Exploring the risk factors of top five malignancies in Bangladesh

S.A.M.F. Rubby,
Sheefta Naz,
A.M. Arefin Khaled,
Ruhul Amin Dicken and
Rashedur M. Rahman*

Department of Electrical and Computer Engineering,
North South University,
Plot-15, Block-B, Bashundhara,
Dhaka 1229, Bangladesh
Email: fazle2712@yahoo.com
Email: sheefta@hotmail.com
Email: akonkshaan@gmail.com
Email: ruhul.amin1125@gmail.com
Email: rashedur.rahman@northsouth.edu
*Corresponding author

Abstract: The number of cancer patients has been increased to a great extent over last few years in Bangladesh. National Institute of Cancer Research and Hospital (NICRH) has confirmed more than 27,000 cancer patients from year 2008–2010. However, the number of researches on cancer patients are limited in the context of Bangladesh. This research aims to explore the key factors that have relationship with cancer. The factors include different demographic and occupational information as well smoking habits of patients. We have generated two adaptive neuro fuzzy inference system (ANFIS) models, one for female and one for male patients of Bangladesh and reported the relationships of different factors with five top malignancies reported in NIRCH data set.

Keywords: adaptive neuro fuzzy inference system; ANFIS; fuzzy inference system; FIS; malignancies; ICD-O; cancer; Bangladesh.

Reference to this paper should be made as follows: Rubby, S.A.M.F., Naz, S., Khaled, A.M.A., Dicken, R.A. and Rahman, R.M. (2017) 'Exploring the risk factors of top five malignancies in Bangladesh', *Int. J. Knowledge Engineering and Soft Data Paradigms*, Vol. 6, No. 1, pp.62–81.

Biographical notes: S.A.M.F. Rubby completed his graduation on Fall 2015 in the Department of Electrical and Computer Engineering of North South University. He is currently working in a private firm as a system administrator. He published papers on data mining and fuzzy inference system under the supervision of Dr. M. Rashedur Rahman and in future he would like to select a career as a research specialist.

Sheefta Naz received her BSc in Computer Science from the North South University of Bangladesh, Dhaka, in 2017. Since then, she has been with the HURDCO International School and College, Dhaka, Bangladesh, where she is currently a high school teacher in Computer Science.

A.M. Arefin Khaled completed his BSc in Computer Science on Fall 2015 in the Department of Electrical and Computer Engineering of North South University. He is currently working at the NextGen Security & Surveillance Ltd. as a Software Engineer. He envisions himself conducting research on application-based algorithmic and optimisation challenges in data science that can empower to fight the challenges of young age.

Ruhul Amin Dicken graduated in 2016 with a BS in Computer Science and Engineering from the Department of Electrical and Computer Engineering (ECE) from the North South University. He currently works as a business analyst at the SELISE Rockin' Software. His research interest includes data mining which aligns with his role at SELISE.

Rashedur M. Rahman is working as a Professor in the Electrical and Computer Engineering Department in the North South University, Dhaka, Bangladesh. He received his PhD from the University of Calgary, Canada in 2007. He published more than 125 research papers in the area of parallel and distributed computing, cloud and grid computing, data and knowledge engineering. He has been serving as a program committee member of many prestigious conferences like ACIIDS, ICCCI, ICECE, etc. He is also in the editorial committee of many international journals. His current research interest is in cloud load characterisation, VM consolidation and application of data mining and fuzzy logic in different decision making problems. He also leads some research teams who work on deep learning for chat service development, recommendation of the quality of fruits etc.

This paper is a revised and expanded version of a paper entitled 'Analysis and classification of respiratory health risks with respect to air pollution levels' presented at 16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015, Takamatsu, Japan, 1–3 June 2015.

1 Introduction

The world around us is evolving daily. Evolution in both human life and animal life brings changes. The changes in the life have its good effects and bad effects. In these bad effects cancer has played a primary role as the worst kind of disease that could end human lives. Some of human diseases in the world are curable and others are not. Cancer is such a disease that hardly has a cure for it. Cancer is a disease which is born from infectious cells. Cancer causes infectious abnormal cell growth in the body. This lump or big cell causes the body space to be invaded by spreading and rapidly growing in different parts of the body (Ghodke et al., n.d.; Latha et al., 2013). Tumours are infectious cell which causes body space invasion but all tumours do not cause cancer. Cancer causes symptoms like abnormal bleeding, extreme coughing, abnormal weight loss, irregular bowel movement etc. These cause the body to react differently than a normal person. In a human body cancer can occur in approximately a hundred different places.

The only way to defeat cancer is to take preventive measures. The cure measure is not very effective in cancer. The process only makes the disease dormant for some time. Cancer can turn dormant but it does not stay dormant for long. Unfortunately, it becomes

active after a short or long period of time (Yilmaz and Ayan, 2013). A preventive measure in cancer includes the separation of infectious part of the body to prevent it from spreading further in the body. The method for curing cancer could be done by different kind of treatments. The treatment includes radiation therapy to help in killing infectious cells in the body or slowing the growth of the bad cells rather making them non-spreading. Another treatment is known as the surgery method to remove the part of infected part from the human body. The most commonly used treatment is known as chemotherapy which is used to kill the infectious cells in the body which is sometimes also known as targeted therapy (Coto et al., 2016). The pain suffered by the patient is also taken care of as well keeping in mind the symptoms of the affected parts of the body. Cancer can also be prevented by avoiding smoking, consumption of tobacco, excessive drinking of alcohol, keeping balanced diet, preventing the radiation exposure and avoiding consumption of processed food.

In the developing countries there is an increase in the number of cancer patients. Although it is not yet discovered the original cause of the cancer in the human body. Some researchers discovered partial causes for cancer in the body. One such is tobacco intake. It is one of the main causes of cancerous cell born in the body. Another main cause of cancer is exposure to extreme radiation. Immense obesity can be a cause of cancer. Some other facts are found like excessive consumption of alcohol, lacking of proper and balanced diet, low physical activity. Exposure to immense amount of insecticides used in farm and pest control could be a cause of cancer. Also exposure to environment effecting pollutions and ionising radiation especially from the mines could be factors of cancer. Genetic change must occur before cancer develops in the human body (Cecere el al., 2012). Approximately with roundabout of ranged 5–10% of the cancer cases are acquired from genetically defects from heredity or in other words inherited from the patients' fore-fathers. In developing countries cancer could occur because of the infectious virus known as the Hepatitis B, Hepatitis C and Human Papilloma Virus. These viruses change the genes of the cells and act as a factor in cancer disease. Cancer can be detected by symptoms of the sick patient and certain screening tests performed in the hospitals. Also cancer can be detected with medical image detection and biopsy.

This research takes in consideration the condition of cancer spread and the state or stage of it in the developing countries. Observing the records from the past few years' in these countries it can be inferred that cancer is not in totally cured treatment status in most cases. This paper considers the detection of cancer as well as the treatment of it. As prevention is the best method of fight against this disease, the first step is to classify the risk of cancer. In this paper occupation, peer habits such as smoking and education level of the people in the country is taken into consideration. This observation of the facts provides a list of possible risk factors and evaluation to gather the top most five malignancies found in the male and female patients of a developing country like Bangladesh.

In Section 2 there is a brief description of related works is presented. In Section 3 the data sources and data used in the paper is presented. The design of the model Adaptive network based fuzzy inference system (ANFIS) and its use in the research is described in detail in Section 4. In Section 5 result and analysis of results are presented. Section 6 concludes the paper with giving direction of further extension of this work.

2 Related works

There are different organs that could be affected by cancer. These organs range from the tongue to the skin. There are more than 100 places in the body where cancer occurs. Cancer can be detectable physically or non-detectable. These non-detectable cancer causes more harm to the body. Cancer caught in the first stage is easier to tackle or can be made dormant or amputated by surgery to avoid the risk of spreading of infected cells in the other parts of the body. Each type of cancer shows abnormal growth in different parts of the body. The tumour growth can be enormous or small. Although tumours do not need to be really big to be cancerous, they need to be infected (Al-Daoud, 2010; Cecere et al., 2012). The infected tumour looks different than the normal. Each type of cancer depends on the stage of the tumour and the place where the abnormal cell growth occurred. The abnormal growth causes the body to suffer a great amount of pain in the body. This makes that part of the body at more risk (Karthikeyeni and Ramya, 2014).

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 (Cecere el al., 2012). Sometimes exposure to pesticides and other manure which are artificially chemical mixture can cause skin cancer. This sulphur exposure in the body is very harmful. Furthermore competition of urbanisation in the community has been very much crucial for the body and the health issues are being overlooked by the authorities. Exposure is not taken as a harmful issue or in other words ignored in the developing countries of the world. More workers who do not know about the situations or the harmfulness of the exposure, harmful chemicals become risky to the infections. Even though we have no specific idea of what causes the cancer infections, these harmful chemicals increase the possibility more than 45% of cancer disease occurring to the human body (Lin, 2009).

In the world different types of cancer dominate in different regions. Lung and liver cancer is common in all around due to obesity and smoking. Thyroid cancer sometimes causes chylous leakage (Li et al., 2016). Lung cancer on the other hand has a division of stages. The further stage is known as advanced lung cancer. However, some patient has only infectious lung cancer while the others have advanced lung cancer. There are different stages which need the study of its regularity in the expansion of the infectious cells (Caponero et al., 2017; Salaken et al., 2017).

There are different types of processes by which detection can be done in finding cancer. For example early detection of breast cancer helps in full removal of the infectious cells from the part of the body. This cancer can be totally cured. However there is a problem in detection of cancer. As mentioned earlier in our paper that cancer is caused by the spreading infectious tumour (Fischer, 2017; Ojha and Goel, 2017) .The tumours can be big or small. The smaller tumours can be so small it is hard to detect in normal process like X-rays (XCT) or (ERT) computed techniques (Alzubaidi et al., 2017). So a physician has to resort into different types of process to find out the cancerous cells in the body. The last part of detecting cancer is biopsy. One research paper (Bilotti et al., 2017) puts forward a model. This model is used by division of detection of cancer infections in the body by using artificial neural network classifier. They took lungs as the main organ for analysis and segmentation.

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 tumour occurrence (Chen et al., 2016; Song and Ma, 2016). 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 (Ghodke et al., n.d.). Illiteracy rate leads to increase of cancer cases in the region (Sammouda, 2016; Sun et al., 2017).

Some of the 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 patients are more prone to the disease due to their habits and less knowledge about the risk factor of the disease (Kamath, 2015; Arita et al., 2006).

3 Data sources

The records of hospital patients diagnosed with cancer are collected and achieved in cancer registries. This could be used for interpretation of statistics and analysis of patient disease history. Such a system was first introduced in Bangladesh by the Department of Cancer Epidemiology of the National Institute of Cancer Research and Hospital (NIRCH, 2016) in 2004 with the support of the World Health Organization.

In this paper, to search for data and 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. It includes statistics of the site of cancer, the top diagnosed malignancies, the smoking habits of the patients, etc. From the report two aspects of information were focused, i.e., the top five malignancies among male and female patients based on the occupation, educational status and smoking habits of the patients as well their demographic information.

3.1 Top five malignancies

International classification for diseases for oncology (ICD-O) are specific codes that are used to classify the type of cancer based on the site of the tumour 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 have decided to focus on the risk assessment of the top five malignancies among males and females of Bangladesh as listed in Table 1.

3.2 Occupation, educational status and smoking habits

As mentioned earlier, from the demographic part of the cancer registry report we have decided to focus on the patient’s occupation, educational status and smoking habits. The statistics of the registry have included these attributes with their numbers separated by

genders and contained within three main attributes such as occupation, educational status and tobacco smoking habit. All of these three attributes can be further categorised in the following way:

1 Occupation
   - Service holder – these group of people are regarded as private/public office employee who generally lives in urban areas.
   - Businessman – this class of people are those which can vary from wealthy class to lower middle-class family who can both reside in rural or urban area.
   - Agri-worker – these are people considered as farmers who live in rural areas.
   - Labourer – these division of people are the day workers who reside in urban areas and perform day to day work, e.g., construction worker, roadside worker, rickshaw puller etc.
   - Housewife – this is only applicable for female patients who generally do not go out of their houses often.
   - Retired/aged – this class of people are those considered as senior citizens.
   - Industrial/factory worker – these clusters of people are solely considered as the industrial workers.
   - Students – these are the young adolescent people.
   - Others.

2 Educational status
   - Not applicable – these are for the infants or babies.
   - Illiterate – individuals who have not received any sort of education.
   - Primary – people who have at least completed their 5th grade.
   - Secondary – this group have their literacy of secondary school certificate, e.g., passed 10th grade.
   - Higher secondary – these people have at least completed their higher secondary examination, e.g., passed 12th Grade.
   - Graduate and above – these group of people are considered as the most literate.

3 Tobacco smoking habit
   - Smoker – based on the smoking habit whether a person is a regular smoker or not.
   - Non-smoker – this are the group who are not related to any sort of smoking.

Table 1 Top five malignancies among male and female patients

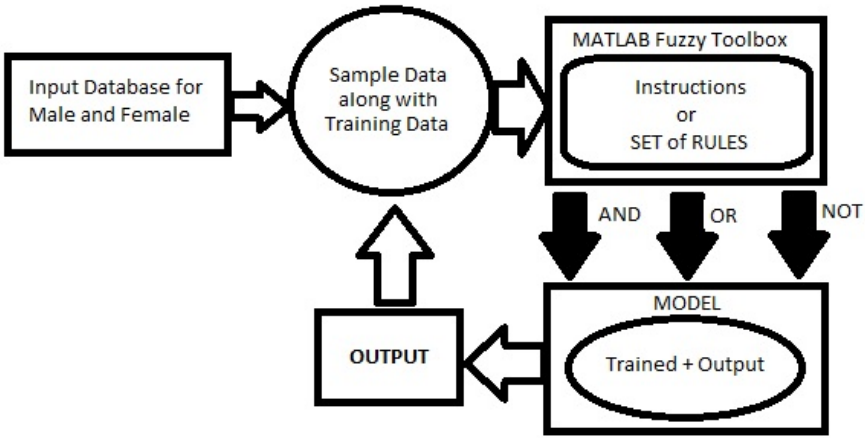
Rank for the period 2008–2010	*Male* (*cancer site – ICD-O*)	*Female* (*cancer site – ICD-O*)
1	Lung – C34	Breast – C50
2	Esophagus – C15	Cervix – C53
3	Stomach – C16	Lung – C34
4	Liver – C22	Ovary – C56
5	Larynx – C32	Esophagus – C15

The reason to focus on these aspects of the cancer patients is 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 is 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 labour 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 is of a higher education background, such as a college or university graduate, is likely to have more knowledge on the suspected dangers of cancer. 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 who 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 is 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.

4 Adaptive network based fuzzy inference system

On the basis of Takagi-Sugeno fuzzy interference system the ANFIS was developed and it is one kind artificial intelligence (AI) system. In the 1990s this technique was developed and it combines the two forms: the fuzzy logic principles and as well as the neural networks (Tsoukalas and Uhrig, 1997), It can gather both the frameworks. The interference system correlates the set of the IF-THEN functions which are nonlinear and to be approximated. The parameters of the membership functions are learned through input-output mapping. Thus it is considered to be the universal estimator (Hong and Lee, 1996).

Figure 2 ANFIS model structure



ANFIS structure is represented in Figure 2. The variables to be used are decided in the next step and are required for creation of the model. The idea is to create an 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 networks and fuzzy logic principles, it is best suited as the 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 (Ya et al., 2014).

4.1 Designing the training dataset

The software used to create the ANFIS is 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. Some sample input is shown in Table 2.

Table 2 Sample of the dataset used to train the ANFIS

ICD-O	*Frequency*	*Smoking habit*	*Occupation*	*Educational status*	*Risk of being diagnosed*
C34	4,505	Smoker	Agri-worker	Illiterate	0.02168
C15	1,002	Non-smoker	Businessman	Graduate	0.00020
C34	4,505	Non-smoker	Service holder	Graduate	0.00056
C22	852	Smoker	Retired	Higher secondary	0.00032

Table 2 depicts four inputs as a sample of the complete dataset. 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 event. 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

$$p(\textit{Risk of an illiterate agri} - \textit{worker who smokes tobacco being with } C34) =$$
$$p(\textit{Patient is diagnosed with } C34) * p(\textit{Patient is a smoker})$$
$$* p(\textit{patient is an agri - worker}) * p(\textit{Patient is illiterate})$$

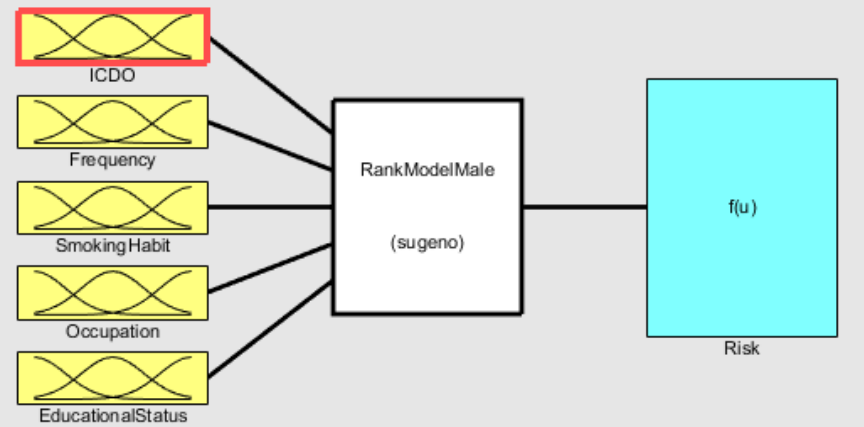
Thus, the measured probability of these set of independent events is 0.02168, which is set at the output in the dataset table.

4.2 Generating the ANFIS

Two separate training data set, as the one depicted in Table 2, were created for male and female and the probability of the risk of being diagnosed cancer is calculated for all possible combinations of the set of demographic events for all the top five ICD-Os.The models are then trained separately for each gender via MATLAB in order to achieve two different risk assessment ANFISs for both male and female. The decision for making the

models separately for the genders was due to the fact that the top malignancies are different for genders, as shown in Table 1, as well as the demographic statistics is varied vastly for each of them.

Figure 2 The FIS generated by the ANFIS toolbox (see online version for colours)



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 male is shown in Figure 2 and it is similar to the FIS generated for female. The subcategories for each attribute as shown in Section 3.2, is taken as members for the corresponding fuzzy variable. The detailed depiction of these membership functions are shown in Figures 3–10. The members of 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 exist two membership structures for the input occupation due to the inclusion of the attribute housewife for females as shown in Figure 13(e) and 13(f). The membership function ICD-O for male and female is shown in Figure 13(a) and 13(b), 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 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 is considering his exposure not only to his own environment but to a certain degree his exposure to other environment as well. Similarly the model lets 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 the 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 times with an effort to reduce the average error of the model as low as possible. The resulting 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 Figures 3–10 and has been discussed in detail in the following section.

5 Results

While making the data tables, numeric values were set for each of the subcategories of 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 shown in Table 1.

Therefore

$1 = Lung, 2 = Esophagus, 3 = Stomach, 4 = Liver, 5 = Larynx.$

- For *occupation*:

  $1 = Service\ holder, 2 = Businessman, 3 = Agri-worker, 4 = Labourer,$
  $5 = Retired, 6 = Industrial/\ factory\ worker,\ 7 = Students, 8 = Others.$

- For *education status*:

  $1 = Not\ applicable, 2 = Illiterate, 3 = Primary,$
  $4 = Secondary, 5 = Higher\ \sec ondary, 6 = Graduate\ or\ above.$

- For *smoking habit*:

  $0 = Non\ smoker, 1 = Smoker$

5.1.1 Risk assessment of ICD-O with respect to occupation

As shown in Figure 3, 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.

5.1.2 Risk assessment of ICD-O with respect to educational status

The risk of the malignancies can be observed to rise significantly in Figure 4 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.

Figure 3 Risk of ICD-O with respect to occupation for male (see online version for colours)

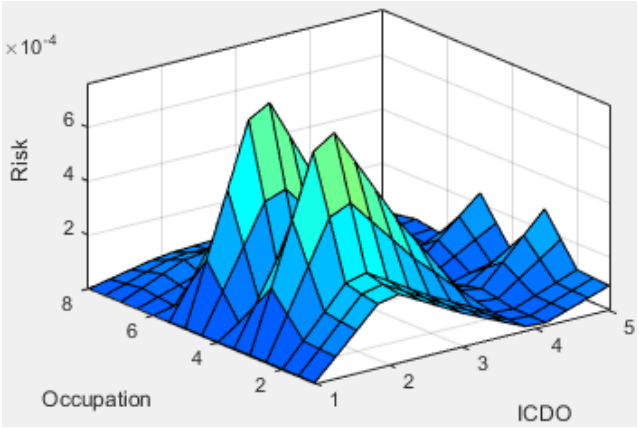


Figure 4 Risk of ICD-O with respect to educational status for male (see online version for colours)

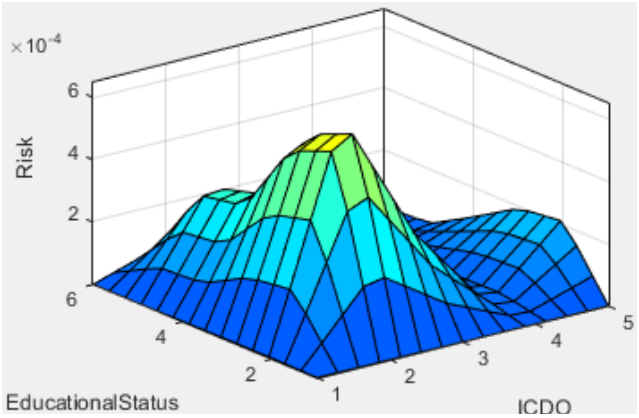
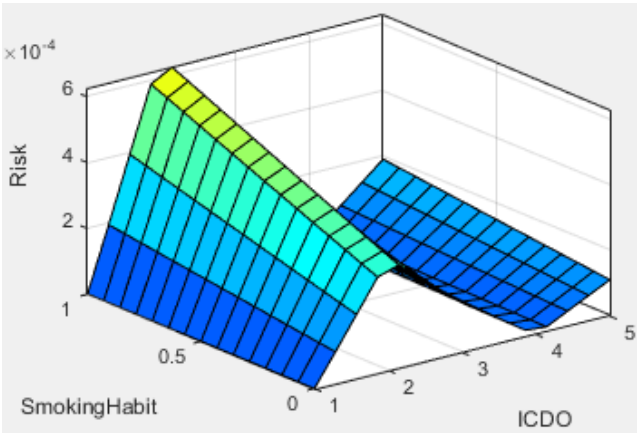


Figure 5 Risk of ICD-O with respect to smoking habit for male (see online version for colours)



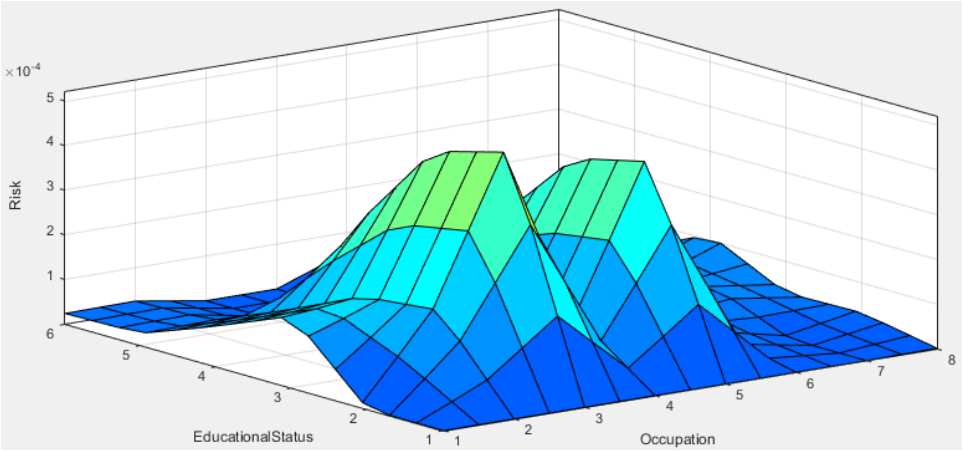
5.1.3 Risk assessment of ICD-O with respect to smoking habit

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 5.

5.1.4 Risk assessment with respect to occupation and educational status

Figure 6 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.

Figure 6 Risk of ICD-O with respect to occupation and educational status for male (see online version for colours)



5.2 Risk assessment of the top five malignancies among females

Similar to the previous section the ICD-Os are 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 = *Labourer*,
  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.

5.2.1 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 7 the model shows an immense elevation in risk around five 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.

Figure 7 Risk of ICD-O with respect to occupation for female (see online version for colours)

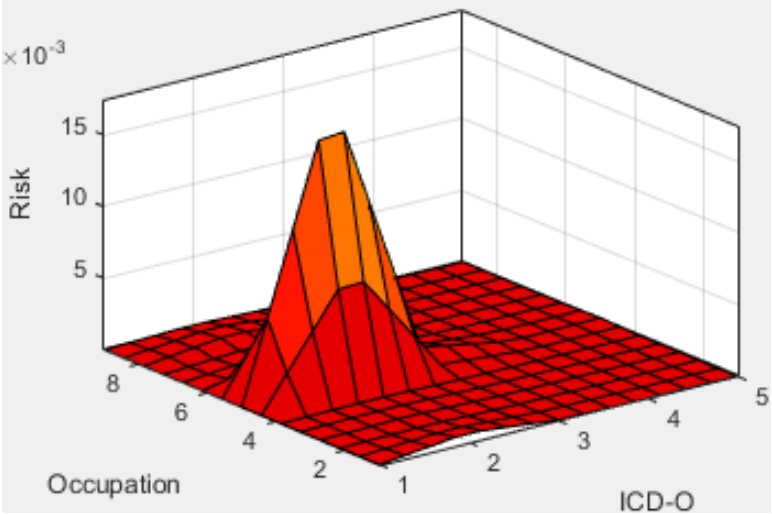
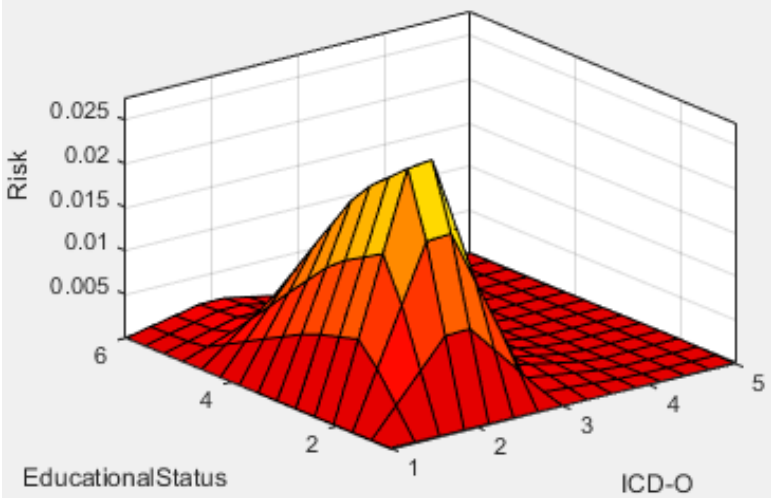


Figure 8 Risk of ICD-O with respect to educational status for female (see online version for colours)



5.2.2 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 8. It seems that it is women who are illiterate or of lower educational background are more at risk of being diagnosed than women with higher educational status.

5.2.3 Risk assessment of ICD-O with respect to smoking habit

Even though Figure 9 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.

Figure 9 Risk of ICD-O with respect to smoking habit for female (see online version for colours)

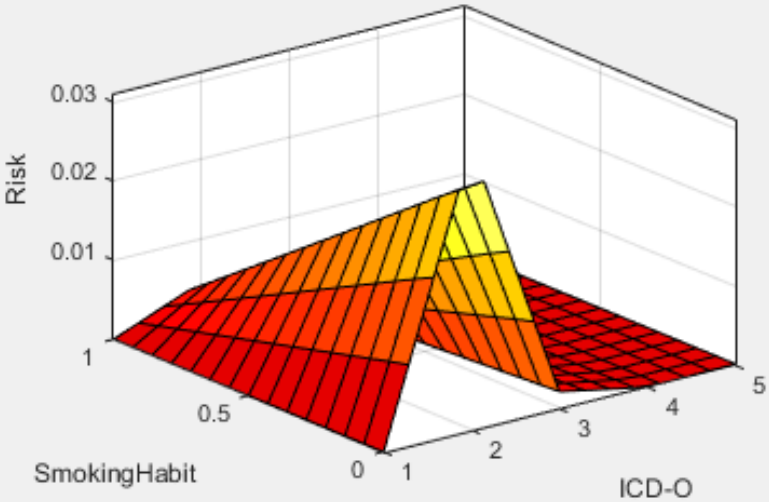
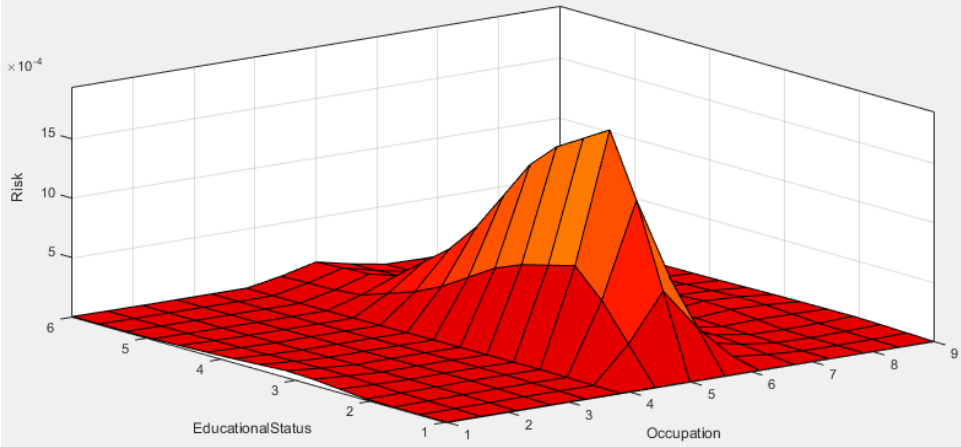


Figure 10 Risk of ICD-O with respect to occupation and educational status for female (see online version for colours)



5.2.4 *Risk assessment of ICD-O with respect to occupation and educational status*

According to Figure 10 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 models for both genders are evaluated. Since the models output a value of the risk using FIS generated rules for each set of independent events, all of these values are first listed. To depict the accuracy of the models a test is carried where the average risk is calculated from the ANFIS. This is done for each ICD-O. The result of this is shown in Figures 11–12.

Figure 11 Accuracy test for each ICD-O for male (see online version for colours)

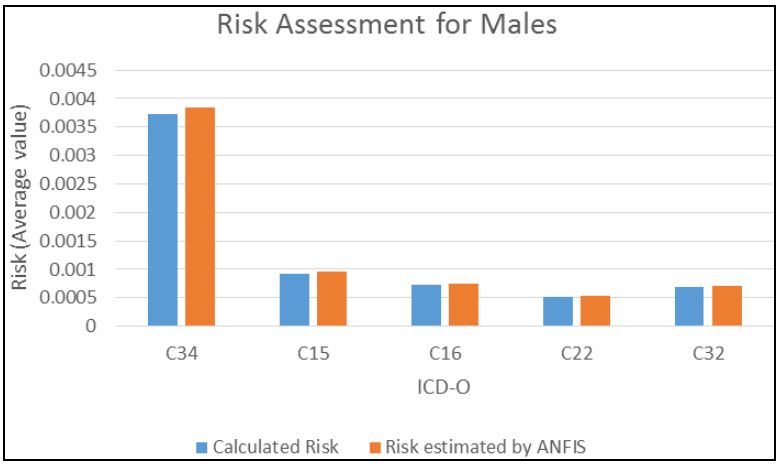
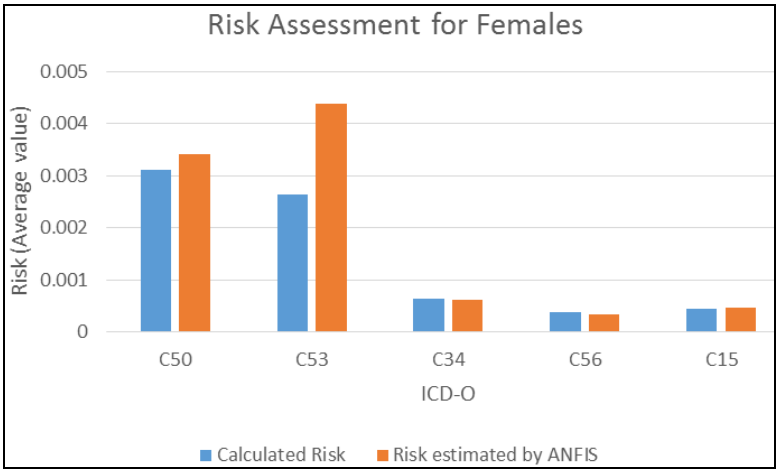
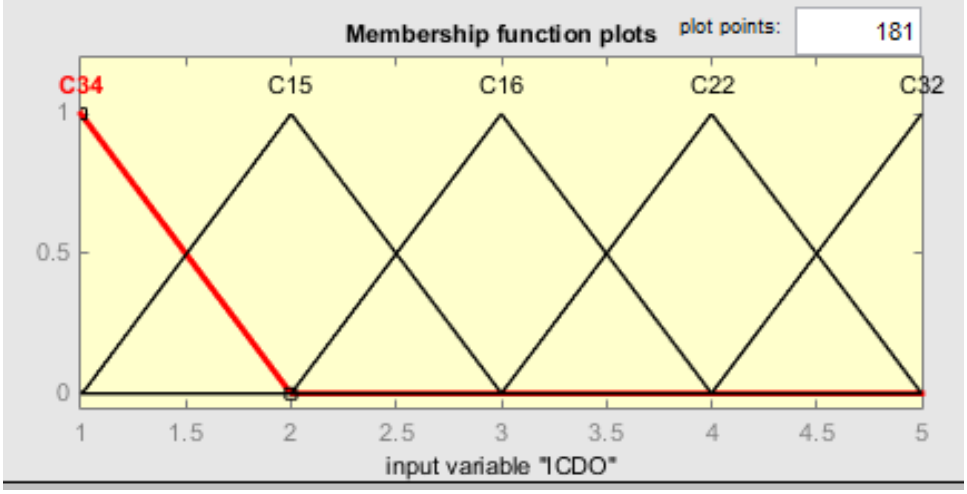


Figure 12 Accuracy test for each ICD-O for female (see online version for colours)

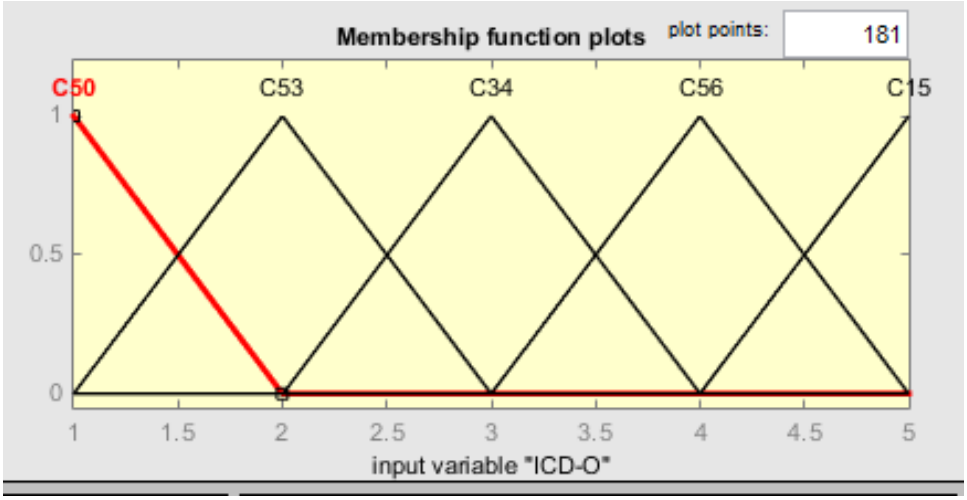


In Figure 11, for each ICD-O, the average risk that is calculated from the actual dataset is compared with the average risk assessed by the ANFIS for the corresponding ICD-O. The blue bars display the actual calculated risk, and the orange ones indicate the estimated risk by the ANFIS. For most of the malignances the estimated risk is similar to the actual risk, the only exception is C53, i.e., cervix cancer, where the ANFIS estimates a higher level of risk for females than the actual risk. The low error percentage demonstrates the validity of the risk assessment of the models for the population of Bangladesh.

Figure 13 Membership function of the FIS input variables, (a) ICD-O (male) (b) ICD-O (female) (c) frequency (d) smoking habit (e) occupation (male) (f) occupation (female) (g) education status (see online version for colours)

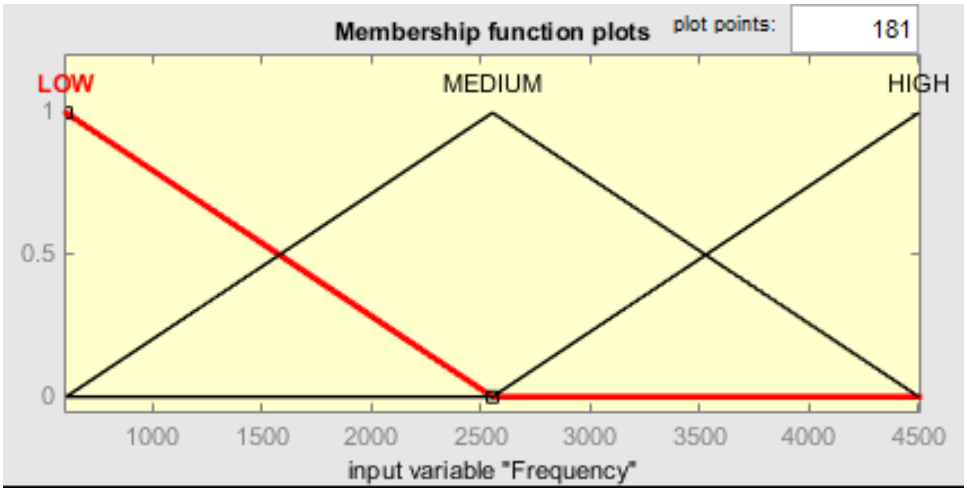


(a)

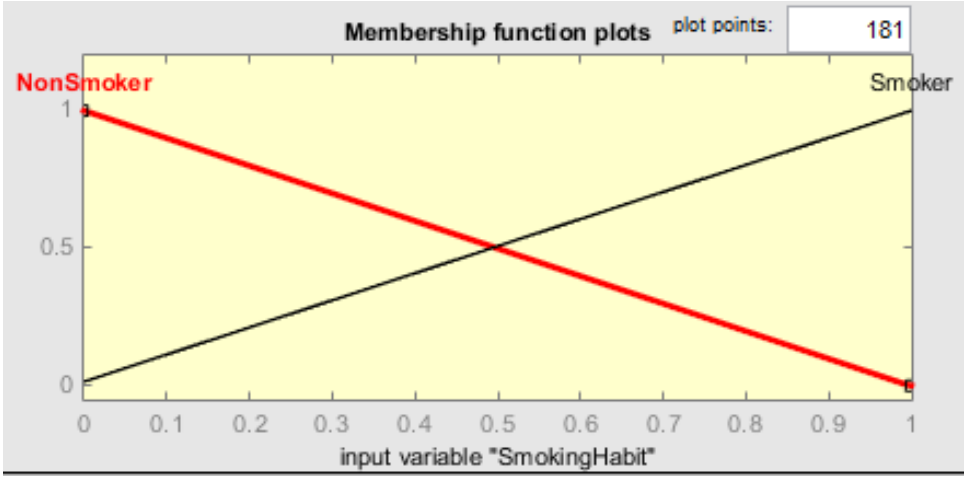


(b)

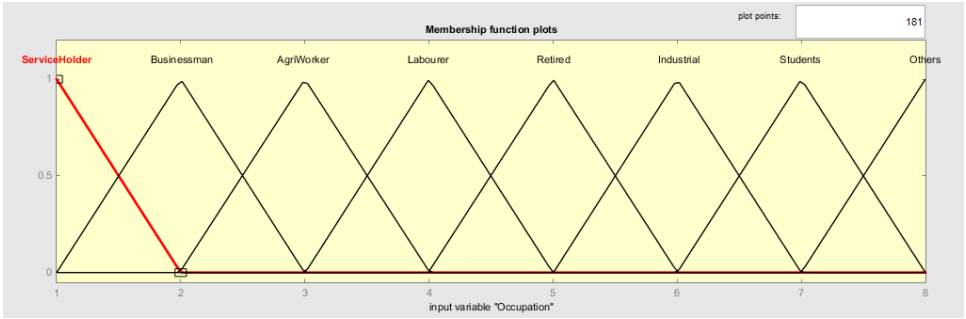
Figure 13 Membership function of the FIS input variables, (a) ICD-O (male) (b) ICD-O (female) (c) frequency (d) smoking habit (e) occupation (male) (f) occupation (female) (g) education status (continued) (see online version for colours)



(c)

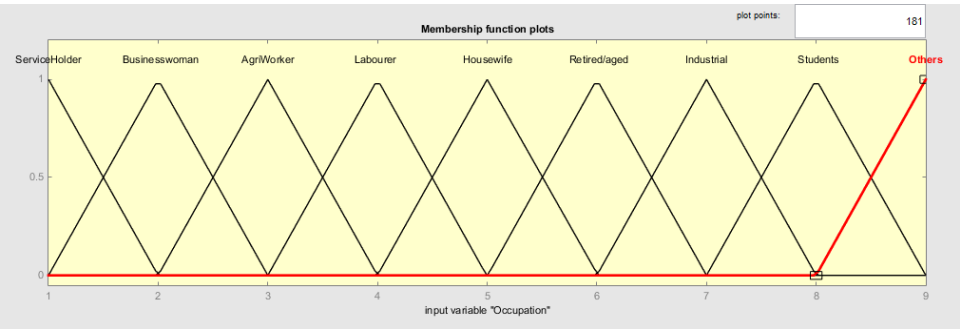


(d)

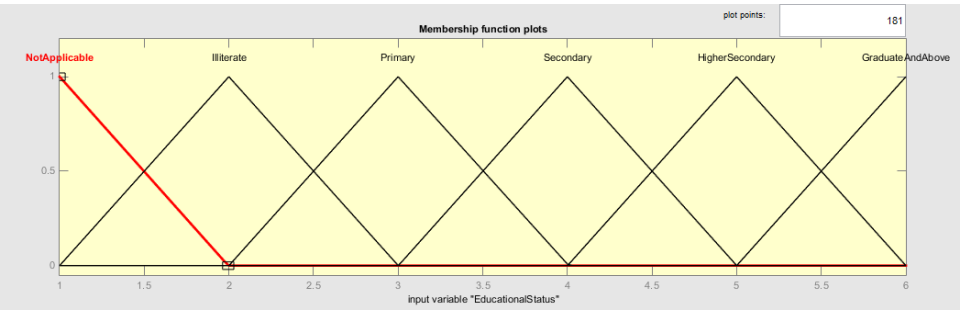


(e)

Figure 13 Membership function of the FIS input variables, (a) ICD-O (male) (b) ICD-O (female) (c) frequency (d) smoking habit (e) occupation (male) (f) occupation (female) (g) education status (continued) (see online version for colours)



(f)



(g)

6 Conclusions and future work

A number of limitations and obstructions had to be overcome in the process of completing the research. Among those the most troublesome was getting access to informative data and statistics which were resulted due to the presence of the cancer registry report. From the vast data, a lot had to be processed to make data ready for our model to work. Moreover, the accumulated data was in huge chunk to process in MATLAB's Fuzzy tool box. 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. This research will create the window of opportunity to generate a similar model for other diseases but presence of valid data is mandatory. Medical policies and health programs could also be designed based on the risk assessment models of different genders. This could aid better health plan in future that the authorities need to focus. 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 for both urban and rural areas. As there is no work for malignancy prediction by machine learning in the context of Bangladesh, we are unable to make a direct comparison of other methods with fuzzy logic. As a future work we plan to compare the performance of fuzzy logic against other machine learning techniques.

References

Al-Daoud, E. (2010) 'Cancer diagnosis using modified fuzzy network universal Journal of Computer Science and Engineering Technology', *Universal Journal of Computer Science and Engineering Technology*, November, Vol. 1, No. 2, pp.73–78.

Alzubaidi, A., Sideseq, F., Faeq, A. and Basil, M. (2017) 'Computer aided diagnosis in digital pathology application: review and perspective approach in lung cancer classification', in *2017 Annual Conference on New Trends in Information & Communications Technology Applications* (*NTICT*).

Arita, S., Nomura, T. and Sonoo, H. (2006) 'Diagnostic system of breast cancer based on imaging data of mammography using fuzzy logic', in *2006 World Automation Congress Conference*, DOI: 10.1109/WAC.2006.375933.

Bilotti, C., Lucena, T., Rodrigues, S. and Bernuci, M. (2017) 'Sketching a mHealth based system to improve breast cancer prevention', in *2017 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges* (*GMEPE/PAHCE*).

Caponero, M., Polimadei, A., Ariano, M., .Schena, E., Massaroni, C., Silvestri, S. and Saccomandi, P. (2017) 'Fabrication and calibration of three temperature probes for monitoring the effects of thermal cancer ablation', in *2017 IEEE International Instrumentation and Measurement Technology Conference* (*I2MTC*).

Cecere, L.M., Williams, E.C., Sun, H.,. Bryson, C.L., Clark, B.J., Bradley, K.A. and Au, D.H. (2012) 'Smoking cessation and the risk of hospitalization for pneumonia', *Respiratory Medicine*, July, Vol. 106, No. 7, pp.1055–1062.

Chen, Z., Xu, W., Yang, Y., Yan, J. and Chen, Q. (2016) 'Study on the infectious regularity of patients with advanced lung cancer', in *2016 8th International Conference on Information Technology in Medicine and Education* (*ITME*).

Coto, J.F., Siles, R. and Rodriguez, M. (2016) 'Biocomputing platform module for cancer genomics and chemotherapy', *2016 IEEE 36th Central American and Panama Convention* (*CONCAPAN*).

Fischer, S. (2017) 'Sniffing for cancer: nano noses hold promise for detecting lung cancer and other diseases', *IEEE Pulse*, July–August, Vol. 8, No. 4, pp.20–22, doi: 10.1109/MPUL.2017.2701488.

Ghodke, L., Naik, A., Konale, R. and Mehta, S. (n.d.) *Brain Cancer Detection Using Neuro Fuzzy Logic*, pp.58–61 [online] http://www.interscience.in (accessed 10th August 2017).

Hong, T.P. and Lee, C.Y. (1996) 'Induction of fuzzy rules and membership functions from training examples', *Fuzzy Sets and Systems*, 25 November, Vol. 84, No. 1, pp.33–47.

Kamath, S. (2015) 'Fuzzy logic for breast cancer diagnosis using medical thermogram images', *Fuzzy Expert Systems for Disease Diagnosis*, *Chapter 7*, pp.168–199, IGI Global, DOI: 10.4018/978-1-4666-7240-6.ch007.

Karthikeyeni, S. and Ramya, S. (2014) 'Comparative analysis of ANFIS and FRBF – survival time prediction of cancer pattern', *International Journal of Advanced Research in Computer and Communication Engineering*, September, Vol. 3, No. 9, pp.7992–7995.

Latha, K.C., Madhu, B., Ayesha, S., Ramya, R., Sridhar. R. and Balassumaran, S. (2013) 'Visualization of risk of breast cancer using fuzzy logic in MATLAB environment', *International Journal of Computational Intelligence Techniques*, Vol. 4, No. 1, pp.114–117.

Li, Y., Li, Y., Wang, P., Sun, S. and Chen, L.G. (2016) 'Sharing experience in the treatment of chylous leakage in thyroid cancer radical resection and central lymph node dissection patients', in *2016 8th International Conference on Information Technology in Medicine and Education* (*ITME*).

Lin, J. (2009) 'Cancers in normal mice exposed to cell phone radiation [health effects]', *IEEE Microwave Magazine*.

National Institute of Cancer Research and Hospital (NIRCH) (2016) *National Institute of Cancer Research and Hospital* [online] http://nicrhbd.org/ (accessed 8 January 2016).

Ojha, U. and Goel, S. (2017) 'A study on prediction of breast cancer recurrence using data mining techniques', in *2017 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence*.

Salaken, S., Khosravi, A., Khatami, A., Nahavandi, S., and Hosen, M. (2017) 'Lung cancer classification using deep learned features on low population dataset', in *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering* (*CCECE*).

Sammouda, R. (2016) 'Segmentation and analysis of ct chest images for early lung cancer detection', in *2016 Global Summit on Computer & Information Technology* (*GSCIT*).

Song, S. and Ma, Y. (2016) 'The research and application of technology in the diagnosis of lung cancer warning association rule mining', in *2016 8th International Conference on Information Technology in Medicine and Education* (*ITME*).

Sun, B., Yue, S., Hao, Z., Cui, Z., Wang, H. and Zhang, W. (2017) 'Early lung cancer identification based on ERT measurements', in *2017 IEEE International Instrumentation and Measurement Technology Conference* (*I2MTC*).

Tsoukalas, L.H. and Uhrig, R.E. (1997) *Fuzzy and Neural Approaches in Engineering*, World Scientific, John Wiley Inc., 605 Third Avenue, New York, NY, USA.

Ya, H.Y, Kun, T.F., Zhong, P.Z. and Li, Q (2014) 'Speed tracking control using an ANFIS model for high-speed electric multiple unit', *Control Engineering Practice*, February, Vol. 23, pp.57–65.

Yilmaz, A. and Ayan, K. (2013) 'Cancer risk analysis by fuzzy logic approach and performance status of the model', *Turkish Journal of Electrical Engineering & Computer Sciences*, Vol. 21, pp.897–912 [online] http://journals.tubitak.gov.tr/elektrik/issues/elk-13-21-3/elk-21-3-20-1108-22.pdf (accessed 5th November 2017).