Lung Cancer Detection using Decision Tree Algorithm

Ms. Leena Patil, Ms. Aparna Sirsat, Ms. Diksha Kamble, Mr.Yogesh Pawar

BE IT, Department of Information Technology, DY Patil Institute of engineering and Technology *Maharashtra, India*
BE IT, Department of Information Technology, DY Patil Institute of engineering and Technology *Maharashtra, India*
BE IT, Department of Information Technology, DY Patil Institute of engineering and Technology *Maharashtra, India*
HOD,Department of Information Technology, DY Patil Institute of engineering and Technology *Maharashtra, India*

---***---

Abstract – *Lung cancer, also known as lung carcinoma a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. If left untreated, this growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. Most cancers that start in the lung, known as primary lung cancers, are carcinomas. The two main types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC).Cigarette smoking is the principal risk factor for development of lung cancer. A Few popular technique are used to Detect the lungs cancer like support vector machine. (SVM), naive bayes classifier. A new approach to detect the lungs cancer by Decision tree algorithm will provide effective result as compare to other algorithm. The proposed system will enhance the performance of prediction and classification.*

Key Words: **Artificial Neural Network, Decision Tree, feed forward Neural Network.**

1. INTRODUCTION

Cancer is a group of diseases involving abnormal cell growth with the potential to spread to other parts of the body. Not all tumors are cancerous; benign tumors do not spread to other parts of the body. Possible signs and symptoms include a lump, abnormal bleeding, prolonged cough, unexplained weight loss and a change in bowel movements. While these symptoms may indicate cancer, they may have other causes.[3] Over 100 types of cancers affect human.

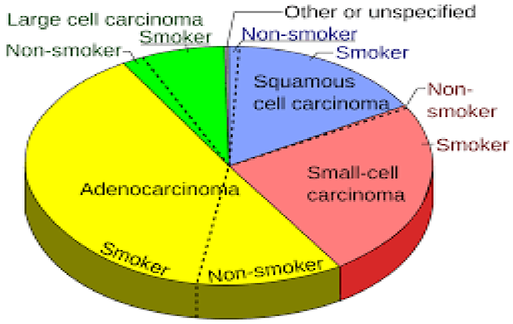

Figure 1.Pia chart with fraction of smokers versus non-smokers

Lung cancer may not produce any noticeable symptoms in the early stages. In approximately 40 percent of people diagnosed with lung cancer, the diagnosis is made after the disease has advanced. In one-third of those diagnosed, the cancer has reached stage 3.

2. DATA SET

The dataset for the project is taken from the UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/Lung+Cancer

In this data set:-
Number of Instances: 32, Number of Attributes: 57 (1 class attribute, 56 predictive)
Attribute Information:
attribute 1 is the class label.

3. Lungs Cancer Detection By ANN

An artificial neural network (ANN) is a massively the parallel distributed processor made up of simple processing units called neurons. The neurons have a natural capability for storing experiential knowledge and making it available for use [2].

Every neural network structure has the do undergo a training phase with the available data or patterns. This training/learning phase uses a suitable learning algorithm. The prime objective of the learning algorithm is to modify the synaptic weights of the network in an the orderly fashion so as to attain a desired design objective and to increase the accuracy of the learning stage minimizing the error. The working of ANN can be divided in two phases one is training phase and other is recalling phase or testing phase. In training phase both the input pattern and its corresponding target output is supplied to the network. Input is given to the network at input node, the input layer neuron processed the input by using activation function and gives its output. The output from this layer is given as input to the next level neuron, and so on up to the output node. The links connected between neurons are having some weight. These weights are updated by some learning algorithm in training phase till the error between the network output and actual output for that input data set or pattern is minimized. The level of error depends on the learning algorithm, quality of data and type of network. Once the minimized error is obtained the other inputs are given to trained network to get the output. This is the recall phase or testing phase

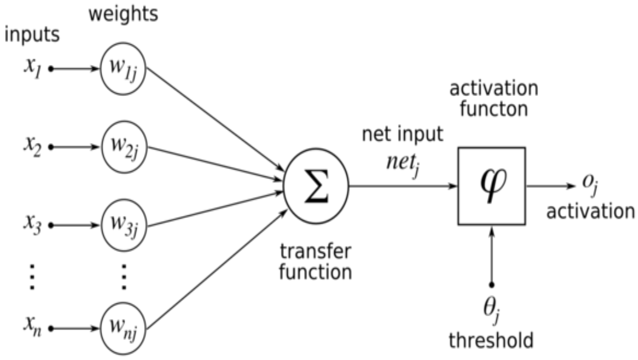

Figure 2. Architecture of neural network

This describes the modules, which should be considered design a good neural network model for prediction and classification.

3.1 Input Selection

The aim of the input selection in the case of ANN is finding optimal input parameters. Using optimal inputs would result in smaller ANN with more accuracy and convergence speed. Parameters, which effect on the prediction and classification of Lung Cancer can be categorized into age, patient history, lung condition value are selected by correlation analysis.

One of the reasons for using fewer features was the limited number of data records (452) compared to 257 features. This helps in avoiding over fitting and also gives insight into the important features which have maximum correlation with the output labels but minimal correlation among themselves.

3.2 Training

The ANN training process requires a set of examples with proper network behaviour (network inputs and target outputs). During training, the weights and biases of the ANN are is the iteratively adjusted to minimize the network performance function. the selected training method for the new ANN models is the Levenberg-Marquart back propagation (LMBP), which is a network training function that updates weight and bias values are according to Levenberg-Marquardt optimization.

This method is an improve Gauss-Newton method that has an extra regularization term is to deal with the additive noise in the training samples. In comparison to LMBP, conventional back propagation methods are often too slow for practical problems

Neurons in the hidden and output layers have nonlinear transfer function is known as the “tangent sigmoid".

The weighted inputs received by a tansig node are summed and passed through this function is to produce an output. The tansig function generates outputs between -1 and +1 and its inputs should be in the same range in the system. So, it is necessary to limit the ANN inputs and target outputs. Mean-standard the deviation and minimum (min)-maximum (max) normalization methods have been tested and min-max method has been selected:

This normalization method has also the advantage of mapping the target output to the non-saturated sector of the tansig function. This is the process helps to improve the accuracy of both the training and Prediction modes [3].

3.3 Output and Hidden Layer

The ANN models have the output layer. In the model of prediction and classification of Lung cancer the output is the weather, patient is having Lung cancer or not and if yes then which type of Lung cancer it is , so the output layer has the only one neuron.

The number of hidden layers and the number of the neurons in each layer are selected whereas the best results are obtained.

4. Decision Tree Approach

4.1 Decision tree algorithm

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes).[4] The paths from root to leaf represent classification rules.

4.2 Steps of Construct a Decision-Tree

There are few steps for construction of Decision tree:

1. First step is check whether all the cases belong to the same class and if Yes then tree is a leaf and that node is labelled by that class.
2. Entropy and information gain are calculated for each and every attribute
3. Assume best selection criteria and accordingly consider the splitting attribute.
4. Counting the information gain: The concept of entropy arrives in this part. Entropy can be stated as its measure of any disordered in the data. Entropy can also be called as a measurement of uncertainty in any random variable.
5. Pruning: For the tree creation process, pruning is an important technique to be performed. The dataset may sometimes contain subsets that are not well defined of instances, so for classification of such a subsets, Pruning can be used [4].
6. Pruning has two types:

   1. Post Pruning: This type of Pruning is performed after the creation of tree.

   2. Online Pruning: This type of Pruning is performed during the process of tree creation.

4.3 Formulas for Entropy Calculation

Entropy= - p(a)*log(p(a)) – p(b)*log(p(b))

P(a) and P(b) is the probability of class (a) and (b) Compute it as the proportion of class a&b in the set.

Information Gain=entropy (after)- entropy (before)

Probability of class: [No of instances of particular class/ Total no of instances]

Example->

```
ends – vowel

        [9m,5f]              <---the [...,...]>

Notation    reprents      the class distribution of

     /            \

Instances that reached a node

=0                =0

--------       ----------

[3m,4f]        [6m,1f]
```

As you can see, before the split we had 9 males and 5 females, i.e. P(m)=9/14 and P(f)=5/14. According to the definition of entropy:

Entropy before = -P(f) * log2 p(f) – p(m) log2 p(m)

Entropy before = - (5/14) * log2(5/14) - (9/14) * log2 (9/14) = 0.9403

Next we compare it with the entropy computed after considering the split by looking at two child branches. In the left branch of ends-vowel=1, we have:

Entropy left= - (3/7) * log2 (3/7) – (4/7) * log2 (4/7) =0.9852

Entropy right = - (6/7) * log2 (6/7) - (1/7) * log2 (1/7)=0.5917

We combine the left/right entropies using the number of instances down each branch as weight factor (7 instances went left, and 7 instances went right), and get the final entropy after the split:

Entropy after = 7/14 * entropy left + 7/14* entropy right=0.7885

Now by comparing the entropy before and after the split, we obtain a measure of information gain,or how much information we gained by doing the split using that particular feature:

Information Gain = [Entropy before-Entropy after]=0.1518

5. PROPOSED SYSTEM

System to automate the classification process for the early detection of Lung cancer.

5.1 System Architecture:-

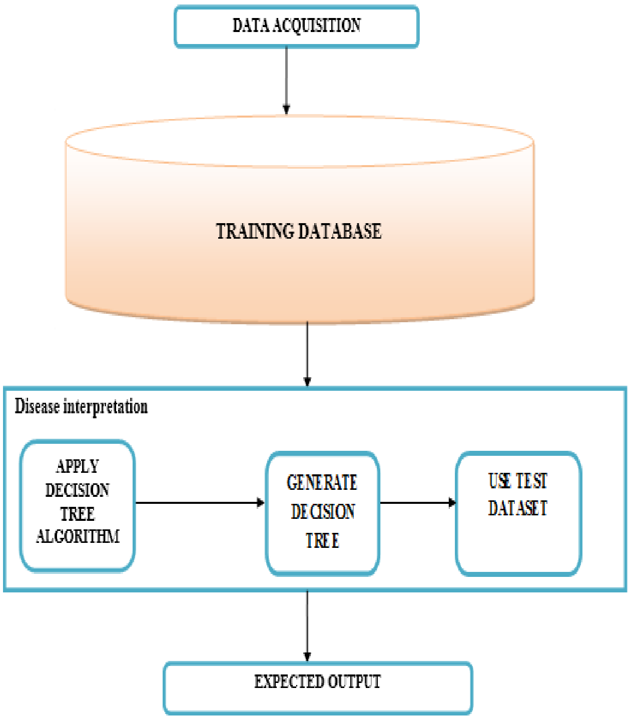

Figure 3.System Architecture

5.2 Modules-:

1. **DATA ACQUISITION**

The Data acquisition is a process of finding a data required for the system and load it into the database with the correct format. By cleaning and transforming a data into readable and understandable format.

2. **STORING INFORMATION IN DATABASE**

The Data comes from the ETL process which are loaded into the different Databases as per requirement.

3. **APPLY DECISION TREE ALGORITHM**

The Data comes into the database is of training data , through which the system is trained. To make proper decision on lung cancer Decision tree algorithm is applied on available data to get system train and ready to take decision for unknown data.

4.GENERATE DECISION TREE

Once Decision tree algorithm is applied on training data, it generates an tree like structure based on data available in training database. Splitting and aggregation of data is done while decision tree is generating.

5. USE TEST DATASET

The testing dataset is a dataset in which result is unknown. This dataset is used for testing a system towards the goals set by the system.

6. EXPECTED OUTPUT

Expected output is an output which user need from the system. It is also called final output from the system. In which user get the answer of whether they have lung cancer or not.

6. CONCLUSIONS

In earlier times for detection of any type of Lung cancer the doctor has to do multiple tests of patient. But this was very time consuming process to get clarify whether the patient is have lung cancer or other type of disease. In a research sometimes patient have to do unnecessary checkups or different tests to identify the disease of lung cancer. To minimize the process time and unnecessary checkups there needs to be a one preliminary test in which patient and doctor both will get clarify with the possibilities of lung cancer. Nowadays the machine learning algorithms plays a vital role in prediction and classification of data. KNN, SVM, decision tree, ELM are the most popular algorithms available in the machine learning. The decision tree algorithm will gone very useful for implementation of Lung cancer Disease with very much accuracy and fast.

REFERENCES

[1]Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. Devi Prasad Bhukya1 and S. Ramachandram2.

[2] A Survey on Decision Tree Based Approaches in Data Mining. Shahrukh Teli M-tech student MPSTME, SVKM'S NMIMS University,Mumbai, India.

[3]Induction of decision tress. J.R. Quinlan centre of advanced computing sciences, new south wales Institute of technology Sydney 2007, Australia.

[4]A New Decision Tree Method For Data Mining In Medicine Kasra Madadipouya1 Department of Computing and Science, Asia Pacific University of Technology & Innovation.