Hybrid neuro-fuzzy system for prediction of stages of lung cancer based on the observed symptom values.

Thiyagarajan Manikandan¹, Nallaiyan Bharathi²

¹Department of Electronics and Communication Engineering, Rajalakshmi Engineering College, Chennai, India
²Department of Electrical and Electronics Engineering, Velammal Engineering College, Chennai, India

Abstract

The lung is an essential organ for humans, which often gets affected by diseases. Common lung diseases are asthma, tuberculosis and lung cancer. Among all these diseases, lung cancer is the deadliest one. Treatment of lung cancer depends on its staging. The objective of this work is to design a Hybrid Neuro-Fuzzy System (HNFS) for the prediction of stages of lung cancer, based on the observed symptom values. In order to predict its stages, the cancer assessment specific (stage-wise) questionnaire for lungs (CASQ-L) was prepared. The study was made by asking 167 lung cancer subjects and 50 normal subjects, aged between 37-81 years, to respond to the CASQ-L. The significant symptoms were identified based on Pearson’s correlations performed on all the observed data. Out of 217 subjects, 129 were used to train the system and other 88 were used for testing. The proposed hybrid system has achieved the highest classification accuracy of 97.7% and with the mean accuracy of 96.5% for a fivefold cross-validation analysis. Our findings suggest that the proposed HNFS would be useful for lung cancer stage-wise prediction.

Keywords: Hybrid neuro-fuzzy system (HNFS), Lung cancer, Stage-wise detection, CASQ-L, Pearson’s correlation.

Introduction

Cancer is an abnormal growth of cells in the living organism. Among a variety of cancers in men and women, lung cancer is the foremost cause of cancerous deaths in the world [1]. According to the most recent figures indicated by World Health Organization (WHO), around 7.6 million lung cancer deaths were recorded in worldwide in every year. It is estimated to be around 17 million worldwide deaths in 2030 [2]. Therefore, it is essential to identify the lung cancer and treat it to minimize the risk.

The lung cancer is majorly classified as, Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC). Only about 15% lung cancers are detected with SCLC in humans [3]. However, more common form of lung cancer is the NSCLC, which is divided into adenocarcinoma, Squamous Cell Carcinoma (SCC), and Large Cell Carcinoma (LCC) [4]. Since they behave in a similar way, they are grouped together. Only a few cells are taken for the biopsy test it is not possible to work out, which type of NSCLC lung cancer patients have. This does not make any difference in treatment; therefore, most NSCLC are treated in the same way. The common symptoms of lung cancer includes cough/persistent cough, weight loss, loss of appetite, shortness of breath, pain/persistent pain in the chest, cough up blood/blood in the sputum, fatigue/persistent fatigue and pain in the bone (back/hips)/shoulder/neck, etc. [5].

The stages of NSCLC are assigned a stage from I to IV, according to the seriousness of the disease [6]. The staging of lung cancer refers to the severity (spread) of the disease in to the human body. Depending on the staging symptoms of cancer will differ [7]. Stage I and II are called early stages in which cancer is limited to the lungs. Stage III and IV are called later stages in which cancer has extended outside of the lungs to other organs of the body. Patient’s survival rate with cancer depends on its staging and treatment. Prediction of stages involves assessment of a cancerous tumors volume, as well as its spread in the other organs. Patients with early stages have better chances of survival than its later stages [8]. But overall, patients with lung cancer survived 5 years after the diagnosis is still low [9]. To accurately stage lung cancer the doctors use various tests, which include blood tests, X-rays, CT scans, bone scans, and Positron Emission Tomography (PET) scans. The volume of a tumor, as well as its spread to other part of the body is documented in the radiological procedure [10]. The treatment of particular tumor depends on its staging. The treatment initial stage of cancer involves surgical resection of the part of an affected lobe [11,12]. Treatment of final stage of the disease includes chemotherapy, radiation therapy, surgery, or some combination of these [13].

Several artificial intelligence techniques were successfully adapted in expert systems which are aimed to wide variety of decision making in the areas of medical detection involving
fuzzy systems and neural networks [14]. Fuzzy systems has proved to be the significant tool for developing intelligent decision support systems by utilizing the expert’s information and interpretation in the clinical background [15]. Physician’s style of thinking and computer assisted expert system has been explored in fuzzy approaches for making final decisions on the diseases. Some research groups developed decision support systems or fuzzy expert systems to diagnose the diseases in the medical field by making use of fuzzy set theory. It helped the practitioners and patients to a great extent to diagnosing the diseases. A detailed study on lung cancer diagnosis based on fuzzy rules was conducted by Durai and Iyengar. The efficiency of their system was low, because of their simple algorithm [16]. Later they developed diagnostic model for the stage-wise lung cancer detection using improved fuzzy rules. In addition, it also suggests the type of treatment for the patients [17]. The key characteristic of their later system was easier modification and updating of the database. Their later system efficiency was better than the former, could be used as the medical diagnosis model for finding the stages of lung cancer patients. However there was no evidence of sensitivity, specificity and accuracy of their detection.

A pre-diagnosis model developed by Vishwa et al. for the lung cancer detection made use of artificial neural network (ANN) [18]. Their study concluded that the developed model was designed with a smaller set of data (symptoms) and could be effective for lung cancer diagnosis. Pre-diagnosing system for lung cancer based on supervised learning techniques was developed by Balachandran and Anitha uses ANN model for making decisions [19]. Statistical parameters, like symptoms and risk factors were used in their model and proved that it can be an effective tool for pre-diagnosing lung cancer in comparison with clinical reports. Later they reported in their work that the ANN performance is superior to that of the simple statistical or rule based models [20].

Although the fuzzy system has structural information in the form of if-then rules, it lacks in flexibility with changing exterior environments. On the other hand neural networks are superior in recognizing patterns than fuzzy systems but failed in explaining how they reach their final decisions. To integrate the neural networks adaptability with human like reasoning, neural networks and fuzzy systems are brought together to make hybrid intelligent systems, called HNFS [21]. In an HFNS, neural networks were used to adjust the membership functions of fuzzy system, which are employed to make the final decision. A number of research groups developed neuro-fuzzy models to detect lung cancer, asthma, Parkinson’s disease, heart beat and patients critical conditions in medical applications [22-27].

Subjects and Methods

Subjects and data collection

To detect the lung cancer and its stages from South Indian population we designed CASQ-L, which focuses on patient’s symptoms who suffers with lung cancer [28]. Evolution committee (EC) has been formed to review the CASQ-L. EC consists of seven radiologists, three oncologists, three TB and chest physicians, two general practitioners, one cardiologist and statistical analyst. Three review meetings held months between December 2013 and April 2014, to refine the questionnaire to meet objective of the study. After all the necessary corrections, the questionnaire was approved by EC and made available both in English and participant’s mother tongue Tamil.

Our study is a sincere attempt to initiate a novel approach to stages related to lung cancer within the framework of public health. The patients recruited in the study were those who had lung cancer and visited the Bharat Scan centre, Chennai, India, between the month of June 2014 and December 2014. Written informed consent was collected from all the patients before the start of the study and ethical committee of Bharat scan approved the study protocol. The volunteer group has been appointed to collect the data’s orally through the interview by explaining the questionnaire. In order to make the automated decision support system and also to detect the efficiency (specificity, sensitivity and accuracy) of the lung cancer, normal subjects were also included. Thus, the study population comprising of 217 individuals (24 stage-I, 32 stage-II, 52 stage-III and 59 stage-IV lung cancer and 50 normal subjects) among them 152 males and 65 females that had mean age of 60.27 ± 8.7 years, aged between 37-81 years were recruited by the department of radiology of Bharat scans. The reliability of the questionnaire is verified with Cronbach’s alpha test. The value of the test is calculated as 0.887 for this study. Hence it is well above the standard measure (0.7), the questionnaire is highly reliable.

CASQ-L consists of 46 questions which are divided into three parts, namely, part1, part 2 and part 3, respectively. Part 1 has the patient’s demographic information, part 2 deals with the lung cancer questionnaire (LCQ), and part 3 reveals the lung cancer specific (stage-wise) questionnaire (LCSQ). The patient’s demographic information consists of 21 questions (Q1-Q21), which are of the fill in the blank, objective, and yes or no types. The LCQ consists of 11 common lung cancer symptoms of the yes or no type (Q21-Q32). Finally, LCSQ consists of 14 lung cancer stage-wise symptoms of objective types (Q33-Q46). Each symptom (input field) is quantified with its frequency of occurrence to find severity of the cancer, which is given in the Appendix. The scale varies from 0 to 1, 0-0.25 represents ‘low’, 0.2-0.5 represents ‘medium’, 0.45-0.75 represents ‘high’ and 0.7-1 represents ‘very high’.

Statistical analysis

The significant symptoms on the observed patient’s data were calculated with SPSS (statistical package for social science) software (version 17.0). The Pearson’s correlation was put in use to find the significant symptoms and p<0.01 was considered as statistically significant.
System description

HNFS consists of neural network and fuzzy system. The neural network is employed with a feed forward-back propagation algorithm, to check whether there is a sign of lung cancer or not. The process will be terminated, if the test result is negative. On the other hand, if the test result is positive, then fuzzy system employed to predict the stages of lung cancer patients. The predicted stage may be of stage-I or stage-II or stage-III or stage-IV.

Neural network

Neural network consists of three layers, namely, the input layer, the hidden layer and the output layer. The first (input) layer receives the input signals. The middle (hidden) layer propagates signals from first layer to the third (output) layer. The third (output) layer produces the result of the process. The feed forward-back propagation algorithm is employed in neural network, which adjusts the weights in the each iteration to reduce the error. We used neural network with 8 neurons in the input layer, which are the significant symptoms of LCQ, 3 neurons in the hidden layer which represent the symptoms that mimic lung cancer, and one neuron in the output layer which represents either the presence of lung cancer or its absence. If the patient has a particular symptom, then the input is taken as 1; otherwise as 0.

In neural network, each neuron receives the input from preceding layer and each of those inputs is multiplied by a weight value. The resulted outputs are summed and conceded through a limiting function which produces the preset range of values at the output. Based on the output value obtained at the output layer, network makes the decision as to whether the patient agony from lung cancer or not. If the output value of neural network exceeds threshold value T (0.6 in our case), then network inhibited and display ‘no sign of cancer’. Otherwise, the network excited and display ‘lung cancer is predicted’. Further it adjusts the fuzzy system to check for the stage of the lung cancer. If the result of network does not match with the target value during the training phase, back propagation algorithm employed which goes back in the connection between the present layer and the earlier layer, and reassigns the weights and weighted sum is again computed. Once training got over, network is capable of predicting the test data’s result.

Fuzzy system

Fuzzy system comprises collection of membership functions, logical operators and if-then rules. The functional operations of the fuzzy systems are fuzzification, inferencing, aggregation, and defuzzification. In the fuzzification process, the degrees of exactness for each rule basis can be determined by the membership functions. During the inference, the exactness value for the basis of each rule is calculated, and applied to the termination part of each rule. During aggregation, the outputs of each rule are pooled into a single fuzzy set. Finally, during the defuzzification, the fuzzy output set is converted into a crisp value.

We assigned fuzzy input membership functions as, low (0-0.25), medium (0.2-0.5), high (0.45-0.75) and very high (0.7-1). To predict the stages of lung cancer we framed 297 fuzzy rules in all possible combinations (corresponding to the significant symptoms of each stage) found from LCSQ. The rules were approved by EC. The fuzzy system produced fuzzy output membership function lung cancer stage score as, low (2-3.5), medium (3.25-5), high (4.25-6.5) and very high (5.5-8). Fuzzy centroid method was used to convert the fuzzy output membership function into a crisp value. The final decision on the stages of lung cancer was made by HNFS based on the crisp value.

Results

Significant symptoms

After finalizing the questionnaire by the EC, the data has been collected from 217 abnormal and normal subjects. The significant symptoms were found separately between the LCQ and LCSQ by calculating Pearson’s correlation coefficients and \( p < 0.01 \) was considered as statistically significant. The Pearson’s correlations coefficients and \( p \) values obtained between LCQ are given in Table 1.

<table>
<thead>
<tr>
<th>Q.No</th>
<th>Symptom</th>
<th>Pearson’s coefficient*</th>
<th>( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q22</td>
<td>Cough/persistent cough</td>
<td>0.708**</td>
<td>0</td>
</tr>
<tr>
<td>Q23</td>
<td>Shortness of breath</td>
<td>0.526**</td>
<td>0</td>
</tr>
<tr>
<td>Q24</td>
<td>Loss of appetite</td>
<td>0.516**</td>
<td>0</td>
</tr>
<tr>
<td>Q25</td>
<td>Weight loss</td>
<td>0.502**</td>
<td>0</td>
</tr>
<tr>
<td>Q26</td>
<td>Pain/Persistent chest pain</td>
<td>0.243**</td>
<td>0</td>
</tr>
<tr>
<td>Q27</td>
<td>Difficulty in swallowing</td>
<td>0.113</td>
<td>0.096</td>
</tr>
<tr>
<td>Q28</td>
<td>Cough up blood/ blood in sputum</td>
<td>0.283**</td>
<td>0</td>
</tr>
<tr>
<td>Q29</td>
<td>Hoarseness of voice</td>
<td>0.097</td>
<td>0.153</td>
</tr>
<tr>
<td>Q30</td>
<td>Fatigue /persistent fatigue</td>
<td>0.372**</td>
<td>0</td>
</tr>
<tr>
<td>Q31</td>
<td>wheezing</td>
<td>0.123</td>
<td>0.071</td>
</tr>
<tr>
<td>Q32</td>
<td>Pain in bone/ shoulder/neck/arm</td>
<td>0.223**</td>
<td>0.001</td>
</tr>
</tbody>
</table>

\*\( p < 0.01 \).

It is evident that from Table 1 that, the symptoms of questions Q27, Q29 and Q31 are having \( p \) values greater than 0.01. The remaining symptom of questions Q22-Q26, Q28, Q30 and Q32 has \( p \) value less than 0.01. Since \( p < 0.01 \) was chosen as significant for this study, the symptoms of questions Q22-Q26, Q28, Q30 and Q32, a total of 8 symptoms were considered for the neural network training and testing. Thus, cough/persistent cough, loss of appetite, weight loss, shortness of breath, pain/
persistent chest pain, cough up blood/blood in sputum, fatigue/
persistent fatigue and pain in bone (back/hips)/ shoulder/neck/arm were selected.

Table 2 shows the Pearson’s correlations between the LCSQ. It is apparent from Table 2 that, the symptoms of questions Q33, Q34, Q35, Q36, and Q37, Q38, Q39, Q40, and Q41, has p value greater than 0.01. The remaining symptom of questions Q31-Q32, Q39, Q41, and Q43, has p value less than 0.01. Thus the symptoms of questions Q33-Q37, Q39, Q41, and Q43, a total of 8 symptoms were selected as significant to frame the fuzzy rules. Totally 297 fuzzy rules; 30 for stage-I, 44 for stage-II, 89 for stage-III and 134 for stage-IV, framed in all possible combinations to make the decision about stages of a lung cancer. Thus, cough/persistent cough, loss of appetite, weight loss, shortness of breath, pain/persistent chest pain, cough up blood/blood in sputum, fatigue/persistent fatigue and pain in bone (back/hips)/shoulder/neck/arm were selected.

Table 2. Pearson’s correlation between LCSQ (n=217).

<table>
<thead>
<tr>
<th>Q. No</th>
<th>Symptom</th>
<th>Pearson’s correlation coefficient (r)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q33</td>
<td>Cough/persistent cough</td>
<td>0.550**</td>
<td>0</td>
</tr>
<tr>
<td>Q34</td>
<td>Shortness of breath</td>
<td>0.449**</td>
<td>0</td>
</tr>
<tr>
<td>Q35</td>
<td>Loss of appetite</td>
<td>0.443**</td>
<td>0</td>
</tr>
<tr>
<td>Q36</td>
<td>Weight loss</td>
<td>0.434**</td>
<td>0</td>
</tr>
<tr>
<td>Q37</td>
<td>Pain/Persistent chest pain</td>
<td>0.232**</td>
<td>0.001</td>
</tr>
<tr>
<td>Q38</td>
<td>Difficulty in swallowing</td>
<td>0.109</td>
<td>0.118</td>
</tr>
<tr>
<td>Q39</td>
<td>Cough up blood/blood in sputum</td>
<td>0.263**</td>
<td>0</td>
</tr>
<tr>
<td>Q40</td>
<td>Hoarseness of voice</td>
<td>0.093</td>
<td>0.181</td>
</tr>
<tr>
<td>Q41</td>
<td>Fatigue /persistent fatigue</td>
<td>0.351**</td>
<td>0</td>
</tr>
<tr>
<td>Q42</td>
<td>wheezing</td>
<td>0.117</td>
<td>0.092</td>
</tr>
<tr>
<td>Q43</td>
<td>Pain in bone (back/hips)/ shoulder/neck/arm</td>
<td>0.211**</td>
<td>0.002</td>
</tr>
<tr>
<td>Q44</td>
<td>Frequent/unexplained fever</td>
<td>0.104</td>
<td>0.135</td>
</tr>
<tr>
<td>Q45</td>
<td>Swelling in the face/neck/feet</td>
<td>0.096</td>
<td>0.169</td>
</tr>
<tr>
<td>Q46</td>
<td>Frequent head ache/dizziness/seizures</td>
<td>0.055</td>
<td>0.432</td>
</tr>
</tbody>
</table>

**p <0.01.

Training and testing in neural network

After the identification of the significant symptoms by the SPSS statistical tool, the neural network was trained to check the presence or absence of lung cancer. The inputs to the neural network were the significant symptoms found from LCQ. Training was carried out using the feed forward algorithm. To reduce the error during the training process, the back propagation algorithm and learning were combined. During the testing phase, the significant symptoms of the subjects test data were fed to neural network, checks for the presence or absence of lung cancer.

Decision making in HNFS

Neural network’s output was combined with the fuzzy system to make the final decision on the stage of a lung cancer. If the output value neural network does not exceed the threshold value, it displays ‘no sign of lung cancer’. Otherwise it display ‘lung cancer is predicted’ and adjusts the fuzzy system to produce the crisp value based on the rule match. The HNFS makes final decision based on the crisp value as follows:

If crisp value ≤ 2.75
{}
Stage-I is predicted
{}
Else if crisp vale >2.75 && crisp value ≤ 4.125
{}
Stage-II is predicted
{}
Else if crisp vale >4.125 && crisp value ≤ 5.365
{}
Stage-III is predicted
{}
Else
{}
Stage IV is predicted
{}

Discussion

The 217 subjects (abnormal and normal) data collected from the Bharat Scans, Chennai, were utilized to design a HNFS. It was trained with 60% of the total records, which were about 129 (14 stage-I, 19 stage-II, 31 stage-III, 35 stage-IV and 30 normal). It was tested remaining 40% records, which were about 88 (10 stage-I, 13 stage-II, 21 stage-III and 24 stage-IV and 20 normal). The records were chosen randomly for training and testing of five-fold cross-validation (Table 3). HNFS performance was analyzed using sensitivity, specificity and accuracy along with Positive Prediction Value (PPV) and Negative Prediction Value (NPV) through the following equations [29]:

Sensitivity = TP/(TP+FN) × 100 → (1)
Specificity = TN/(TN +FP) × 100 → (2)
Accuracy = (TP+TN)/(TP+ TN+ FP+FN) × 100 → (3)
PPV = TP/(TP+FP) × 100 → (4)
NPV = TN/(FN+TN) × 100 → (5)
True positive (TP) is defined as the situation where by the subject projected as lung cancer disease when actually at lung cancer risk. True negative (TN) is phenomenon thereby the subject is predicted as healthy person and has healthy in reality. Correspondingly, false positive (FP) is the condition of incorrect lung cancer prediction when subject is healthy in reality. False negative (FN) is the situation of healthy prediction when subject is prone to lung cancer risk in reality.

**Table 3.** Performance of the HNFS for randomly divided data sets (training and independent testing sets) via a stratified 5-fold cross validation.

<table>
<thead>
<tr>
<th>Fold</th>
<th>Size of train data set (n)</th>
<th>Size of test data set (n)</th>
<th>HNFS output</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
<th>PPV (%)</th>
<th>NPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold ≠ 1</td>
<td>129</td>
<td>74 14 88</td>
<td>100 70 93.2</td>
<td>91.9</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fold ≠ 2</td>
<td>129</td>
<td>71 17 88</td>
<td>100 85 96.6</td>
<td>95.8</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fold ≠ 3</td>
<td>129</td>
<td>73 15 88</td>
<td>100 75 94.3</td>
<td>93.2</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fold ≠ 4</td>
<td>129</td>
<td>72 16 88</td>
<td>100 80 95.5</td>
<td>94.4</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fold ≠ 5</td>
<td>129</td>
<td>70 18 88</td>
<td>100 90 97.7</td>
<td>97.1</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>100 80 95.5</td>
<td></td>
<td>94.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

'n' indicates number of subjects.

The HNFS is attempts to screen lung cancer subjects; outstanding to its ability to show 95.5 (accuracy) they can be recognized correctly. Similar study has reported by Oana et al. for Parkinson’s disease detection exhibited 95.46% accuracy [25]. Our study depicted sensitivity of 100%, specificity of 80%, PPV of 94.5% and NPV of 100%. Similarly, Haryo et al. reported that the neuro-fuzzy system for the chest X ray images achieved accuracy of 90% [22]. Classification of lung cancer nodules from CT images using neural fuzzy classifier achieved 95% reported by Tariq et al. [23]. The early diagnosis of critical condition of a patient using FPGA based fuzzy neural network system provided an accuracy of 95.2% reported by Chowdhury and Saha [27]. Thus, the proposed hybrid system performance using neuro-fuzzy model is comparable with that of the existing methods.

**Conclusion**

Major cause of cancerous death in the world is the lung cancer. The commonness of lung cancer disease is high in India, especially in rural areas, but did not get noticed by the people because of lack of awareness about its symptoms. It is also not possible by the volunteer agencies to carry the screening for all the people in the rural areas. This work attempted to design a screening tool, using HNFS, which could help both patients and doctors to predict the lung cancer and its stages. The sensitivity, specificity, accuracy, PPV and NPV of the proposed HNFS using neuro-fuzzy model are calculated as 100%, 80%, 95.5% and 94.5%, 100%, respectively. These experimental results indicate HNFS performance is satisfactory and may useful for lung cancer stage-wise detection.

**Acknowledgement**

The study was conducted at Bharat Scans, Royapettah, Chennai, following approval (Ref:IEC-BERF/Approval Lr./Date: 4-6-2014) by the ethical committee of the Bharat Education and Research Foundation. The authors wish to express their gratitude to the authorities of Bharat Scans for providing required facilitative infrastructure and invaluable help during oral interactions with patients during the interviews.

**References**


*Correspondence to*  
Thiyagarajan Manikandan  
Associate Professor  
Department of Electronics & Communication Engineering  
Rajalakshmi Engineering College  
Chennai-602106  
India