Tech-Review of Lung Cancer Prediction

Raghumanda Kavya Sree[1], Sai Prasad Ravi[2], Raveena Babu[3]
[1,2,3]*Lovely Professional University*

Abstract **- Lung cancer is increasing at a rapid rate and due to the present pollution, it is increasing in non-smokers also. We really cannot predict anymore who will become the next victim of this deadly disease. But we can always take help from the technology. Early diagnosis of lung cancer can be very helpful in saving many lives, so the techniques to predict and detect early-stage cancer are increasing rapidly. Using generic lung cancer symptoms can help in predicting the possibility of getting lung cancer. In this study we list out few of the techniques involved in predicting lung cancer. Techniques such as Data Mining, Machine Learning, Automatic lung cancer prediction from X-ray images using Deep learning, are few of the technologies we will list out in this paper. In this paper we can see the use of classification-based data mining techniques such as Naïve Bayes, Bayesian Network and Artificial Neural Network for the prediction of lung cancer. We also state how Machine Learning classifier techniques such as, Artificial Neural Network (ANN) and Support Vector Machine (SVM), are used to determine the data sets as cancerous and non-cancerous.**

Index Terms **- Data mining, lung cancer prediction, Naïve Bayes, Deep Learning, machine Learning, ANN.**

I.INTRODUCTION

In this paper we discuss the various ways in which we can use technologies to predict lung cancer. One of the techniques is using Data Mining. Data Mining is basically analysing raw data to get useful information. Data Mining has been proven to be significantly useful in the healthcare field. We can use the symptoms as the data set and determine severity of the disease on a particular person according to the symptoms. WEKA tool is also used for predicting the risk level of the lung cancer [1]. WEKA is a collection of machine learning algorithms for solving data-mining problems. Data Mining tools are used for prediction. Data Mining techniques can be assorted and be applied to find the association in the data and derive knowledge and predict the value off the dependent variables [2]. Few of the data-mining techniques which are commonly used are Naïve Bayes, Decision Tree, Artificial Neural Network (ANN), Bagging Algorithm, K-Nearest Neighborhood (KNN) and Support Vector Machine (SVM). Knowledge discovery in databases (KDD) is used in data mining. Data mining consists of clustering, classification, statistical analysis and prediction of data. Data mining applications include analysis of data for better policy making in health, prevention of various errors in hospitals, detection of fraudulent insurance claims early detection and prevention of various diseases, value for more money, saving costs and saving more lives by reducing death rates [2].

There are two major types of lung cancer. They are Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Each type of lung cancer grows and spreads in different ways and is treated differently. If the cancer has features of both types, it is called mixed small cell/large cell cancer [3]. Screening is a secondary form of prevention of lung cancer. As mentioned earlier, we need to systematically study the symptoms to determine the risk factors. But the reason for contracting lung cancer has changed a lot recently. We cannot say that only smokers are vulnerable to this disease anymore. The air we breathe has become much more polluted and carcinogens may be present in the air we breathe and in the food we eat. The exposure to carcinogens presents in the environment is unavoidable due to which prediction of diseases such as lung cancer has become really complicated. Lung cancer research generally uses clinical tests but the recent advancements in the technology has made it possible to make things more easier in the healthcare industry.

As everything is being computerized and all the data is stored digitally, many ways have come up to use the huge volumes of data in medical research and come up with relations in the data. Knowledge Discovery Databases (KDD), which includes data mining techniques, uses the data to find the patterns and relationships among large number of variables, and the

outcome could be predicted using the historical cases stored in the dataset.

Also various Machine Learning techniques are also being used to detect lung cancer. Machine Learning is now slowly becoming an integral part of healthcare industry. Machine Learning has many applications that can be used in the medical industry. For example, Image Processing can be used to help doctors in predicting the cancer.

We can use Machine Learning classifier techniques such as Artificial neural Network (ANN) and Support Vector Machine (SVM), Naïve Bayes. It has also been observed that application of integration of feature selection and classifier will provide a better result in analysis of cancer data [3].

Various Deep Learning strategies have also been used for obtaining accuracy in the prediction of lung cancer. There are also studies where we see that using computer aided diagnosis process which involves three phases segmentation, detecting and staging process are used for classification of CT images of lung cancer with more accuracy [3][7].

LDCT (low dose computed topography) is used in lung cancer screening and reduced lung cancer deaths and is also recommended for high-risk demographic characteristics. But LDCT availability has been reduced and chest X-rays on the other hand are easily available almost everywhere. But the quality of X_rays is not quite good compared to LDCT. So, chest x-rays are being used with computer-aided diagnosis (CADx) system to improve lung cancer diagnostic performance [6].

The convolutional neural network (CNN) is used to detect abnormalities in chest x-rays. In practice, researchers often pre-train CNNs on ImageNet, a standard image dataset containing more than one million images. The trained CNNs are then adjusted on a specific target image dataset. Unfortunately, the available lung cancer image dataset is too small for this transfer learning to be effective, even with a data augmentation trick. To alleviate this problem, the idea of applying transfer learning was proposed several times to gradually improve the performance of the model. In this work, transfer learning is applied twice. Firstly to transfer the model from a general image into the chest x-ray domain, and secondly to transfer the model for lung cancer. Such multitransfer learning can solve the problem of a small sample size and achieves a better task result compared to the traditional transfer learning technique[6].

II.DATA MINING TECHNIQUES

The classification techniques and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. A particular algorithm may not be applied to all the applications due to the complexity of the data types. Therefore, the choice of the data mining algorithm depends not only upon the purpose of the application but also on the compatibility of the dataset.

1.Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is utmost commonly utilized as dimensionality lessening method in the pre-processing stage for machine learning applications in addition to design-classification [2]. We can project a particular dataset on top of a lower-dimensional space using virtuous class reparability so as to decrease computational prices. Linear Discriminant analysis is controlled and also calculates the guidelines which would signify the axes that are applied to make separation amongst multiple types of classes. We have mentioned the steps used for implementing a LDA technique [2]:

Step 1: Calculate the d-dimensional mean vectors intended for the dissimilar classes from the specific dataset.

Step 2: Calculate the disseminate matrices i.e. between-class as well as within-class scatter matrix.

Step 3: Evaluate the eigen vectors (e1, e2... ed) as well as corresponding eigen values ($\lambda 1$, $\lambda 2$, ...,λd) for the disseminate matrices.

Step 4: Sorts the eigenvectors by diminishing eigenvalues as well as select k eigenvectors using the leading eigenvalues in the direction of forming a d×k-dimensional matrix W i.e. where every particular column exemplifies an eigenvector.

Step 5 : Afterwards, utilize this d×k eigenvector matrix towards transforming the samples onto the new subspace. This could be précised by utilizing the equation $Y = X \times W$ i.e. where X is an n×d-dimensional matrix; the ith row signifies the ith sample, and Y is the converted n×k- dimensional matrix using the n samples anticipated into the new subspace[2].

2. Knowledge Discover And Data Mining

We list out the analysis tasks that can be goals of a discovery process and lists methods and research areas that are promising in solving various tasks.

2.1 Knowledge Discovery Process

KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. The below image shows data mining as a step in an iterative knowledge discovery process.

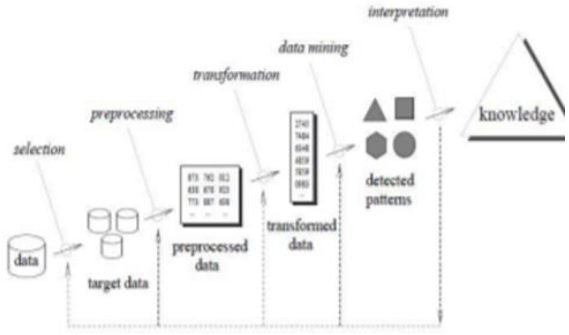

Figure 1.1. Iterative discovery process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps[3]:

1. Data cleaning: also known as data cleansing it is a phase in which noise data and irrelevant data are removed from the collection.
2. Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
5. Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6. Pattern evaluation: this step, strictly interesting patterns representing knowledge are identified based on given measures.
7. Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

2.2 Data Mining Process

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms.

Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects). Following are the few steps involved in data mining[3]:

1. Problem definition: The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioral model.
2. Data exploration: If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.
3. Data preparation: The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.
4. Modeling: Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule-based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.
5. Evaluation and Deployment: Based on the results of the data mining algorithms, an analysis is

conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

The following image shows the data mining process.

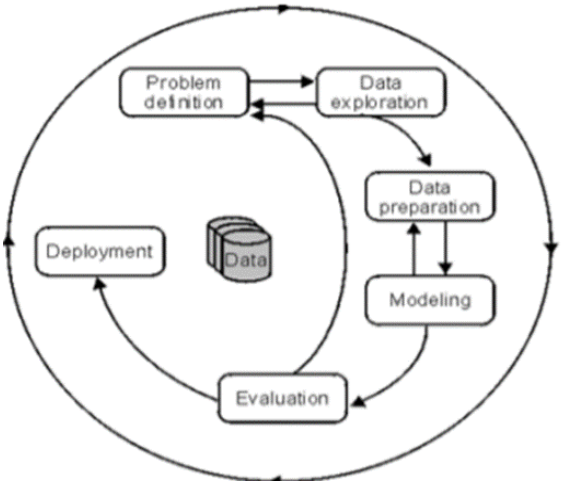

Figure 1.2. Data Mining Process

2.3 Data Mining classification methods

The data mining consists of various methods. Different methods serve different purposes, each method offering its own advantages and disadvantages. In data mining, classification is one of the most important task. It maps the data into predefined targets. It is a supervised learning as targets are predefined.

The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The most used classification algorithms exploited in the microarray analysis belong to four categories: IFTHEN Rule, Decision tree, Bayesian classifiers and Neural networks.

2.1.1 IF-THEN RULE

This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction.

In the health care system it can be applied as follows: (Symptoms) (Previous--- history) → (Cause—of---disease).

2.1.2 Decision Tree Algorithm:

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree based models which include classification and regression trees, are the common implementation of induction modeling. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications [3].

2.1.3 Bayesian classifiers and Naive Bayesian:

From a Bayesian viewpoint, a classification problem can be written as the problem of finding the class with maximum probability given a set of observed attribute values. Such probability is seen as the posterior probability of the class given the data and is usually computed using the Bayes theorem.

Estimating this probability distribution from a training dataset is a difficult problem, because it may require a very large dataset to significantly explore all the possible combinations. Conversely, Naive Bayesian is a simple probabilistic classifier based on Bayesian theorem with the (naive) independence assumption. Based on that rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation. Despite its simplicity, the Naive Bayes classifier is known to be a robust method, which shows on average good performance in terms of classification accuracy, also when the independence assumption does not hold

2.1.4 Neural Network Architecture:

Especially, the neural network approach has been widely adopted in recent years. The neural network has several advantages, including its nonparametric nature, arbitrary decision boundary capability, easy

adaptation to different types of data and input structures, fuzzy output values, and generalization for use with multiple images.

Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. (Actual biological neural networks are incomparably more complex.) Neural nets may used in classification problems (where the output is a categorical variable) or for regressions (where the output variable is continuous).

The architecture of neural network is shown below and it consists of three layers such as input layer, hidden layer and output layer. The nodes in the input layer linked with a number of nodes in the hidden layer. Each input node joined to each node in the hidden layer. The nodes in the hidden layer may connect to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables.

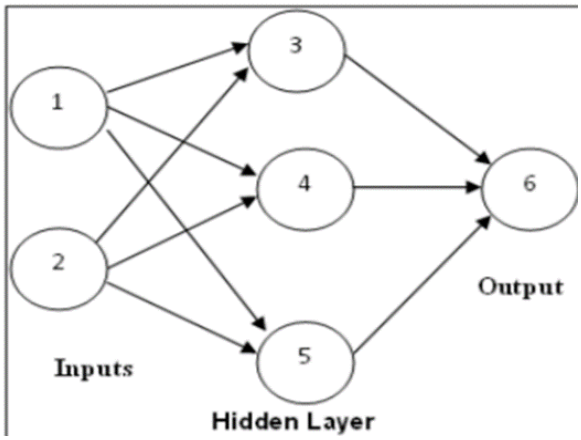

Figure 1.3. Layers of Neural Network

III.MACHINE LEARNING CLASSIFICATIONS

CLASSIFICATION COMES UNDER SUPERVISED LEARNING PROCESS IN ORDER TO PREDICT GIVEN INPUT DATA TO A CERTAIN CLASS LABEL. THE NOVELTY IN CLASSIFICATION RELIES ON MAPPING INPUT FUNCTION TO A CERTAIN OUTPUT LEVEL. VARIOUS LEARNING CLASSIFIERS ARE DESCRIBED AS PERCEPTRON, NAÏVE BAYES, DECISION TREE, LOGISTIC REGRESSION, K NEAREST NEIGHBOUR, ARTIFICIAL NETWORK, SUPPORT VECTOR MACHINE. CLASSIFICATION IN MACHINE LEARNING IS ONE OF PRIOR DECISION MAKING TECHNIQUES USED FOR DATA ANALYSIS[4].

1)Radial Basis Function Network

Radial basis function network comes under neural network that uses radial basis function as its threshold function. RBF network has advantage of easy of design and strong tolerance to input noises. Radial basis Function is characterized by feed forward architecture which comprises of an one middle layer between input and output layer. It uses a series of basis function that are centered at each sampling point[4].

The X network can be written as follows:

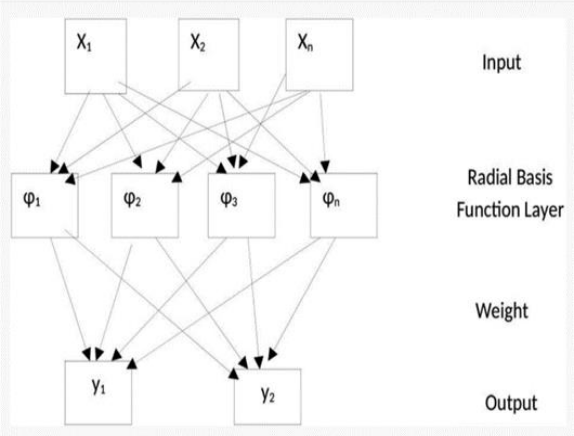

Figure 1.4. Network of Radial Basis Function

Where

$y=\prod Ni=1wiRi(x)+w0y=\prod i=1NwiRi(x)+w0$

wi: weight, w0: biasterm, R: Activation function

$Ri(x)=\varphi[\|xi-ci\|]Ri(x)=\varphi[\|xi-ci\|]$

φ: radial function, ci: RBF centre

In RBF architecture the weight that connects to input unit and middle layer represents the centre of the corresponding neuron where as weights connecting to middle layer and output layer are used to train the network.

2)Support vector classifier

One of the simple and useful approaches in supervised learning is support vector classification. Support vector classifier (SVC) is usually preferred for data analysis because of its computational capability with in very less time frame. This classifier works on the decision boundary concept Recognized as hyper plane. The hyper plane is used to classify the input data in to required target group. But in order to fit the decision boundary in a plane maximize distance margin is

chosen from data points for classification. User defined support vector classifier can be framed using various kernel function to improve the accuracy. Support vector classifier is well suited for both structured and unstructured data. Support vector classifier is not affected with over fitting problem and makes it more reliable[4].

3)Logistic Regression Classifier

Logistic Regression classifier is brought from statistics. These classifiers is based on the probability of outcome from the input process data. Binary logistic regression is generally preferred in machine learning technique for dealing with binary input variables. To categorize the class in to specific category sigmoid function is utilized. Advantages of Logistic Regression classifier.

- Logistic regression classifier is very flexible to implement
- Suitable for binary classification
- Depend on probabilistic model

4) KNN Classifier

Knn classifier comes under lazy learning process in which training and testing can be realized on same data or as per the programmer's choice. In the process, the data of interest is retrieved and analyzed depending upon the majority value of class label assigned as per k, where k is an integer. The value of k is based on distance calculation process. The choice of k depends on data. Larger value of k minimizes the noise on classification. Similarly Parameter selection is also a prominent technique to improve the accuracy in classification. Weighted Knn classifier:

A mechanism in which a suitable weight can be assigned to the neighbor's value so that its contribution has great impact to neighbors than distant ones. In the weighted knn approach the weight has a significant value in evaluating the nearest optimistic value. Generally the weight is based on reciprocal of distance approach. The weight value of attribute is multiplied with distance to obtain the required value[4].

Pseudo code for Knn

- Take the input data
  Consider initial value of k
- Divide the train and test data
- For achieving required target iteration for all training data points
- Find the distance between test data and each row of training data. (Euclidean Distance is the best Choice approach)
- Arrange the calculated distance in ascending order based on distance values.
- Consider the Top k value from sorted value.
- Find the Majority class label
- Obtain the target class.

IV.RESULT

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques. More recently machine learning has been applied to cancer prognosis and prediction. A growing dependence on protein biomarkers and microarray data, a strong bias towards applications in prostate and breast cancer, and a heavy reliance on "older" technologies. A number of published studies also appear to lack an appropriate level of validation or testing, authors say. The research shows that machine learning methods can be used to substantially (15–25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality, they say.

Linear Discriminant Analysis is utmost commonly utilized as dimensionality lessening method in the pre-processing stage for machine learning applications in addition to design-classification. We can project a particular dataset on top of a lower-dimensional space using virtuous class reparability so as to decrease computational prices. This could be précised by utilizing the equation $Y = X \times W$ i.e. where X is an n×d-dimensional matrix; the ith row signifies the i th sample, and Y is the converted n×k- dimensional matrix using the n samples anticipated into the new subspace.

The Naive Bayes algorithm is a straightforward probabilistic classifier that computes a set of probabilities by calculating the frequency and combinations of individuals in a given data set. The probability of a particular element in the data appears as a member throughout the set of possible outcomes and is calculated by measuring the correlation of each feature value within such a training data set class. The training dataset is a subset which is used to train a

classifier algorithm by predicting future, unknown values using known values. The algorithm employs the Bayes theorem implies that all attributes are independent of the class variable's value.

Data mining methods are used in the KDD process to extract patterns from data.The extraction of implicit, previously unknown, and potentially useful information from data in databases is referred to as KDD..The Knowledge Discovery in Databases process is made up of several steps that lead from raw data collections to some form of new knowledge. The patterns that can be discovered are determined by the data mining tasks used. In general, data mining tasks are classified into two types: descriptive data mining tasks that describe the general properties of existing data and predictive data mining tasks that attempt to make predictions based on available data.

Knn classifier belongs to the lazy learning process, in which training and testing can be done on the same data or on data of the programmer's choice. The relevant data is retrieved and analysed based on the majority value of the class label assigned as per k, where k is an integer. The value of k is determined by the distance verification stage. The value of k is determined by the data. A higher k value reduces classification disturbance. Comparably, methodology is a popular technique for improving classification accuracy.

ANNs analyse risk of mortality using a proactive attitude, and their internal configuration can be updated around a functional target using bottom-up calculation (the data are used to generate the model itself). Given their inability to cope with incomplete information, ANN models can process several variables at the same time. Outliers and nonlinear relations between variables can be taken into account by ANNs. As a result, while current figures show variables that are only meaningful for the population overall, the ANN model contains parameters that are relevant at the person level even though they are not substantial for the population overall.[8]

The most recent works relevant to cancer prediction/prognosis using ML techniques are presented in this review. Following a brief description of the ML branch and the concepts of the data preprocessing methods, feature selection techniques, and classification algorithms used, we mainly focused on the prediction of cancer susceptibility, cancer recurrence, and cancer survival. Clearly, there have been a large number of ML studies published in the last decade that provide accurate results regarding specific predictive cancer outcomes.

However, identifying potential drawbacksHowever, identifying potential flaws in the experimental design, collecting appropriate data samples, and validating the classified results is critical for clinical decision-making. It's also worth noting that, despite the claims stating that these Machine Learning classification techniques can lead to adequate and effective decision-making, only a few have made it into clinical practice. Recent advances in omics technologies have paved the way for a better understanding of a variety of diseases, but more precise validation results are required before gene expression signatures can be used in clinics. One of the most common limitations noted in the studies surveyed in this review is the small number of data samples. The size of the training datasets, which must be sufficiently large while using classification schemes to model a disease, is a necessary prerequisite while using classification schemes to model a disease. A significantly larger dataset allows for adequate partitioning into training and test datasets, resulting in reasonable estimation validation. Misclassifications can occur when the training sample is too small in comparison to the data dimensionality, and the estimators can produce unstable and biassed models. It is obvious that a larger sample size of patients who are up for survival prediction can improve the generalizability of the predictive model. We discovered that SVM and ANN classifiers were widely used as one of the most commonly used ML algorithms relevant to cancer patient forecast outcomes. As we mentioned in the introduction, ANNs have been widely used for nearly 30 years .[8]

Furthermore, SVMs are a more recent approach in cancer prediction/prognosis and have been widely used due to their accurate predictive performance. However, the most suitable algorithm is determined by a number of factors, including the type of data collected, the size of the data samples, the time constraints, and the type of prediction interventions. Regarding the long term of cancer modelling, new techniques for overcoming the limitations discussed above should be investigated. A more accurate analysis of the data of the heterogeneous datasets used would yield more accurate results and provide reasoning for disease outcomes. More research is needed based on the creation of more public databases

that accumulate valid cancer datasets from all patients who have been diagnosed with the disease. The researchers' use of them would make modelling studies easier, resulting in more accurate results and interconnected clinical decision making.

V.CONCLUSION

Lung cancer being the deadliest causing the most number of deaths out of all types of cancer. Ther are literally more than 100 types of cancer lung cancer being one of them. Since the symptoms of lung cancer appear at the advanced stages , it is hard to diagnose properly at the early stages if we succeed in doing so gives a greater chance in treating the patients. hence early stage prediction of cancer is mandatory to diagnose the patients with cancer properly and it also would probably save the life of the cancer patient.it is expected that machine learning systems in general can increase the speed and accuracy of the diagnosis and decision making among doctors thus decreasing the costs of the treatment and also in giving proper or required medication to thee victims. In this paper we will be going through the advancements being made or that has been made in predicting lung cancer in patients at early stages. with the advancements being made in the present-day technology we can diagnose cancer in its early stage. Machine learning based predictions for lung cancer have been proved effective in the medical field by making it easier to identify the roots of the cancer. By using Support Vector Machines help in recognizing the patterns and analyzing data.

ANN reasoning can be used in tasks requiring attention focusing. ANNs have a place in clinical care as well, but their effectiveness is dependent on greater integration of clinical guidelines, as well as an understanding about the need to integrate various frameworks in order to achieve the easiest and most straightforward overall rationale framework, and the willingness to test this in a specific clinical setting.

Data mining also helps in predicting the chance of a cancer in early stages by analyzing raw data of the victims, we can use their symptoms as a dataset and predict the chances of their condition getting worse. Artificial Neural Network also helps in predicting early-stage cancer symptoms as it is cluster of neuron network similar to that of a biological one like in human brain. and these neurons are connected as a system and are interconnected with each other and these exchange information or messages between them like a biological neuron these neurons can adapt with experience thus creating neural nets that are adaptive to inputs which are capable of learning . Artificial Neural Network has many advantages like having long training time, weight adjustment and high cost in computing. Our main aim through this process is to determine the cancer at early stage thus saving money and the life of the victim. The performance of this method that is proposed shows effective results which in turn proves that usage of neural networks in the field of science in determining cancer in patients might help the doctors in diagnosing in early stages by giving proper medication to the victims.

In the KDD process, the data mining methods are for extracting patterns from data. There are two types of data mining tasks: descriptive data mining and predictive data mining. Data mining can be done on data which are in quantitative, textual, or multimedia forms. The patterns that can be discovered depend upon the data Mining tasks applied. Data preparation which is one of the few steps in data mining having the purpose of this step is to clean and transform the data so that missing and invalid values are treated, and all known valid values are made consistent for more robust analysis. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.

Though screening would help in detection of cancer in patient it takes lot of time or it detects after it is too late too diagnose as screening only shows us the cancer causing part through the scans only after when it is in the severe or final state as it is hard to identify in some cancer types the lumps being formed or the cancer part inside wouldn't be that easy to show up in the scan. And the technological advancements play a vital role in this point by giving access to full data of the patients their previous medical records and with the help of machine learning techniques we can easily predict early symptoms of lung cancer.

Now we know about the Machine learning and Data mining techniques which are being used to predict and prognosis cancer.

Machine learning methods clearly improve the performance or predictive accuracy of most forecasts. We believe that if study quality continues to improve, the use of machine learning classifiers will become much more common in many clinical and hospital settings.

REFERENCES

[1] Dr.T.Christopher, J,Jamera Banu, Study Of Classsification Algorithm for Lung Cancer Prediction, IJISET- International journal Of Innovative Science, engineering & technology, Vol.3 Issue 2, February 2016.

[2] Varun Jaiswal, Divya Chauhan, An Effieicent Data Mining Classification Approach for Detecting Lung Cancer Disease, International Conference on Communication and Electronics Systems(ICCES), 2016.

[3] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra, Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, (IJCSIT) International Journal of Computer Science and Information Technologies, 2013.

[4] Radhanath Patra, Prediction of Lung Cancer using Machine Learning Classifier, International Conference on Computing Science, Communication and Security, 2020.

[5] Ahmad S. Ahmad, Ali M. Mayya, A new tool to predict lung cancer based on risk factors, cell.com/heliyon, 2020

[6] Worawate Ausawalaithong, Sanparith Marukatat, Arjaree Thirach, Theerawit Wilaiprasitporn, Automatic lung cancer prediction frm chest X-ray images using the deep learning approach, Biomedical Engineering International Conference, 2018.

[7] Paing, M.P., Hamamoto, K., Tungjitkusolmun, S., Pintavirooj, C.: Automatic detection and staging of lung tumors using locational features and double-staged classifications. Appl. Sci. 9(11), 2329 (2019)

[8] Shi HY, Hwang SL, Lee KT, et al. In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. J Neurosurg2013;118:746-52.10.3171/2013.1.JN S121130.