CHAPTER 3

HYBRID NEURO-FUZZY SYSTEM FOR PREDICTION OF LUNG CANCER AND ITS STAGES BASED ON THE OBSERVED SYMPTOMS

3.1 INTRODUCTION

Lung cancer is the major cause of cancerous deaths in the United States and worldwide (American cancer society, 2014). In India, it is the foremost cause of deaths in men; and ranked ninth in all cancer related deaths reported in women (Behera and Balamugesh, 2004). Two main types of lung cancer are Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). About 15 out of 100 (15%) lung cancers are diagnosed with SCLC. It is aggressive, that is, it often grows rapidly and spreads to other regions, including lymph nodes, bone, brain, adrenal glands, and liver. The risk of developing SCLC is highly associated with cigarette smoking. Less than 5% of the patients diagnosed with the disease never smoked (Jackman and Johnson, 2005). SCLC is so called, because when viewed through the microscope the cancer cells look small.

The more common form of lung cancer is the NSCLC, namely, Squamous Cell Carcinoma (SCC), adenocarcinoma and Large Cell Carcinoma (LCC) (Gomathi and Thangaraj, 2011). SCC develops in the central region of the lung and accounts for approximately 25-30% of all lung cancer cases. Adenocarcinoma accounts for approximately 40% of all lung cancer cases and develops in the outer region of the lung. LCC accounts for
approximately 10-15% of all lung cancer cases, and is associated with rapid tumor growth (American Cancer Society, 2010). They are grouped together, because they behave in a similar way. It is not possible to work out, which type of NSCLC lung cancer patients have, because only a few cells are taken for the biopsy test. This does not make any difference in treatment; therefore, most NSCLC are treated in the same way. In spite of decades of research on diagnosis and therapy, the survival of lung cancer patients remains poor, which is reported about 13-15% at five-year survival (Kuruvilla and Gunavathi, 2014).

The stages of NSCLC are assigned a stage from I to IV, according to the seriousness of the disease (Consonni et al. 2015). 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 differ (Cancer Facts and Figures, 2013). Stage I and II are called early stage in which cancer is limited to the lungs. Stage III and IV are called later stage in which cancer extends to other organs of the body. Patient’s survival rate with cancer depends on its staging and treatment. In order to identify the lung cancer in its early stage (through its symptoms), generally questionnaires are prepared to use in the preclinical tools and data are collected by asking patients to respond to the questionnaire.

This chapter is organized as follows. Section 3.1 gives review of questionnaires available for lung cancer detection. Section 3.2 deals about the formation of Evaluation Committee (EC). Section 3.3 describes on Cancer Assessment Stage-wise Questionnaire for Lung (CASQ-L). Section 3.4 elaborates on subjects and methods and section 3.5 gives experimental results and discussions. Finally, section 3.6 brings the summary of this chapter.
3.2 REVIEW OF QUESTIONNAIRES AVAILABLE FOR LUNG CANCER DETECTION

Only limited screening tools are available, to diagnose lung cancer at an early stage from patients, by making use of questionnaires. The lung cancer screening tool developed by the American Lung Association makes use of the age, smoking history and number of packs smoked per year for predicting the lung cancer (American Lung Association 2010). The lung cancer screening decision tool available at the Memorial Sloan Kettering Cancer Center helps the clinicians and patients to determine the chance that screening will be beneficial based on the patient’s age, smoking history and environmental exposure (Memorial Sloan Kettering Cancer Center 2013). The lung cancer risk profiler developed by Saint Peter’s University hospital predicts the risk of cancer, based on the smoking history, and other lung cancer risk factors such as diagnosed Chronic Obstructive Pulmonary Disease (COPD), diagnosed pulmonary fibrosis, exposure to asbestos, radon gas and dust, etc. (Memorial Sloan Kettering Cancer Center 2013). David Cella developed subsets of questionnaires before making the final version of his Functional Assessment Cancer Therapy for Lungs (FACT-L). It consists of 36 questions grouped into physical well-being, social/family well-being, emotional well-being, and functional well-being and additional concerns for assessing the patients with lung cancer (David Cella 1999). Ganti and Mulshine (2006) demonstrated that, the lung screening tool based on the Computer-aided Diagnosis Systems (CAD) system could be significant for early lung cancer detection. Members of the national lung screening trial team reported that their implemented preclinical screening tool could reduce deaths among the lung cancer patients (Denise et al. 2011).

The limitations of the existing screening tools used in hospitals are that, they were designed using only a few possible lung cancer related
questions. To make the screening tool very effective, it is necessary to combine all possible lung cancer related questions along with their symptoms. The questionnaires reported in the literatures were developed by considering the living condition and life style of American and Spanish populations. No specific questionnaire is available for detecting lung cancer and its stages from the south Indian population. Therefore, the present study is aimed to design questionnaire for lung cancer detection and its stages of the patients and to use in the preclinical screening tool to meet the objectives of the study.

3.3 EVALUATION COMMITTEE

Questionnaire was prepared based on the pilot study which focuses on patient’s symptoms, who suffered with lung cancer (Manikandan et al. 2014). Based on patient’s inputs the questionnaires were reorganised to meet the objectives of the study. An Evaluation Committee (EC) has been formed to review the questionnaires. Research supervisor and DC members authorized and authenticated to constitute the EC in connection with this research. It consists of three oncologists, seven radiologists, three TB and chest physicians, two general practitioners, one cardiologist and statistical analyst.

Three review meetings were held periodically, before the approval of the final questionnaires. The first review meeting was held in December 2013. The second and third meetings were conducted during the months of February 2014 and April 2014, respectively, to refine the questionnaires to meet the objectives of the study. After the approval by the EC, the questionnaires were approved by the institutional ethics committee of Bharat Education and Research Foundation (Academic wing of Bharat Scans), Royapettah, Chennai, affiliated to the Tamil Nadu Dr. M.G.R Medical University, Chennai in May 2014 (Appendix 1). The questionnaire was made
available both in English and participant’s mother tongue Tamil before the start of the study.

3.4 CANCER ASSESSMENT STAGE-WISE QUESTIONNAIRE FOR LUNG

CASQ-L was prepared, to develop a preclinical screening tool to predict the lung cancer and its stages based on the observed symptoms. The CASQ-L is given in Appendix 2. The consent letter to be obtained from the patients during the data collection is given in Appendix 3.

CASQ-L consists of 46 questions which are divided into three parts, namely part 1, 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 \((Q_1-Q_{21})\), which are of the fill in the blanks, objectives, and yes or no types. The LCQ consists of 11 common lung cancer symptoms of the yes or no type \((Q_{22}-Q_{32})\). Finally, LCSQ consists of 14 lung cancer stage-wise symptoms of objective types \((Q_{33}-Q_{46})\). Each symptom (input field) is quantified with its frequency of occurrence to find severity of the cancer. The scale varies from 0 to 1, where 0-0.25 represents ‘low’, 0.2-0.5 represents ‘medium’, 0.45-0.75 represents ‘high’ and 0.7-1 represents ‘very high. The frequency of occurrence (scale range) of each symptom in LCSQ is given in Appendix 4.

3.5 SUBJECTS AND METHODS

3.5.1 Study Population

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 (confirmed through biopsy test) and visited the Bharat Scans, 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 Scans approved the study protocol. The volunteer group has been appointed to collect the data orally through interview by explaining the questionnaire. Before the start of the study, the questionnaire was explained to all the participants and they were given enough time to understand the importance of their contribution in this study.

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 (with lung infections) were also included. Out of the study population of 217 individuals (24 stage-I, 32 stage-II, 52 stage-III and 59 stage-IV lung cancer and 50 normal subjects), 152 were males and 65 were females (Mean±SD age=60.27±8.7 years), aged between 37-81 years. The reliability of the questionnaire is verified with Cronbach’s Alpha test. The value of the test is calculated as 0.887. Hence it is well above the standard measure (0.7), the questionnaire is highly reliable.

### 3.5.2 Hybrid Neuro-Fuzzy System

Several artificial intelligence techniques were successfully adapted in hybrid systems which are aimed for a wide variety of decision making in the areas of medical detection involving fuzzy system and Artificial Neural Networks (ANNs) (Gliwa and Byrski 2011). Fuzzy system 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 (Kannan and Ramat 2008). Physician’s style of thinking and computer assisted expert system have 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 diagnose the diseases.

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 ANNs 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, ANNs and fuzzy systems are brought together to make hybrid intelligent systems, called Hybrid Neuro-Fuzzy Systems (HNFS). In an HFNS, ANNs 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 (Haryo et al. 2011; Tariq et al. 2013; Patra and Thakur, 2013; Oana et al. 2013; Braojos et al. 2014).

### 3.5.3 Statistical Analysis

The significant symptoms on the observed patients’ data were calculated with Statistical Package for Social Sciences (SPSS) software (version 17.0). The p-value of the Pearson’s correlation test was put in to find the significant symptoms and p<0.01 was considered statistically highly significant. The p-value is the probability value, which determines the statistical significance between the variables. The p value is chosen as statistically significant if p<0.05 (i.e., 95% confidence level) and statistically highly significant if p<0.01 (i.e., 99% confidence level).
3.5.4 System Description

The flow diagram of HNFS is shown in Figure 3.1, which consists of ANN and fuzzy system or Fuzzy Inference System (FIS). The ANN 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 is 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.

![Flow diagram of the HNFS system](image)

**Figure 3.1 Flow diagram of the HNFS system**
3.5.4.1 Artificial neural network

ANN 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 architecture of the feed forward-back propagation network is shown in Figure 3.2, which adjusts the weights in each iteration to reduce the error. We used ANN 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.

![Figure 3.2 Architecture of feed forward-back propagation algorithm](image_url)

In ANN, each neuron receives the input from the 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 agonises from lung cancer or not. If the output value of neural network exceeds threshold value T, then network is inhibited and display ‘no sign of cancer’. Otherwise, the ANN is excited and displays ‘lung cancer is predicted’. Further it adjusts the fuzzy system to check the stage of the lung cancer. If the result of network does not match with the target value during the training phase, the employed back propagation algorithm goes back in the connection between the present layer and the earlier layer, and reassigns the weights and weighted sum which is again computed.

The steps involved in the neural networks training phase is given as follows:

Step 1: Create a neural network with 8 neurons in input layer, 3 neurons in hidden layer and one neuron in output layer

Step 2: Initialize the weights and bias to the small random values.

Step 3: While stopping condition is false do the steps 4-10.

Step 4: For each training sample $x_i : t_k$, i.e. input the input vectors and target outputs and do the steps 5-11.

/*Feed forward phase*/

Step 5: Apply vector $x_i$ (significant symptoms from LCQ) to the input layer, where

$$x_i = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\} \quad (3.1)$$
Step 6: Compute the input \((z_{inj})\) and output \((z_o)\) for each hidden unit \(z_j\). The input to the each hidden layer is given by,

\[ z_{inj} = v_o + \sum x_i \cdot v_{ij} \]  \hspace{1cm} (3.2)

where \(v_o\) is the bias value on the hidden layer and \(v_{ij}\) is the weights between input and hidden layer. Therefore output from each hidden layer is calculated by,

\[ z_o = f (z_{inj}) \]  \hspace{1cm} (3.3)

The hidden layer uses sigmoid activation function to calculate its output value which is given by,

\[ z_j = \frac{1}{1+e^{-z_{inj}}} \]  \hspace{1cm} (3.4)

Step 7: For the output layer compute the input \((y_{ink})\) and output \((y_o)\). The input to the output layer \((y_{ink})\) is calculated as,

\[ y_{ink} = v_o + \sum z_o \cdot w_{jk} \]  \hspace{1cm} (3.5)

where \(v_o\) is the bias value on output layer and \(w_{jk}\) is the weights between hidden layer and output layer. Therefore output from output layer is calculated by,

\[ y_o = f (y_{ink}) \]  \hspace{1cm} (3.6)

The output layer uses linear activation function to calculate its output value \((y_k)\). The output value obtained from the output layer is given by,

\[ y_k = \begin{cases} 1, & \text{if } y_k > T \\ 0, & \text{if } y_k \leq T \end{cases} \]  \hspace{1cm} (3.7)

where \(T\) is a non-negative threshold.
/*Error back propagation phase*/

Step 8: For the output layer \((y_k)\), compute the error term \((\delta_k)\) and weight change \((\Delta w_{jk})\) between output and hidden layer, which are given by,

\[
\delta_k = (t_k - y_k) \cdot f'(y_{ink})
\]  \hspace{1cm} (3.8)

where \(f'(y_{ink})\) is the derivative of \(y_{ink}\) and

\[
\Delta w_{jk} = \alpha \cdot \delta_k \cdot z_o
\]  \hspace{1cm} (3.9)

where \(\alpha\) is called learning rate.

Step 9: For each hidden unit \((z_j)\), compute the error term \((\delta_j)\) and weight change \((\Delta v_{ij})\) between input and hidden layer, which are given by,

\[
\delta_j = \delta_{inj} \cdot f'(z_{inj})
\]  \hspace{1cm} (3.10)

where, \(\delta_{inj} = \sum \delta_k \cdot w_{jk}\)  \hspace{1cm} (3.11)

\(f'(z_{inj})\) is the derivative of \(z_{inj}\) and

\[
\Delta v_{ij} = \alpha \cdot \delta_j \cdot x_i
\]  \hspace{1cm} (3.12)

Step 10: Update the weights between hidden layer and output layer and the input layer and hidden layer as,

\[
w_{jk} + 1 = w_{jk} + \Delta w_{jk} for all j and k
\]  \hspace{1cm} (3.13)

\[
v_{ij} + 1 = v_{ij} + \Delta v_{ij} for all i and j
\]  \hspace{1cm} (3.14)

Step 11: Test the stopping condition, which is given by,

\[
E = \sum (t_k - y_k)^2
\]  \hspace{1cm} (3.15)
The training phase is continued till the network reaches the stopping condition, which is the minimum total mean squared error ($E$) between the neural network output value ($y_k$) and the target value ($t_k$). Once training got over, network is capable of predicting the test data’s result.

The steps involved in neural networks testing phase is given as follows:

Step 1: Feed the significant symptoms from LCQ of the subject’s test data to the neural network.

Step 2: Calculate the net output value ($y_k$) from the output layer.

Step 3: If the net output value is less than the predefined threshold value $T$, then the output neuron inhibited and display ‘no sign of lung cancer’. Else, go to step 4.

Step 4: The output neuron fires and display ‘lung cancer is predicted’. Further, the neural output adjusts the fuzzy system to check for a specific stage of the cancer.

3.5.4.2 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.
To characterize the fuzzy set, we necessitate describing its membership function. It represents how every point in the input space is mapped to a degree of membership between 0 and 1. It is denoted by $\mu$. For all the input symptoms, the membership function is defined as shown below.

For example, in Figure 3.3, considering the symptom persistent cough, input fuzzy membership functions are: Low, medium, high and very high. The triangular function is used to represent the input membership function. The range of values assigned for input membership functions are given in Table 3.1. Assuming a universe of discourse $y$ and fuzzy set $A$ defined on it. We also assume discrete set of $y$ elements. The membership function $\mu_A(y)$ defined for fuzzy set $A$ maps the elements $y_i$ of $y_n$ to the degree of membership in the interval 0 and 1.

$$\mu_A(y) = \text{Degree}(y \in A), 0 \leq \mu_A(y) \leq 1$$  \hspace{1cm} (3.16)

![Figure 3.3](image.png)  

**Figure 3.3** Input fuzzy membership functions for persistent cough (low, medium, high and very high)
Table 3.1 A range of values assigned for fuzzy sets of persistent cough

<table>
<thead>
<tr>
<th>Input symptom</th>
<th>Range</th>
<th>Fuzzy sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistent cough</td>
<td>0-0.25</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>0.2-0.5</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>0.45-0.75</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>0.7-1</td>
<td>Very high</td>
</tr>
</tbody>
</table>

Fuzzy set for a discrete set of elements represented as vector:

\[ A = (A_1, A_2, \ldots, A_n), \quad A_i = \mu_A(y_i) \]  \hfill (3.17)

For every input variable the membership function is assigned as follows. For example, the input membership functions for the symptom persistent cough are:

\[ \mu_{A\text{low}}(y) = \begin{cases} 
\frac{y}{0.125}, & 0 \leq y < 0.125 \\
1, & y = 0.125 \\
\frac{0.25-y}{0.125}, & 0.125 < y \leq 0.25 
\end{cases} \]  \hfill (3.18)

\[ \mu_{A\text{medium}}(y) = \begin{cases} 
\frac{y-0.2}{0.15}, & 0.2 \leq y < 0.35 \\
1, & y = 0.35 \\
\frac{0.5-y}{0.15}, & 0.35 < y \leq 0.5 
\end{cases} \]  \hfill (3.19)

\[ \mu_{A\text{high}}(y) = \begin{cases} 
\frac{y-0.45}{0.15}, & 0.45 \leq y < 0.6 \\
1, & y = 0.6 \\
\frac{0.75-y}{0.15}, & 0.6 < y \leq 0.75 
\end{cases} \]  \hfill (3.20)

\[ \mu_{A\text{veryhigh}}(y) = \begin{cases} 
\frac{y-0.7}{0.15}, & 0.7 \leq y < 0.85 \\
1, & y = 0.85 \\
\frac{1-y}{0.15}, & 0.85 < y \leq 1 
\end{cases} \]  \hfill (3.21)
Similarly, the membership functions were assigned for the remaining nine significant symptoms. The input membership functions are mapped to an output membership functions based on the rule match. In Figure 3.4, it is evident that, output fuzzy membership function for lung cancer score has the range, low, medium, high and very high.

![Output fuzzy membership function for lung cancer score](image)

**Figure 3.4** Output fuzzy membership function for lung cancer score (low, medium, high, very high)

The output fuzzy membership functions are defuzzified by fuzzy centroid method to produce the crisp value, which is calculated by the formula (Oana et al. 2013):

\[
\text{Crisp}(\text{medium}) = \frac{\int y \mu_{A\text{medium}}(y) \, dy}{\int \mu_{A\text{medium}}(y) \, dy}
\]  

(3.22)

### 3.6 RESULTS AND DISCUSSION

#### 3.6.1 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 using SPSS software separately between the LCQ and LCSQ by calculating Pearson’s correlation coefficients and p<0.01 was considered statistically significant. Since the researcher is considered to have 99% confidence level p<0.01 chosen as significant for this study (Richard and Rubin, 2009). The Pearson’s correlations coefficients and p values computed using SPSS software between LCQ is given in Table 3.2. Table 3.3 shows the Pearson’s correlations coefficients and p values between the LCSQ are computed using SPSS software.

### Table 3.2 Pearson’s correlation between LCQ (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>Q22</td>
<td>Persistent cough</td>
<td>0.708**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q23</td>
<td>Shortness of breath</td>
<td>0.526**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q24</td>
<td>Loss of appetite</td>
<td>0.516**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q25</td>
<td>Weight loss</td>
<td>0.502**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q26</td>
<td>Persistent chest pain</td>
<td>0.243**</td>
<td>0.000</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>0.000</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>Persistent fatigue</td>
<td>0.372**</td>
<td>0.000</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 (back/hips)/shoulder/neck/arm</td>
<td>0.223**</td>
<td>0.001</td>
</tr>
</tbody>
</table>

It is evident that from Table 3.2 that, the symptoms of questions Q27, Q29 and Q31 are having p values greater than 0.01. The remaining symptoms of questions Q22-Q26, Q28, Q30 and Q32 have 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 ANN training and testing. Thus, persistent cough, loss of appetite, weight loss, shortness of breath, persistent chest pain, cough up blood/blood in sputum, persistent fatigue and pain in bone (back/hips)/shoulder/neck/arm were selected.

### Table 3.3 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>Persistent cough</td>
<td>0.550**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q34</td>
<td>Shortness of breath</td>
<td>0.449**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q35</td>
<td>Loss of appetite</td>
<td>0.443**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q36</td>
<td>Weight loss</td>
<td>0.434**</td>
<td>0.000</td>
</tr>
<tr>
<td>Q37</td>
<td>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>0.000</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>Persistent fatigue</td>
<td>0.351**</td>
<td>0.000</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 headache/dizziness/seizures</td>
<td>0.055</td>
<td>0.432</td>
</tr>
</tbody>
</table>
It is apparent from Table 3.3 that, the symptoms of questions Q38, Q40, Q42, and Q44-46 have p value greater than 0.01. The remaining symptom of questions Q33-Q37, Q39, Q41, and Q43, have 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. Thus, persistent cough, shortness of breath, loss of appetite, weight loss, persistent chest pain, cough up blood/blood in sputum, persistent fatigue and pain in bone (back/hips)/shoulder/neck/arm were selected. 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 lung cancer, which are given in Appendix 5.

3.6.2 Training and Testing in ANN

After the identification of the significant symptoms by the SPSS statistical tool, the ANN was trained to check the presence or absence of lung cancer. The inputs to the ANN were the significant symptoms found from LCQ. During the training phase, the input and desired output values are presented to the ANN in the HNFS. If the result of the network does not match with the desired output, then the error (which is the difference between the input value and desired output value) is computed. The error is propagated backwards using back propagation algorithm (gradient descent) from the output layer to the input layer. The weights between the input and output layers are adjusted till it reaches a minimum value. During the testing phase, the significant symptoms of the subjects test data were fed to ANN, checks for the presence or absence of lung cancer.

3.6.3 Decision Making in HNFS

ANNs output was combined with the fuzzy system in HNFS to make the final decision during the stages of a lung cancer. If the output value
of the ANN does not exceed the threshold value, it displays ‘lung cancer is not predicted’. 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 $\leq 2.75$
{
    Stage-I is predicted
}
Else if crisp value $> 2.75$ and $\leq 4.125$
{
    Stage-II is predicted
}
Else if crisp value $> 4.125$ and $\leq 5.365$
{
    Stage-III is predicted
}
Else
{
    Stage IV is predicted
}

Thus, the HNFS system with a heterogeneous structure was developed for predicting the lung cancer from the patients’ symptom values. In this model, the neural network and fuzzy system work as independent components. When a new case is presented to the diagnostic system, the trained neural network determines inputs to the fuzzy system. Then the fuzzy system using predefined fuzzy sets and fuzzy rules, maps the given inputs to an output, and thereby obtains the stages of lung cancer.
Totally, 217 subjects (167 cancerous and 50 non-cancerous) those who visited the Scan center between June 2014 and December 2014 were analysed using HNFS. It was trained with 60% of the total records, which were about 100 cancerous subjects (14 stage-I, 19 stage-II, 32 stage-III, 35 stage-IV) and 30 non-cancerous subjects. It was tested remaining 40% records, which were about 67 cancerous subjects (10 stage-I, 12 stage-II, 21 stage-III and 24 stage-IV) and 20 non-cancerous subjects. HNFS implemented here classified all the 67 cancerous patients as cancerous, therefore True Positive (TP) is 67 and False Negative (FN) is zero. Out of 20 non-cancerous subjects the developed system correctly classified 16 as non-cancerous subjects and 4 as cancerous subjects. Therefore, the True Negative (TN) is 16 and False Positive (FP) is 4. The implemented HNFS performance was analysed using sensitivity, specificity, accuracy, precision, recall and F-measure through the following equations (Kuruvilla and Gunavathi, 2015):

\[
\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{3.23}
\]

\[
\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{3.24}
\]

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \tag{3.25}
\]

\[
\text{Precision(P)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{3.26}
\]

\[
\text{Recall(R)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{3.27}
\]

\[
\text{F-measure} = \frac{2 \times \text{P} \times \text{R}}{(\text{P} + \text{R})} \times 100 \tag{3.28}
\]
TP is defined as the situation where by the subject projected as lung cancer disease when actually at lung cancer risk. TN is phenomenon thereby the subject is predicted as healthy person and has healthy in reality. Correspondingly, FP is the condition of incorrect lung cancer prediction when subject is healthy in reality. FN is the situation of healthy prediction when subject is prone to lung cancer risk in reality.

The sensitivity (recall), specificity, accuracy, precision and F-measure of the proposed HNFS using neuro-fuzzy model were calculated as 100%, 80%, 95%, 94% and 97%, respectively. After the implementation of HNFS with the trained data set, the system performance was validated with 20 subjects (10 cancerous and 10 normal) by the radiologist, which yielded a sensitivity, specificity and accuracy of 100%, 90% and 95% respectively (Appendix 6). Thus, the experimental result indicated that the performance of implemented HNFS was satisfactory and could be used as a pre-clinical screening tool for lung cancer detection. Similar study has reported by Durai et al. (2011) for the lung cancer detection based on symptom values using fuzzy rule based inference system yielded a sensitivity and specificity of 89% and 78% respectively. Similarly, Balachandran and Anitha (2014) reported that lung cancer detection using ANN produced a sensitivity of 90%. Pre-diagnosis of lung cancer using feed forward neural network and back propagation algorithm described by Abhinav et al. (2011) provided a sensitivity, specificity and accuracy of 93%, 78% and 90% respectively.

3.7 SUMMARY

This chapter elaborated a new CAD system for lung cancer and its stage-wise detection based on the observed symptoms of patients using HNFS. First, CASQ-L was prepared. It has demographic information, LCQ and LCSQ. The significant symptoms between LCQ and LCSQ found separately using SPSS software by computing Pearson’s correlation and
p<0.01 was chosen as statistically significant. The significant symptoms of LCQ were used to design the ANN of HNFS and significant symptoms of LCSQ were utilized to frame the fuzzy rules for fuzzy system of HNFS. During the testing process, ANN was employed to check the presence of lung cancer and fuzzy system was employed to check the stage of a predicted cancer based on the training given during the training process. Totally, 297 fuzzy rules were framed in all possible combinations to detect the stage of a lung cancer patient. The proposed HNFS using neuro-fuzzy model produced a sensitivity, specificity, accuracy, precision, recall and F-measure of 100%, 80%, 95%, 100%, 94% and 97%, respectively. The following chapter clearly analyses the segmentation of suspected lung nodules from CT scan images of patients by Auto-seed Region Growing with Morphological Masking (ARGMM).