A New Approach For Prediction of Lung Carcinoma Using Back Propogation Neural Network with Decision Tree Classifiers

Ching-Hsien Hsu[1], Gunasekaran Manogaran[2], Parthasarathy Panchatcharam[3] and Vivekanandan S.[3]

[1]Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan
[2]University of California, Davis, USA
[3]VIT University, Vellore, India

***Abstract* — In this paper an investigation was made to examine the lung tumour expectation utilizing classification algorithms, for example, Back Propagation Neural Network and Decision Tree. At first 20 tumour and non-disease patients' examples information were gathered with 30 qualities, pre-prepared and dissected utilizing classification algorithms and later a similar methodology was actualized on 50 occurrences (50 Cancer patients and 10 non growth patients). The informational indexes utilized as a part of this examination are taken from UCI data sets for patients affected by lung cancer and Michigan Lung Cancer patient's informational index .The principle point of this paper is to give the prior notice to the clients and to quantify the execution investigation of the classification algorithms utilizing WEKA Tool. Test comes about demonstrate that the previously mentioned calculation has promising outcomes for this reason with the general forecast exactness of 94% and 95.4%, separately. Another way to deal with identifies the lungs tumour by Decision tree and BPNN calculation will give viable outcome as contrast with other calculation. The proposed framework will improve the execution of prediction and classification.**

Keywords—Back Propogation, Decision tree, Lung cancer, Classification algorithms

I. INTRODUCTION

Lung malignancy is the one of the main source of disease passings in the two ladies and men. Appearance of Lung growth in the body of the patient uncovers through early manifestations in the vast majority of the cases (Ahyaningsih et al. 2017). Treatment and visualization rely upon the histological kind of disease, the stage (level of spread), and the patient's execution status. Conceivable medicines incorporate surgery, chemotherapy, and radiotherapy Survival relies upon organize, general wellbeing, and different elements, yet general just 14% of individuals determined to have lung malignancy survive five years after the analysis. Manifestations that may recommend lung malignancy include: dyspnea (shortness of breath with movement), haemoptysis (hacking up blood), incessant hacking or change in consistent hacking design, wheezing, chest torment or agony in the midriff, cachexia (weight reduction, weariness, and loss of craving), dysphonia (rough voice), clubbing of the fingernails (extraordinary), dysphasia (trouble gulping), Pain in bear, chest, arm, Bronchitis or pneumonia, Decline in Health and unexplained weight reduction. Mortality and bleakness because of tobacco utilize is high. Generally lung tumour creates inside the divider or epithelium of the bronchial tree. In any case, it can begin anyplace in the lungs and influence any piece of the respiratory framework. Lung disease for the most part influences individuals between the ages of 55 and 65 and frequently takes numerous years to create (Alatas et al. 2011 and Ben David et al. 2005).

There are two noteworthy sorts of lung tumour are Non-little cell lung malignancy (NSCLC) and little cell lung growth (SCLC) or oat cell disease. Each sort of lung tumour develops and spreads in various ways, and is dealt with in an unexpected way. On the off chance that the tumour has highlights of the two kinds, it is called blended little cell/vast cell disease. Non-little cell lung tumour is more typical than SCLC and it by and large develops and spreads all the more gradually. SCLC is relatively related with smoking and develops all the more rapidly and frame extensive tumours that can spread generally through the body. They regularly begin in the bronchi close to the focal point of the chest. Lung disease passing rate is identified with aggregate sum of cigarette smoked (Benjaafar et al. 2002, Benlic U 2013, and Sang Min et al. 2006).

Information mining assumes an indispensable part in the disclosure of learning from substantial databases. Data mining has discovered its noteworthy hold in each field including healthcare (Hornik K et al. 1990). Data mining has its real part in removing the concealed data in the restorative information base. Mining process is more than the information investigation which incorporates grouping, bunching, and affiliation governed mining and expectation. Lung disease is the most well known reason for malignancy demise overall (Karaolis MA et al. 2010, Parthasarathy P et al. 2018 and Rao BS et al. 2014). A patient influenced with Lung Cancer may feel side effects in different places in the body. The lung malignancy side effects are utilized to anticipate the hazard level of the growth illness. The principle point of this examination is to anticipate the hazard level of lung malignancy (Parthasarathy P et al. 2018).

II. RELATED WORK

The approach that is being taken after here for the forecast procedure depends on precise investigation of the measurable elements, side effects and hazard factors related with Lung disease. Non-clinical side effects and hazard factors are a portion of the non-specific pointers of the tumour ailments. At first the parameters for the pre-analysis are gathered by cooperating with the obsessive, clinical and medicinal oncologists (Domain specialists) (Subbulakshmi et al. 2015).

978-1-7281-0236-8/18/$31.00 ©2018 IEEE
DOI 10.1109/SC2.2018.00023

There is incalculable existing framework that shows different techniques for distinguishing the lung growth. One such framework proposed by (Weng C et al. 2016) propounds the utilization of fluffy rationale for the testing and exactness of the highlights removed and the ACO system to increase grouping of the malignant pictures. In this framework Histogram-Equalization is connected to the DICOM pictures for their upgrade. After the picture improvement, the procedure of highlight extraction is done utilizing the Binarization Approach that involves ordinary and strange pictures. The framework identifies lung disease in its beginning times utilizing fluffy rationale and ACO strategy. Similarly (Sandeep Kumar Saini 2014) proffers the method of summed up Neural Networks and Association Rule Mining with a specific end goal to distinguish the highlights of the pictures and characterize them into malignant and non-harmful. In this framework the picture initially experiences pre-handling stage. In the later stage the component extraction process is done trailed by Rule-age process and utilizing which the pictures are delegated carcinogenic or non-harmful with the assistance of Neural Network for the same. (V. Krishnaiah et al. 2013) proposes a framework that utilizes strategies, for example, If-Then Rules, Decision-Tree, Naïve Bayes Classifier, and Artificial Neural Network. The Decision tree is utilized to penetrate through the database and decipher it. The Naïve Bayes, IF-THEN run and Artificial Neural Network are utilized to order the database. Along these lines the outcomes acquired after grouping are utilized by the lung malignancy forecast framework for early discovery that can spare a patient's life.

(Rahman RM et al 2011) proposes a framework that performs Association Rule mining investigation on lung disease information to recognize hotspots in the tumour information, to analyze the patients' survival time that is altogether higher than or lower than the normal survival time. The framework utilizes SEER database that is produced using the SEER site on permit understanding, the information mining technique for affiliation run mining is actualized to characterize the pictures as dangerous or non-carcinogenic. A two phase affiliation manage mining is utilized where the nonessential guidelines from arrange 1 are disposed of in organize 2 and order is implemented to distinguish hotspots in lung growth information. For programmed recognition of lung disease, (Rao BS et al. 2014) proposes a framework that utilizes receptive Neuro fluffy framework. The framework separates the influenced locale from the lung CT picture utilizing morphological reproduction took after by knob division that is finished utilizing worldwide edge and morphological activities. In the propelled organize the element extraction is executed on these sectioned knobs and in light of these highlights the arrangement of carcinogenic or non-dangerous picture is completed with a help of ANFIS based classifier.

The clinical and imaging symptomatic guidelines of fringe lung tumour by information mining method, and to investigate new thoughts in the conclusion of fringe lung growth, and to acquire beginning time innovation and learning backing of PC supported recognizing (CAD). The information were foreign into the database after the institutionalization of the clinical and CT discoveries properties were distinguished. The analysis rules for fringe lung growth with three information-mining innovation is same as clinical demonstrative tenets, and these guidelines additionally can be utilized to construct the learning base of master framework (Kumar PM et al. 2018). They showed the potential estimations of information mining innovation in clinical imaging conclusion and differential finding.

III. Data Mining Classification

The information mining comprises of different techniques. Distinctive techniques fill diverse needs, every strategy offering its own particular focal points and hindrances. In information mining, grouping is a standout amongst the most critical assignments. It maps the information in to predefined targets. It is a managed learning as targets are predefined. The point of the arrangement is to construct a classifier in view of a few cases with a few ascribes to portray the articles or one credit to depict the gathering of the items. At that point, the classifier is utilized to anticipate the gathering qualities of new cases from the area in view of the estimations of different traits.

The most utilized classification algorithms misused in the microarray investigation have a place with decision tree and neural systems. Decision tree gets from the basic partition and vanquish calculation. In these tree structures, leaves speak to classes and branches speak to conjunctions of highlights that prompt those classes. At every hub of the tree, the quality that most adequately parts tests into various classes is picked. To foresee the class mark of info, a way to a leaf from the root is discovered relying upon the estimation of the predicate at every hub that is gone by. The most well known calculations of the decision trees are ID3 (Chandra et al. 2018) and C4.5. An advancement of decision tree abused for microarray information investigation is the arbitrary woods (Diaz-Uriate R et al. 2006), which utilizes an outfit of grouping trees. (Rahman RM et al 2011) Showed the great execution of arbitrary woodland for boisterous and multi-class microarray information. A manufactured neural system is a scientific model in light of organic neural systems. It comprises of an interconnected gathering of fake neurons and procedures data utilizing a connectionist way to deal with calculation. Neurons are sorted out into layers. The info layer comprises basically of the first information, while the yield layer hubs speak to the classes. At that point, there might be a few concealed layers. A key element of neural systems is an iterative learning process in which information tests are exhibited to the system each one in turn, and the weights are balanced keeping in mind the end goal to foresee the right class name. Preferences of neural systems incorporate their high resistance to loud information, and their capacity to characterize designs on which they have not been prepared. In (J R Quinlan 1986) an audit of points of interest and drawbacks of neural systems with regards to microarray examination is displayed.

IV. The Proposed Methodology

In several previous studies various decision tree models and unsupervised clustering algorithms were employed to identify the most important protein attributes and obtaining the best classification of lung tumours based of them. In this study we proposed Back Propagation and decision tree methods to predict the type of lung tumour based on machine learning and training capabilities.

4.1 Lung Cancer Detection by BPNN

An Artificial neural system (ANN) is a parallel-dispersed processor made up of basic preparing units called neurons. The neurons have a characteristic capacity for putting away experiential learning and making it accessible for utilize.

Each neural system structure has a preparation stage with the accessible information or examples. This preparation/learning stage utilizes a reasonable learning calculation. The prime target of the learning calculation is to alter the synaptic weights of the system in a deliberate mold to accomplish a coveted outline objective and to build the precision of the learning stage limiting the mistake. The working of ANN can be partitioned in two stages, one is preparing stage and other is reviewing stage or testing stage. In preparing stage, both the example and its relating target yield are provided to the system. Information is given to the system at input hub; the information layer neuron handled the contribution by utilizing actuation capacity and gives its yield. The yield from this layer is given as contribution to the following level neuron. The connections associated between neurons are having some weight. These weights are refreshed by some learning calculation in preparing stage till the blunder between the system yield and real yield for that info informational index or example is limited. The level of blunder relies upon the learning calculation, nature of information and kind of system. Once the limited blunder is gotten alternate sources of info are given to prepared system to get the yield. This is the review stage or testing stage. This portrays the modules, which ought to be thought about plan as a decent neural system demonstrates for forecast and characterization.

4.2 Training Model

The ANN preparing process requires an arrangement of cases with appropriate system conduct (organize data sources and target yields). Amid preparing, the weights and inclinations of the ANN are is the iteratively changed in accordance with limit the system execution work. The chose preparing strategy for the new ANN models is the Levenberg-Marquart back spread (LMBP), which is a system preparing capacity that updates weight and inclination esteems are as per Levenberg-Marquardt improvement. This technique is an enhance Gauss-Newton strategy that has an additional regularization term is to manage the added substance commotion in the preparation tests. In contrast with LMBP, traditional back engendering techniques are frequently too moderate for handy issues Neurons in the covered up and yield layers have nonlinear exchange work is known as the "digression sigmoid".

The weighted sources of info got by a tansig hub are summed and gone through this capacity is to create a yield. The tansig work produces yields between - 1 and +1 and its data sources ought to be in a similar range in the framework. Along these lines, it is important to restrict the ANN information sources and target yields. Mean-standard the deviation and least (min) - most extreme (max) standardization techniques have been tried and min-max strategy has been chosen: This standardization technique has additionally the benefit of mapping the objective yield to the non-immersed area of the tansig work. This is the procedure enhances the precision of both the preparation and Prediction modes.

4.3 Network Layers

The point of the input selection on account of ANN is finding ideal info parameters. Utilizing ideal sources of info would bring about littler ANN with more precision and union speed. Parameters, which impact on the expectation and arrangement of Lung Cancer can be sorted into age, persistent history, lung condition esteem is chosen by relationship investigation. The ANN models have the yield layer. In the model of expectation and order of Lung growth the yield is the climate, tolerant is having Lung tumour or not and if yes then which sort of Lung malignancy it is, so the yield layer has the just a single neuron. The quantity of shrouded layers and the quantity of the neurons in each layer are chosen though the best outcomes are acquired.

BPNN is a learning calculation utilized for preparing the counterfeit neural system. Basically, the back engendering calculation comprises of two phases i.e. - forward pass and in reverse go through which the different layers or areas of the system are prepared. A Decision tree is a decision help mechanical assembly that uses a tree-like graph or model of decisions and their possible outcomes, including chance event comes about, resource costs, and utility. It is one way to deal with demonstrate a figuring. A decision tree is a flowchart-like structure in which each internal center addresses a "test" on a quality (e.g. despite whether a coin flip comes up heads or tails), each branch addresses the after effect of the test and each leaf center point addresses a class name (decision taken in the wake of enrolling all qualities). The algorithm of BPNN is given as follows

Algorithm 1 Back Propagation Neural Network

Input: Initial weights & information vector

```
1. Compute $\partial_j = (y_j - d_j)$ for all yield neurons, where $d_j$ is the coveted
   yield of neuron $j$ and $y_j$ is its present yield: $y_j = g(\sigma_i w_{ij} x_i) = (1 +
   e\text{-}\Sigma w_{ij} x_i) - 1$, expecting a sigmoid actuation work.
2. For residual neurons (from last concealed layer to first), process
   $\partial_j = \Sigma kwjkg'(x)\partial_k$, where $\partial_k$ is the $\partial_j$ of the succeeding layer, and
   $g'(x) = y_k(1 - y_k)$.
3. Refresh the weights as indicated by: $w_{ij}(t + 1) = w_{ij}(t) - y_i y_j(1-
   y_j)\partial_j$ , where, $w_{ij}$ is a parameter called the learning rate.
4. Repeat until the point that the blunder is diminished to a pre-
   specified esteem.
```

Algorithm 2 Steps for Decision Tree Construction

```
1. The initial step is to check whether every one of the cases have a
   place with a similar class and if yes at that point tree is a leaf
   and that hub is marked by that class.
2. Entropy and data pick up are computed for every single
   characteristic.
3. Accept best determination criteria and as needs be think about
   the part quality.
4. Tallying the data pick up: The idea of entropy touches base in
   this part. Entropy can be expressed as its measure of any
```

confused in the information. Entropy can likewise be called as an estimation of vulnerability in any irregular variable.

5. Pruning: For the tree creation process, pruning is a vital strategy to be performed. The dataset may here and there contain subsets that are not all around characterized of occasions, so for arrangement of, for example, subsets, Pruning can be utilized.

V. Performance Analysis

Performance investigation was made utilizing the characterization calculations, for example, Neural Network, Decision Tree and J48 calculation for anticipating the Lung Cancer Disease from the given dataset occurrences and the above proposed calculations are connected on Lung Cancer Disease dataset in the Weka (Waikato Environment for Knowledge Analysis) instrument and the execution is estimated. Furthermore, the yield of this calculation can be effectively comprehended by the end client and fulfills the execution measure. Along these lines, J48 choice tree calculation has been utilized as a part of this examination and underneath is the classifier yield subsequent to running in Weka.

Analysis configuration was made on different classification algorithms on the UCI Machine Learning Repository Lung Cancer Dataset. The comparative analysis is given in Tables 1 and 2. In Table 1, Lung cancer data set has 30 instances and 30 attributes. It shows that the decision tree algorithm builds the prediction in 0.1 seconds and the 25 instances were correctly classified and 5 are incorrectly classified.

Table 1. Comparative analysis (30 instances, 30 attributes)

Algorithm	Execution time	Number of instances	Correctly and in-correctly classified instances	Error
Bayes	0.25	30	15, 15	0.234
C 4.5	1.5	30	10, 20	0.267
Multilayer Perceptron	0.75	30	14, 16	0.198
Decision tree	0.1	30	25, 5	0.116

Table 2. Comparative analysis (50 instances, 50 attributes)

Algorithm	Execution time	Number of instances	Correctly and in correctly classified instances	Error
Bayes	0.25	50	25, 25	0.267
C 4.5	1.5	50	20, 30	0.288
Multilayer Perceptron	0.75	50	24, 26	0.180
Decision tree	0.1	50	40, 10	0.101

Figure 1 shows the prediction accuracy of the decision tree based on different number of trainings.

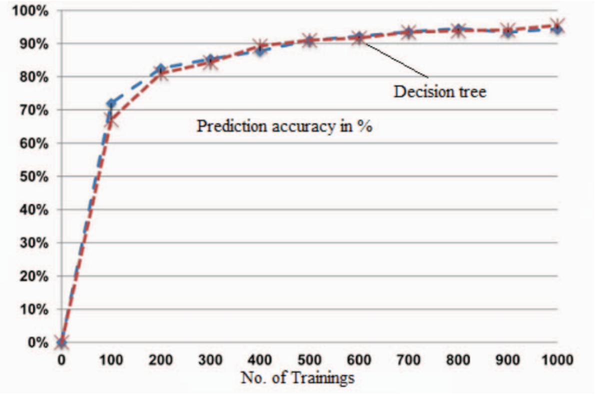

Figure 1. Precision accuracy vs. number of trainings

Ten different tests have been conducted on the same dataset using the decision tree algorithm J48 and Back Propagation model for neural network. In J48, the dataset was split using Pareto principle ratio, 50% training set and 20% test data. As for BP neural network, the data was split into 20 folds using cross validation. Both algorithms predicted at least 91% cases each test. However, Back Propagation Neural network model was able to correctly classify more cases on average as shown in Figure 2.

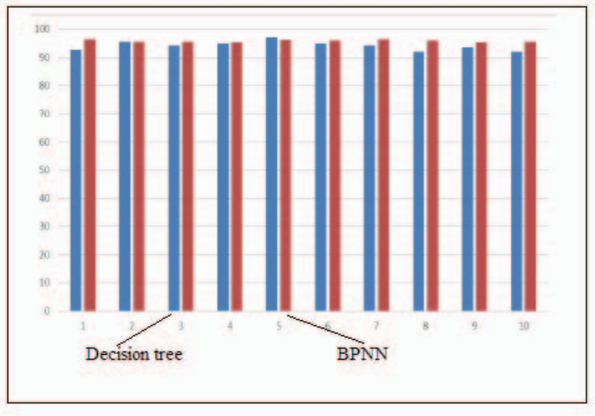

Figure 3 Performance analysis graph between decision tree and back propagation neural network

VI. Conclusions

Cancer is conceivably lethal infection. Identifying malignancy is as yet trying for the specialists in the field of medication. Indeed, even now the real reason and finish cure of growth isn't imagined. Identification of tumour in prior stage is reparable. In an exploration some of the time tolerant needs to do pointless checkups or distinctive tests to distinguish the infection of lung tumour. To limit the procedure time and pointless checkups there should be a one preparatory test in which patient and specialist both will get clears up with the potential outcomes of lung malignancy. These days the machine learning calculations assumes a fundamental part in expectation and characterization of information. KNN, SVM, choice tree, ELM are the most well known calculations accessible in the machine learning. The choice tree calculation will go exceptionally helpful for execution of Lung tumour Disease with especially exactness and quick. In this work we proposed a framework called Classification based Lung disease expectation framework. The primary point of this model is to give the prior notice to the clients, and it is likewise cost and time advantage to the client. It predicts three particular malignancy dangers. The investigation has been performed utilizing WEKA tool with a few information mining-grouping strategies and it is discovered that the Back Propagation and Decision tree gives a superior execution in all perspectives over the other arrangement calculations. The framework removes concealed information from a chronicled lung malignancy ailment database. The best model to anticipate patients with Lung growth illness seems, by all accounts, to be Decision Trees and Neural Network. Decision Trees comes about are less demanding to peruse and translate. The bore through element to get to point-by-point patients' profiles is just accessible in Decision Trees. BPNN fared superior to anything Decision Trees as it could recognize all the huge therapeutic indicators. The expectation framework can be additionally improved and can be extended in future for investigation and forecast of different sicknesses. It can likewise be joined with other information mining procedures, e.g., Time Series, Clustering and Association Rules.

References

[1] Ahyaningsih F (2017) A combined strategy for solving quadratic assignment problem. In: Proceedings of AIP conference 1867(1):020006

[2] Alatas B (2011) ACROA: artificial chemical reaction optimization algorithm for global optimization. Expert Systems with Applications 38(10):13170–13180

[3] Ben-David G, Malah D (2005) Bounds on the performance of vector-quantizers under channel errors. IEEE Transactions on Information Theory 51(6):2227–2235

[4] Benjaafar S (2002) Modeling and analysis of congestion in the design of facility layouts. Management Science 48(5):679–704

[5] Benlic U, Hao JK (2013) Breakout local search for the quadratic assignment problem. Applied Mathematics and Computation 219(9):4800–4815

[6] Sang Min Park, Min Kyung Lim, Soon Ae Shin, Young Ho Yun (2006) Impact of prediagnosis smoking, Alcohol, Obesity and Insulin resistance on survival in Male cancer Patients: National Health Insurance corporation study. Journal of clinical Oncology, 24(31):132-140

[7] Hornik K, Stinchcombe M, White H (1990) Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Netw 3:551–560.

[8] Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS (2010) Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Transactions on Information Technology in Biomedicine 14:559–566.

[9] Rao BS, Rao KN, Setty SP (2014) An approach for heart disease detection by enhancing training phase of neural network using hybrid algorithm. 2014 IEEE International Advance Computing Conference, pp. 1211–1220

[10] Subbulakshmi CV, Deepa SN (2015) Medical dataset classification: a machine learning paradigm integrating particle swarm optimization with extreme learning machine classifier. The Scientific World Journal 2015:1–12.

[11] Weng C-H, Huang TC-K, Han R-P (2016) Disease prediction with different types of neural network classifiers. Telematics and Informatics 33:277–292.

[12] Sandeep Kumar Saini, Gaurav, Amita Choudhary (2014) Detection of Lung Carcinoma using fuzzy Logic and ACO Techniques, IJERT, 3(8):903-906

[13] Krishnaiah,V., Narsimha,G., Subhash Chandra,N (2013) Diagnosis of LungCancer Prediction System Using Data Mining Classification Techniques. International Journal of Computer Science and Information Technologies 4 (1)

[14] Rahman RM, Md. Hasan FR (2011) Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data. Expert System with Applications 38:11421–11436.

[15] Kumar, P. M., Lokesh, S., Varatharajan, R., Babu, G. C., & Parthasarathy, P. (2018). Cloud and IoT based disease prediction and diagnosis system for healthcare using Fuzzy neural classifier. Future Generation Computer Systems.

[16] Kanisha, B., Lokesh, S., Kumar, P. M., Parthasarathy, P., & Chandra Babu, G. (2018). Speech recognition with improved support vector machine using dual classifiers and cross fitness validation. Personal and Ubiquitous Computing, 1-9.

[17] J.R. Quinlan (1986) Induction of decision trees. Machine learning, 1(1):81–106

[18] Ramón Díaz-UriarteEmail author and Sara Alvarez de Andrés (2006) Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1):3