Risk Assessment of the Top Five Malignancies among Males and Females with Respect to Occupation, Educational Status and Smoking Habits

Ruhul Amin Dicken, S.A.M Fazle Rubby, Sheefta Naz, A. M. Arefin Khaled, Ashraful Azad, Rashedur M Rahman
Department of Electrical and Computer Engineering, North South University,
Plot-15, Block-B, Bashundhara, Dhaka 1229, Bangladesh
ruhul.amin1125@gmail.com, fazle2712@yahoo.com, sheefta@hotmail.com, akonkshaan@gmail.com,
aumit_600mph@gmail.com, rashedur.rahman@northsouth.edu

Abstract: **The diagnosis of the different types of cancers has greatly increased among the population of Bangladesh over the last decade. From 2008 – 2010 the National Institute of Cancer Research and Hospital (NICRH) has confirmed 27281 cases of cancers among male and female patients. This research paper works to find the relationship of the diagnosis of such cancer patient in accordance with their occupation, educational status and tobacco smoking habits. Data was collected from the Cancer registry Report of 2008-2010 published by the Department of Epidemiology at NICRH. Using the dataset two ANFIS models were generated to evaluate the risk of the top five malignancies among male and female population of Bangladesh.**

Keywords: **Adaptive Neuro Fuzzy Inference System (ANFIS); Fuzzy Inference System (FIS); malignancies; ICD-O; cancer.**

I. INTRODUCTION

Cancer is a dangerous disease which is caused due to abnormal growth of the cell in different parts of the body. It occurs due to the abnormal and infectious growth in the cells causing the body to get abnormally infectious and leading to tremendous pain. The cancer is first developed from the tumor in the body. Tumors are the infectious cell in the body, which grows abnormally big and infectious, and later with rapid growth turns cancerous [1][3]. The triggers for these cancerous cells can be of different kinds. Tobacco chewing is one of the causes of cancer deaths which is about 22 %. Also the scarcity of physical activity and the intake of alcohol and other liquid drugs are also major reasons for cancer. Some other causes of cancer are due to certain infections from diseases or cuts, exposure to chemicals, exposure to radiation which ionizes, and environmental pollutants which are harmful for the body.

In developing countries around the world, 20% of cancers cases are caused by infections which can be named as Hepatitis-B, Hepatitis-C, and HPV. Such factors play a part, at which partially changes the genes of definite cell. Such genetic change must occur before cancer develops in the human body [8]. Around 5–10% of cancer cases are from genetic defects which are heredity or inherited from the person's fore-fathers.

This paper considers the condition of the developing countries of the world. It has been seen for the past few years cancer has no curable treatment. The treatment that is given to the cancer patient is to lessen the cancerous infectious cells down to such a point to turn it dormant for a limited time. [9] Prevention is more favorable for such incurable diseases. The first step is to classify the risk of cancer.

This paper takes into consideration the occupation, education and smoking habits of a person and provides a risk assessment for the top five malignancies for male and female patients of Bangladesh. Rest of the paper is organized as follows. Section 2 briefly presents related works. Section 3 describes the data sources used in this paper. Then design of the ANFIS is detailed in Section 4. Section 5 analyzes results. Finally, Section 6 concludes the paper.

II. RELATED WORKS

Each type of cancer shows abnormal growth in different parts of the body. The tumor growth can be enormous or small. Although tumors do not need to be really big to be cancerous, they need to be infected [5][6]. Each type of cancer depends on the stage of the tumor and the place where the abnormal cell growth occurred. This makes that part of the body at more risk. [7].

A miner can get multi- radiation exposure in the mines if the mine is Sulphur exposed. Also different kind of coal residue can form black elemental dust to store and infect the lungs or stomach due to tar or gun powder for the blasts used to mine [8]. Sometimes exposure to pesticides and other manure which are artificially chemical mixture can cause skin cancer.

The type of cancer and the causes can be slotted to different reasons. The lung cancer is more occurring in people with smoking habits. Chewing tobacco causes more people to have buccal mucosa cancer or at least tumor occurrence. In Bangladesh buccal mucosa, lung, liver, blood, kidney carries cancerous cell increase due to farm pesticides. Similarly, nuclear exposure or mechanical workloads gives radiation exposure to the workers without enough protection against the exposure. Along with smoking habits another important issue is the literacy rate [1]. Illiteracy rate leads to increase of cancer cases in the region.

Some doctors deduce that it requires one small exposure to make the dormant cancerous cell active. However, others ignore this fact saying that it is a slow poison like disease, and patient who is more prone to the disease are due to their habits and less knowledge about the risk factor of the disease [2][4].

III. DATA SOURCES

Patients' disease history is collected and recorded in cancer registries. A system was first introduced to collect this information in Bangladesh by the Department of Cancer Epidemiology of the National Institute of Cancer Research and Hospital (NICRH) [10] in

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

2004 with the support of the World Health Organization. In search for the data, to develop a model for the research, the Department of Cancer Epidemiology at NICRH was approached. Data was then collected from the Cancer Registry Report 2008 -2010 which was published by the department in 2013 with courtesy of Square Pharmaceuticals LTD. The report contained information on the demography of the 27281 cancer patients diagnosed at NICRH for the period of 2008-2010 including statistics of the type of cancer, the top diagnosed malignancies, the smoking habits of the patients, etc. From the report we are interested in top five malignancies among male and female patients, the occupation, educational status and smoking habits of the patients with demographic information.

3.1 The Top Five Malignancies:

The International Classification for Diseases for Oncology (ICD-O) are specific codes that are used to topography classify the type of cancer based on the type of the tumor and have been used in cancer registries for the last 35 years. There are a total of 80 classifications denoted with a 'C' and a number, such as C34 for Lungs, each with further subdivisions with respect to the corresponding organ. From the report we decided to focus on the risk assessment of the top five malignancies among males and females of Bangladesh as listed in Table 1.

Rank for the period 2008-2010	MALE (Cancer Site – ICD-O)	FEMALE (Cancer Site – ICD-O)
1	Lung – C34	Breast – C50
2	Esophagus – C15	Cervix – C53
3	Stomach – C16	Lung – C34
4	Liver – C22	Ovary – C56
5	Larynx – C32	Esophagus – C15

Table 1 Top Five Malignancies among male and female patients

3.2 Occupation, Educational Status and Smoking Habits:

As mentioned earlier, from the demographic part of the cancer registry report we decided to focus on the patient's occupation, educational status and smoking habits. The statistics of the registry included these attributes with their numbers separated by genders and contained the categories as shown in figure 1.

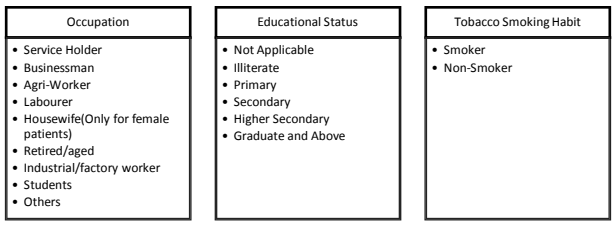


Figure 1 Demographic Attributes of the cancer patients

The reason to focus on these aspects of the cancer patients was to determine the risk of the malignancies described in Table 1 with respect to the patient's daily lifestyle.

Focusing on a patient's occupation gives us an insight to environment the person was exposed to before the diagnosis. A housewife will spend much of her time in a cleaner environment compared to the polluted environment of a female labor at a construction site. A factory worker is more likely to come into contact with carcinogens on a daily basis than a businessman.

The educational status of a person relates to one's health awareness in their everyday lives. A person who has a higher education background, such as a college or university graduate, is likely to have more knowledge on the suspected dangers of cancer risk.

He/she is therefore prone to avoid such dangers and will lead a healthier lifestyle. An illiterate person lacks in understanding of a healthy lifestyle and harmful environment. A very good example of this can be presented by the daily practice of some villagers of rural areas to use river water in their cooking. Coming from a lower education background they know little of the harmful substance or pollution of the water, therefore, leaving them more exposed to the dangers on cancer causing carcinogens.

It was also important to include the smoking habits of these patients as tobacco smoking is a widely known cause for some of the top five selected malignancies.

IV. ANFIS

The variables to be used were decided in the next step. The idea is to create an Adaptive Neuro Fuzzy Inference System (ANFIS) that would estimate the risk of attaining any of the selected cancers based on the person's occupation, educational status and smoking habit. Since ANFIS incorporates both neural network and fuzzy logic principles, it is best suited as the choice of model due to its flexible structure, its compatibility to find solutions for insufficiently defined problems and its ability to manage the obscurity of some attributes in a dataset.

4.1 Designing the Training Dataset

The software used to create the ANFIS was MATLAB which requires a training dataset to generate the fuzzy inference system which will then be used to train the model. The training dataset had to be designed in the form of a table where each column is considered as an input variable and the last column is considered as an output variable. The table for this input is shown in Table 2.

ICD-O	Frequency	Smoking Habit	Occupation	Educational Status	Risk of being diagnosed
C34	4505	Smoker	Agri-worker	Illiterate	0.02168
C15	1002	Non-Smoker	Businessman	Graduate	0.00020
C34	4505	Non-Smoker	Service Holder	Graduate	0.00056
C22	852	Smoker	Retired	Higher Secondary	0.00032

Table 2 Sample of the dataset used to train the ANFIS

The table depicts two inputs as a sample of the complete dataset. As can be seen with the first two columns list the ICD-O and the frequency, i.e. the number of times that ICD-O has been diagnosed at NICRH. The next three columns list the demographic attributes and should be considered as independent events. The output column titled as Risk of being diagnosed is a probability calculation of the combination of events for each set of inputs. In order to provide a better explanation of the process, consider the first input from Table 2 as an example.

Here we need to measure the Risk of being diagnosed with C34 where the patient is an illiterate agri-worker who smokes tobacco. Therefore,

p(Risk of an illiterate agri-worker who smokes tobacco being diagnosed with C34) =

*p(Patient is diagnosed with C34) * p(Patient is a Smoker) * p(patient is an agri-worker) * p(Patient is illiterate)*

Thus the measured probability of these set of independent events is 0.02168, which is set at the output in the dataset table.

4.2 Generating the ANFIS

Two separate training data sets, as the one depicted in table 2, were created for male and female and the probability of the risk of being diagnosed is calculated for all possible combinations of the set of demographic events for all top five ICD-Os. The models were then trained separately for each gender via MATLAB in order to achieve two different risk assessments. ANFIS was used for both male and female patients. The decision for making the models separately for the genders was due to the fact that the top malignancies were different for male and female, as shown in Table 1, as well as the demographic statistics varied vastly for each.

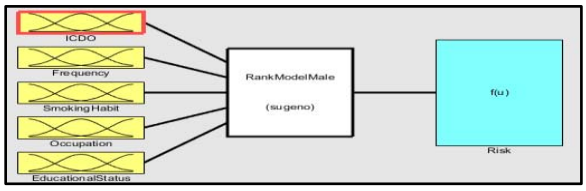


Figure 2 The FIS generated by the ANFIS toolbox

Using the dataset, the ANFIS editor toolbox generates a FIS that will take the ICD-O, frequency, smoking habit, occupation and education as fuzzy variables. The FIS for males is shown in figure 2 and it is similar to the FIS for females. The subcategories for each attribute as shown in Figure 1, is taken as members for the corresponding fuzzy variable. The detailed depiction of these membership functions are shown in Figure 5. The members of most input variables remain the same for the FIS of both male and female, with the exception of the input variables ICD-O and occupation. There exists two membership structures for the input occupation due to the inclusion of the attribute housewife for females as shown in 5e and 5f. The membership function of ICD-O for male and female is shown in 5a and 5b, the structure differs as the top five malignancies for each gender is different in Bangladesh.

Due to consideration of the inputs as a fuzzy variable the model becomes for suited in nature to that of real world situation. For example since all the subcategories of the occupation are fuzzy in nature, for a patient who is a service holder, the model considers his exposure not only to his own environment but to a certain degree his exposure to other environment as well. Similarly the model helps us consider that a non-smoking patient is also exposed to a smoking environment to some degree as a passive smoker. The FIS creates a structure that deduces rules for all possible set of attributes that describes a diagnosed cancer patient and produces a value for the risk of the corresponding ICD-O according to said attribute types as the output.

After all the rules are generated by the FIS, the ANFIS uses the rules to train the model by itself. The value of epochs is set to 3, so that ANFIS would train the model at least three to reduce the average error of the model as low as possible. The trained model outputs the risk of the malignancies as a numeric value for each set of independent events. This output can best be represented by the graphs shown in figure 4 and has been discussed in detail in the following section.

V. RESULTS

While making the data tables, numeric values were set for each of the subcategories of the input variables. Thus the model generated by the ANFIS uses the assigned numeric values to represent each of the attributes. In this section the graphic representation of the risk assessment but the trained ANFIS will be discussed.

5.1 Risk Assessment of the top five malignancies among males.

For male patients the numeric values of the ICD-O are set according to their rank as a malignancy as shown in table 1. Therefore 1 = Lung, 2 = Esophagus, 3 = Stomach, 4 = Liver, 5 = Larynx.

For Occupation: 1 = Service holder, 2 = Businessman, 3 = Agri-worker, 4 = Laborer, 5 = Retired, 6= Industrial/Factory Worker, 7 = Students, 8 = Others.

For Education Status: 1 = Not Applicable, 2 = Illiterate, 3 = Primary, 4 = Secondary, 5 = Higher Secondary, 6 = Graduate or above.

For Smoking Habit: 0 = Non-Smoker, 1 = Smoker.

Risk assessment of ICD-O with respect to occupation: As shown in figure 4a, a significant rise in the risk of the top three malignancies is observed as we move towards the range of 3 to 6 of the occupation axis. These represent the agri-workers, labourers and factory workers. Since these people are more common to work in harmful environments their risk of being diagnosed with the top three malignancies increases as well. Compared to that, the businessman or students are at a much lower risk of these malignancies. The risk of liver cancer also seems to be less affected by the patient's environment.

Risk assessment of ICD-O with respect to Educational Status: The risk of the malignancies can be observed to rise significantly in figure 4b for people of lower educational background and those who are illiterate. As we move along the axis towards higher educational status, the risk also decreases due to a person's increasing knowledge of health hazards.

Risk assessment of ICD-O with respect to Smoking Habit: Smokers are at alarmingly high risk for the top malignancies, except liver cancer. Even for males who do not smoke the value for the risk is pretty high, mostly due to exposure to a smoking environment. This is observed in figure 4c.

Risk assessment with respect to Occupation and Educational Status: Figure 4d depicts the risk of the malignancies with respect to a person's occupation and educational status. The risk is alarmingly high among males who are argi-workers or factory workers with low educational backgrounds.

5.2 Risk Assessment of the top five malignancies among females.

Similar to the previous method the ICD-Os were assigned numeric values according to their rank shown in Table 1. Thus the ICD-Os are 1 = Breast, 2 = Cervix, 3 = Lung, 4 = Ovary, 5 = Esophagus.

For Occupation: 1 = Service holder, 2 = Businesswoman, 3 = Agri-worker, 4 = Laborer, 5 = Housewife, 6 = Retired, 7 = Industrial/Factory Worker, 8 = Students, 9 = Others.

For Education Status and Smoking Habit the assigned values are same as given in section 5.1

Risk assessment of ICD-O with respect to occupation: Housewife surpasses the number of female patients from other professions by a huge margin. Thus in figure 4e the model shows an immense elevation in risk around 5 on the occupation axis which represents female who are housewives. It should also be noted that the risk for females are most for the top two malignancies, i.e. breast cancer and cervix cancer.

Risk assessment of ICD-O with respect to Educational Status: Again the risk for females seem to remain high for the top two malignancies as seen in figure 4f. It seems that it is women who are illiterate or of low educational background are more at risk of being diagnosed than women with higher educational backgraound.

Risk assessment of ICD-O with respect to Smoking Habit: Even though the figure 4g shows the model to higher risk levels for females who are non-smokers, it is also due to the fact that dataset used to train the model contains few number of diagnosed female patients with a history of smoking.

Risk assessment of ICD-O with respect to Occupation and Educational Status: According to figure 4h women who are housewife, labourers or aged and with low education status are at high risk of have been diagnosed. The risk gradually decreases for women of these professions if they belong to a higher level of educational status.

5.3 Performance Evaluation

The ANFIS for both the genders were evaluated. Since the model outputs a value for the risk using FIS generated rules for each set of independent events, all of these values were first listed. To depict the accuracy of the models a test was carried where the average risk is calculated from the ANFIS. This is done for each ICD-O. The result of this is shown in figure 3.

Figure 3 Accuracy Test for each ICD-O

In the figure 3, for each ICD-O, the average risk that was calculated from the actual dataset is compared with the average risk assessed by the ANFIS for the corresponding ICD-O. The blue bars display the actual calculated risk, and the orange ones indicate the estimated risk by the ANFIS. For most of the malignances the estimated risk are similar to the actual risk, the only exception being C53, i.e. cervix cancer, where the ANFIS estimated a higher level of risk for females than the actual risk.

.

VI. CONCLUSION

A lot of obstructions were faced in the process of completing the research. Among those the most troublesome was getting access to informative data and statistics which were resolved due to the presence of the cancer registry report. The ANFIS created in this research paper provides various possibilities of applications if it could be more efficiently implemented. Authorities will now know better in which environments they should focus more if they wish to reduce the percentage of population being effected by cancer. Medical policies and health programs could also be designed based on the risk assessment models of both genders. This initial study will give a direction of the possible areas to work with to reduce the risk of cancer among the people of Bangladesh.

REFERENCES

[1] Ghodke, L., Naik, A., Konale, R., & Mehta, S. (n.d.). Brain Cancer Detection Using Neuro Fuzzy Logic, 58-61. Retrieved from www.interscience.in

[2] Kamath, S. (n.d.). Fuzzy Logic for Breast Cancer Diagnosis Using Medical Thermogram Images. Fuzzy Expert Systems for Disease Diagnosis, 168-199.

[3]Latha K.C., Madhu B., Ayesha S., Ramya. R., Sridhar. R. and Balassumaran. S "Visualization of risk of breast cancer using fuzzy logic in matlab enviornment, International Journal of Computational Intelligence Techniques, vol. 4, no. 1, 2013.

[4] Arita, S., Nomura, T., & Sonoo, H. (n.d.). Diagnostic System of Breast Cancer based on Imaging Data of Mammography using Fuzzy Logic. 2006 World Automation Congress.

[5]E. Al-Daoud, "Cancer Diagnosis Using Modified Fuzzy NetworkUniversal Journal of Computer Science and Engineering Technology, 2010. Retrieved from www.elsevier.com

[6]Laura M. Cecere , Emily C. Williams, Haili Sun, Chris L. Bryson, Brendan J. Clark, Katharine A. Bradley, David H. Au "Smoking cessation and the risk of hospitalization for pneumonia", *Respiratory Medicine*, 2012.

[7]S. Karthikeyeni, S. Ramya, "Comparative analysis of ANFIS and FRBF-survival time prediction of cancer pattern", *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 3, Issue 9, September 2014 Retrieved from www.elsevier.com

[8]Laura M. Cecere , Emily C. Williams, Haili Sun, Chris L. Bryson, Brendan J. Clark, Katharine A. Bradley, David H. Au "Smoking cessation and the risk of hospitalization for pneumonia", *Respiratory Medicine*, Vol. 3, Issue 9, September 2014.

[9]Yilmaz .A and Ayan .K, "Cancer risk analysis by fuzzy logic approach and performance status of the model", *Turkish Journal of Electrical Engineering & Computer Sciences*, 2013. Retrived from www.elsevier.com.

[10] National Institute of Cancer Research and Hospital, http://nicrhbd.org/, retrieved on 8th. January, 2016.

[11] Lefteri H. Tsoukalas, Robert E. Uhrig, Fuzzy and Neural Approaches in Engineering, World Scientific, 1997.

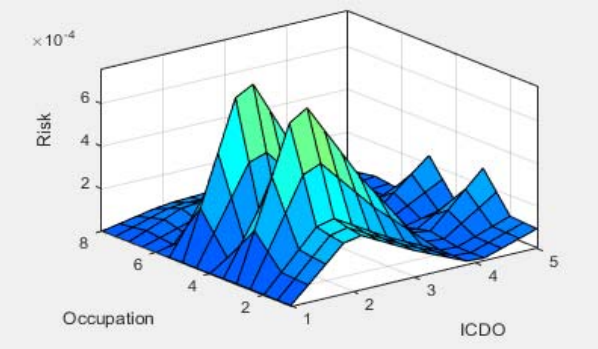
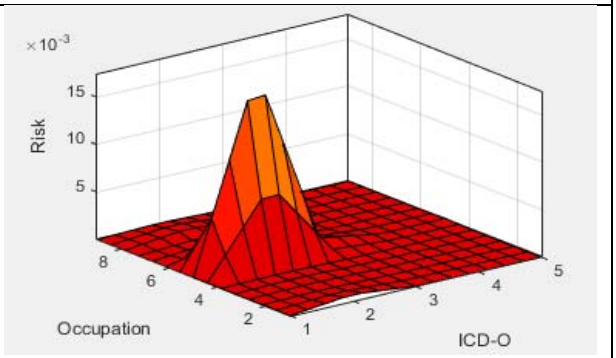
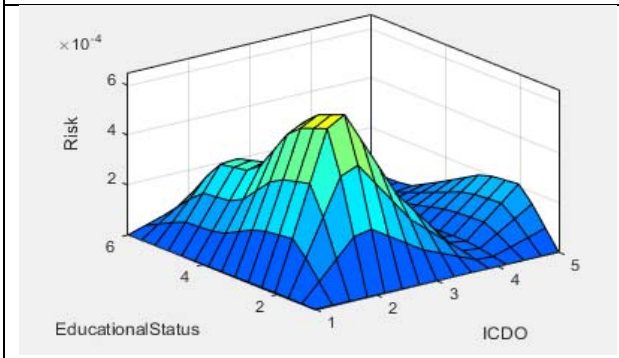
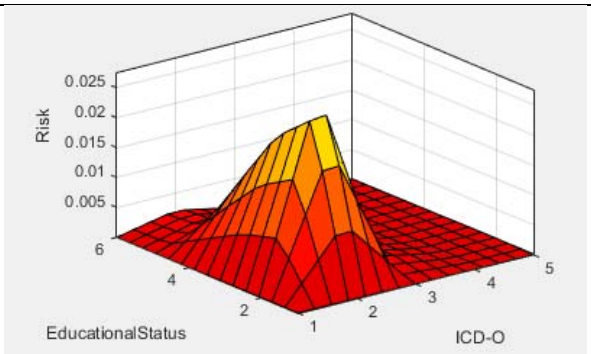
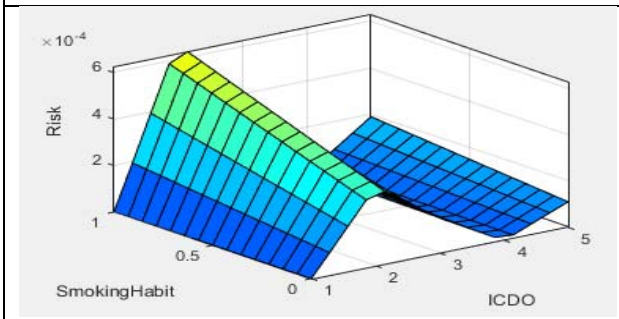
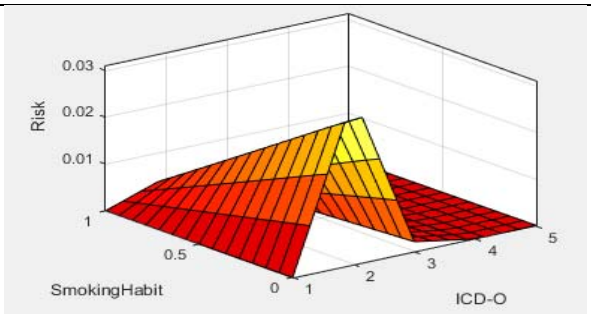
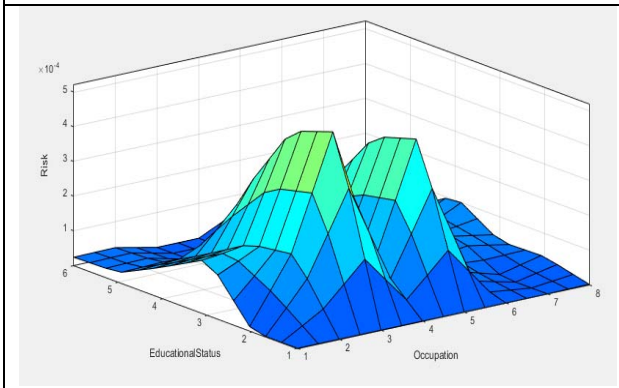
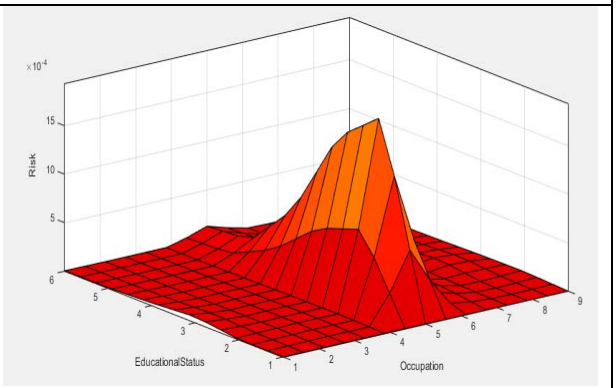
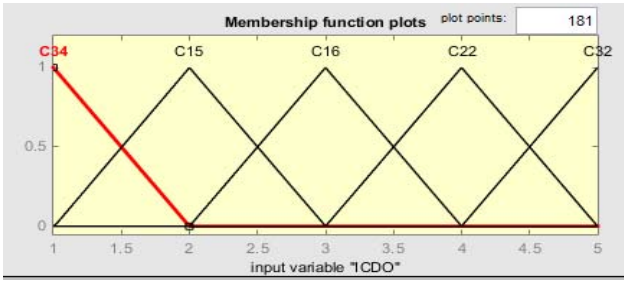
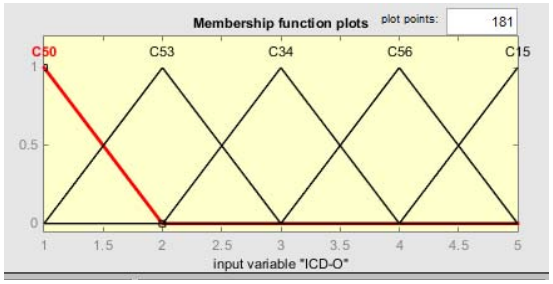
MALE	FEMALE
	
4a. Risk of ICD-O with respect to Occupation	*4e. Risk of ICD-O with respect to Occupation*
	
4b. Risk of ICD-O with respect to Educational Status	*4f. Risk of ICD-O with respect to Educational Status*
	
4c. Risk of ICD-O with respect to Smoking Habit	*4g. Risk of ICD-O with respect to Smoking Habit*
	
4d. Risk of ICD-O with respect to Occupation and Educational Status	*4h. Risk of ICD-O with respect to Occupation and Educational Status*

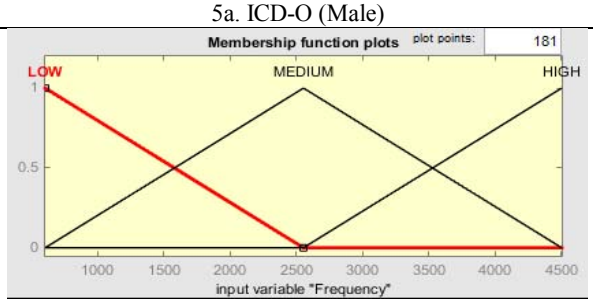
Figure 4 Surface View of the Risk Assessment by the ANFIS



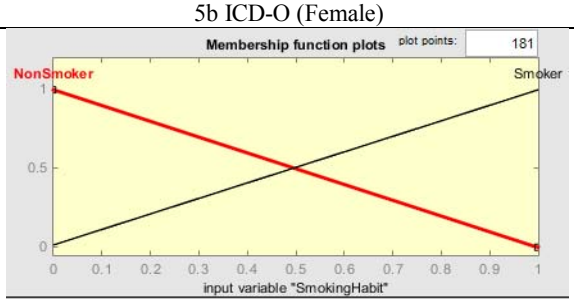
5a. ICD-O (Male)



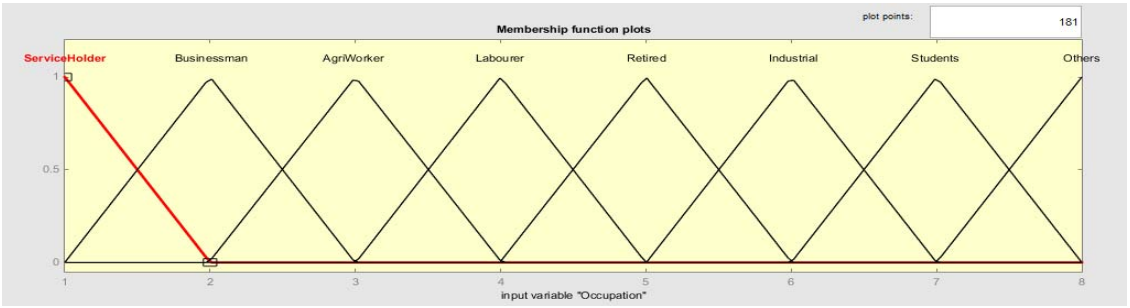
5b ICD-O (Female)



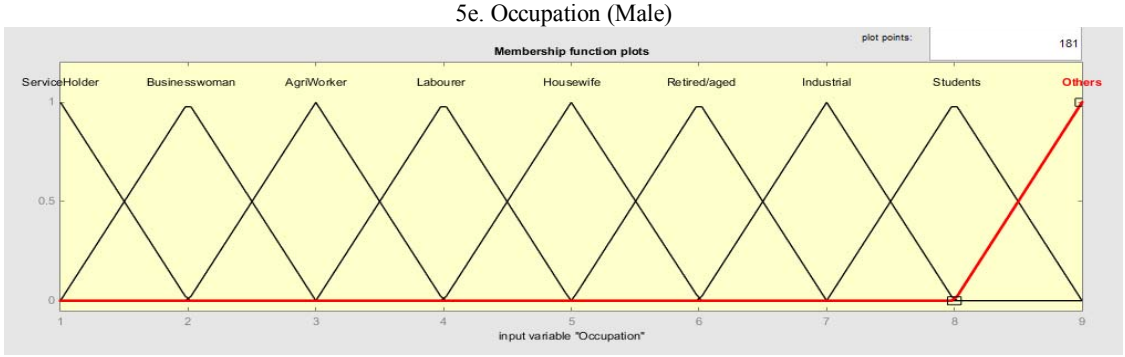
5c. Frequency



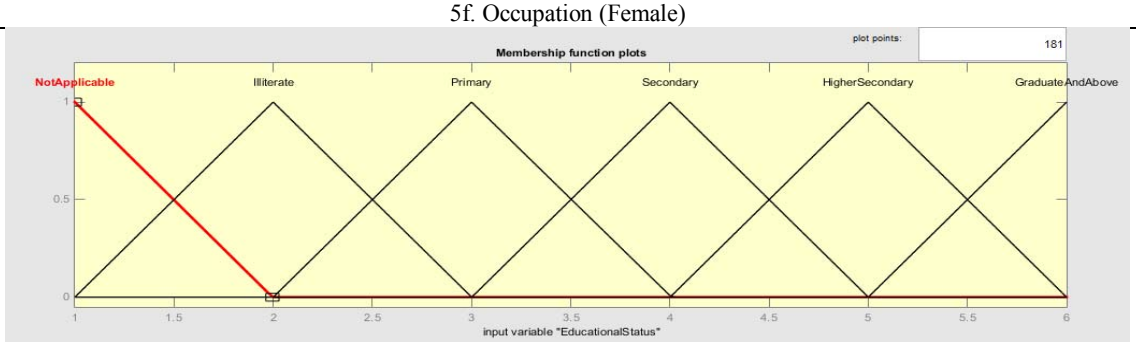
5d. Smoking Habit



5e. Occupation (Male)



5f. Occupation (Female)



5g. Education Status

Figure 5 Membership Function of the FIS Input Variables