_test,y_pred=y_pred)
plot_confusion_matrix(cm, level,title='Confusion matri
```

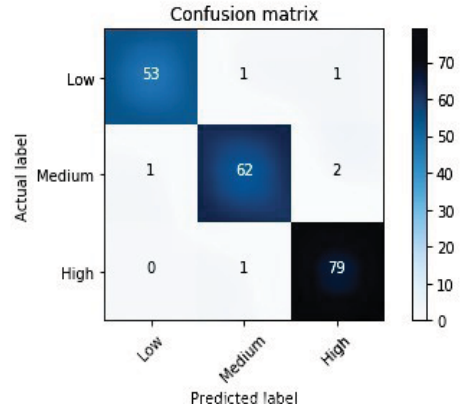

Fig. 9. Confusion matrix for the K-Nearest Neighbor model

D. Neural Networks

The model gives an output of 95.99% after completion of 18 epochs. The prototype involves one input layer , one output layer along with four hidden layers with 3 units. The plotted confusion matrix shows the model‘s performance in determining the final result. The outcome shows the presence of 8 wrong predictions out of 200 test records.

```
In [18]: score = model.evaluate(x= x_test, y= y_cat_test, batch_size=32)
         acc = score[1]
         err = 1 - acc
         print("Loss Value : ", score[0])
         print("Accuracy : ", score[1]*100)

         200/200 [==============================] - 0s 65us/step
         Loss Value :  0.22773556519299745
         Accuracy :  95.99999785423279
```

Fig. 10. Accuracy of Artificial Neural Network model

```
cm = confusion_matrix(y_true=y_true, y_pred=y_predict)
plot_confusion_matrix(cm, level, title='Confusion Matrix')
```

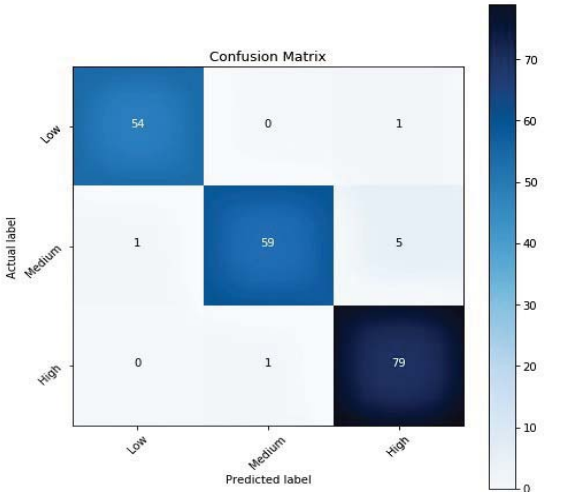

Fig. 11. Confusion matrix for Artificial Neural Network model

E. Voting Classifier

The models used to build voting classifiers are Support Vector Machine, K-Nearest Neighbors, and Random Forest classifiers. The model gives an output of 99.5% accuracy. The plotted confusion matrix shows the model‘s performance in determining the final result. The outcome shows the presence of only 1 wrong prediction out of 200 test records.

```
In [13]: y_pred = clf.predict(x_test)

In [14]: acc_vc = accuracy_score(y_true=y_test, y_pred= y_pred)
         print("Overall accuracy of Voting model  is : %f" %(acc_vc*100))

         Overall accuracy of Voting model  is : 99.500000
```

Fig. 12. Accuracy of the Voting Classifier

```
cm = confusion_matrix(y_true=y_test,y_pred=y_pred)
plot_confusion_matrix(cm, level,title='Confusion matrix')
```

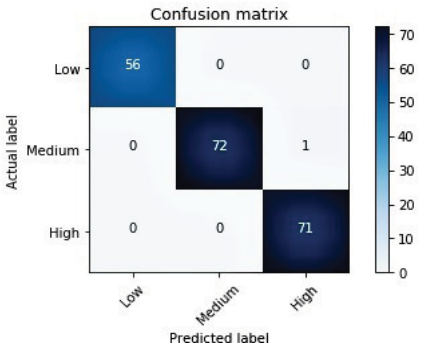

Fig. 13. Confusion matrix for Voting Classifier

TABLE I. PERFORMANCE OF DIFFERENT CLASSIFIERS

Model	Accuracy(%)
Support Vector Machine	95
Random Forest	97.5
K-Nearest Neighbor	97
Artificial Neural Network	95.99
Voting Classifier	99.5

V. CONCLUSION AND FUTURE WORK

The experiment is carried to predict the best fitting model for lung cancer patient data set which gives the greatest accuracy. The aim of conducting the experiment is to predict the early stage lung cancer in a person utilizing the various machine learning classification algorithms. During the course of the project the research on various papers helped in learning various aspects of the project. The study helped in understanding that SVM is not suitable for data having more noise along with large data sets. Also it is understood that with a large number of trees, the random forest algorithm becomes slow for real time prophecy and it acts as a predictive model like a black box. The study gives an idea about how K-NN works well only with balanced data and is sensitive to outliers. Furthermore, the working of artificial neural networks with numerical data makes it difficult to show the problem to the network and the network duration is unknown. This also helped in the consideration of techniques and features based on various factors. The experiment helped in learning various machine learning algorithms, its implementation, uses, pros and cons. It also gave a complete insight on preprocessing techniques and numerous activation functions used in neural networks. Building various models, training and testing them tells that the voting classifier is the best suitable model on this dataset for the early stage prediction of lung cancer in a patient. There is a scope of enhancement in the experiment by using various other models like Logistic Regression, Extra trees classifier and boosting methods. The accuracy might vary depending on the volume of the data collected by applying various other preprocessing techniques and tools which can be explored further.

REFERENCES

[1] T. Babu, T. Singh, D. Gupta and S. Hameed, "Colon Cancer Detection in Biopsy Images for Indian Population at Different Magnification Factors Using Texture Features," 2017 Ninth International Conference on Advanced Computing (ICoAC), Chennai, 2017, pp. 192-197, doi: 10.1109/ICoAC.2017.8441173.

[2] T. Babu, D. Gupta, T. Singh, S. Hameed, R. Nayar and R. Veena, "Cancer Screening On Indian Colon Biopsy Images Using Texture and Morphological Features," 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 0175-0181, doi: 10.1109/ICCSP.2018.8524492.

[3] Pandey S., M S., Shrivastava A. (2018) Data Classification Using Machine Learning Approach. In: Thampi S., Mitra S., Mukhopadhyay J., Li KC., James A., Berretti S. (eds) Intelligent Systems Technologies and Applications. ISTA 2017. Advances in Intelligent Systems and Computing, vol 683. Springer, Cham.

[4] Va Dominic, Dr. Deepa Gupta, Sangita Khare, and Aggarwal, Ab, "Investigation of chronic disease correlation using data mining techniques", in 2015 2nd International Conference on Recent Advances in Engineering and Computational Sciences, RAECS 2015, 2015.

[5] S.Sivakumar, Dr.C.Chandrasekar. Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines, International Journal of Engineering and Technology (IJET)

[6] Neesha Jothi, Nur Aini Abdul Rashid, Wahidah Husain, Data Mining in Healthcare - A Review, Procedia Computer Science 72:306-313 · December 2015

[7] Dr. S. Senthil, B. Ayshwarya, Lung Cancer Prediction using Feed Forward Back Propagation Neural Networks with Optimal Features, International Journal of Applied Engineering Research, 13(1), pp.318-325.

[8] Jennifer P. Cabrera , Lung Cancer Classification Tool using MicroarrayData and Support Vector machines, In 2015 6th International Conference on Information, Intelligence,
Systems and Applications (IISA) (pp. 1-6). IEEE

[9] S. Sasikala, M. Bharathi, B. R. Sowmiya, Lung Cancer Detection and Classification Using Deep CNN, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-2S December 2018

[10] Syed Saba Raoof , M.A.Jabbar, Syed Aley Fathima, Lung Cancer Prediction using Machine Learning: A comprehensive Approach

[11] Janee Alam, S., & Hossan, A. Multi-Stage LungCancer Detection and Prediction Using Multi-class SVM Classifier. 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)

[12] Kakeda, S., Moriya, J., Sato, H., Aoki, T., Watanabe, H., Nakata, Doi, K. 2004. Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer-Aided Diagnosis System. American Journal of Roentgen logy, 182(2), 505– 510. doi:10.2214/ajr.182.2.1820505.

[13] Wasudeo Rahane, Yamini Magar, Himali Dalvi ,Anjali Kalane ,Satyajeet Jondhale ,Lung Cancer Detection Using Image Processing and Machine Learning HealthCare, Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India

[14] Rohit Y. Bhalerao, Rachana K. Gaitonde, Harsh P. Jani ,Vinit Raut, A novel approach for detection of Lung Cancer using Digital Image Processing and Convolution Neural Networks, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)

[15] Divya Chauhan, Varun Jaiswal , An Efficient Data Mining Classification Approach for Detecting Lung Cancer Disease.