A Fuzzy Neuro Approach to Identify Diarrhea Epidemic in Bangladesh

Muhammad Swadhin Shahriar Faruque
Department of Electrical and Computer Engineering
North South University
Plot-15, Block-B, Bashundhara, Dhaka, Bangladesh
swadhin.shahriar.alpha@gmail.com

Shourav Banik, Rashedur M Rahman
Department of Electrical and Computer Engineering
North South University
Plot-15, Block-B, Banshundhara, Dhaka, Bangladesh
shourav.b2283@gmail.com, rashedur@northsouth.edu

***Abstract*— The paper presents the identification of probability of diarrhea epidemic occurring in Bangladesh per month using a competitive neural network and fuzzy logic. Here we have divided the months into six seasons: spring, summer, rainy, early fall, late fall, winter. The infected rate is divided into four parts: low, medium, high, very high. At first infection rate in each season is learned by using a competitive neural network and then the identification of the percentage of an epidemic occurrence is done by fuzzy algorithm (specifically by the Mamdani Min). The centroid function was later used to get a crisp value that corresponds to the probability of epidemic in a certain year.**

Keywords— diarrhea; epidemic; competitive neural network; fuzzy logic.

I. INTRODUCTION

Thousands of people are infected and hospitalized because of diarrhea in this country. Sometimes it becomes rather tedious to identify whether an epidemic is truly occurred due to diarrhea. This paper presents a model which is trained with previous records in such a way that the trained model can identify the diarrhea epidemic by knowing the current time frame and the total number of infected people in the current time frame. To train our model using the data set, competitive neural network was used. Fuzzy algorithm also helped to identify the high and low points of the infection rate. Instead of going for crisp values which changes from time to time fuzzy values were used due to its flexibility. Finally the Fuzzy Inference Algorithm (FIS) was used to show the epidemic rate. Using fuzzy neuro approach we addressed the property of vagueness in identifying epidemic instead of the usual approximations in the Bayesian context.

Section II will briefly discuss related work already done in this field. Section III depicts the data acquisition part. In section IV we have described the methodology. Section V discusses and analyses the result gained from the system. Lastly section VI concludes and gives direction of future work.

II. RELATED WORK

A lot of works have been conducted on disease detection and epidemic predictability model over the years. A research done on cholera trend in Matlab, Bangladesh found that although cholera occurrences vary every year, the seasonality remain more or less the same every year. The research showed that the seasonal outbreaks could be predicted by increased water temperature and other environmental factors [4]. It was found that combination of LGCA (Lattice Gas Cellular Automaton) and Fuzzy Logic could help figuring out the spatial characteristic of epidemic [7]. In [6] Rupasinghe et al. worked to detect dengue risk probability based on Fuzzy Interference System (FIS). Precipitation, humidity, temperature and urbanization of a certain area were collected and used to build the FIS. Bell shaped functions were used as membership function for the attributes. The authors argued that the true nature of the membership function was unknown. To overcome this issue they used an adaptive neuro fuzzy system to train and optimize the membership function [6]. Again in [8] instead of using traditional fuzzy logic, multiresolution analysis and fuzzy system was used to build a model of dengue and severe dengue case. This model helped to predict dengue fever occurrence in the upcoming future. Fuzzy system is quite generally used in diagnosis and detection of diseases, but in case of epidemic identification usage of fuzzy system is quite rare. Therefore, the research presented in this paper is unique.

III. DATA ACQUISITION

The data of diarrhea infected people of Bangladesh in 10 consecutive years starting from 1998 to 2007 is used in order to establish our epidemic identification model. Fig. 1 depicts the infected people of Bangladesh every month in those 10 years. This data set was collected from Institute of Epidemiology, Disease Control and Research (IEDCR). Fig. 2 is the total population of Bangladesh [1]. We then pre-processed the data by dividing each month's infected number by that year's total population. This gave us the percentage of the infected people per month. Fig. 3 depicts this.

	A	B	C	D	E	F	G	H	I	J	K
1		1998	1999	2000	2001	2002	2003	2004	2005	2006	200
2	January	85782	129566	102236	141620	169553	171312	136726	153005	153971	13807
3	February	84917	114606	98604	119875	162922	149504	130365	140679	145718	14222
4	March	101846	127884	103656	134035	177094	164890	150876	160854	163824	16310
5	April	129437	147436	116019	183712	199406	209527	182982	197937	180705	19239
6	May	132072	159268	124531	166249	200342	209414	200698	211260	187210	22212
7	June	139671	134843	124625	160483	214710	192949	183577	194278	174369	20660
8	July	143442	134591	139238	160880	205959	225130	206211	176502	181708	20489
9	August	199038	126151	138895	149115	308032	206017	286524	183028	163253	24274
10	September	373525	125991	141166	155949	231466	202839	199120	182357	153603	23384
11	October	283481	124505	195153	171739	227326	197335	223695	209497	164993	20573
12	November	195396	115642	144226	162286	305999	200014	168283	181977	154116	20222
13	December	189810	107803	126903	156308	196416	158332	178102	160547	136380	18135

Figure 1. Number of Infected people Per Month

	A	B
1	1998	135692000
2	1999	138235000
3	2000	140767000
4	2001	143289000
5	2002	145797000
6	2003	148281000
7	2004	150726000
8	2005	153122000
9	2006	155463000
10	2007	157753000

Figure 2. Total Population of Bangladesh

IV. METHODLOGY

We divide twelve months into six seasons. Given Bangladesh has six seasons it was easier to identify the infection rate over the period of seasonal changes. Using seasons rather than months to build our time membership function, the changes could be observed more clearly. Center for Disease Control and Prevention (CDC) defines endemic as, "the amount of a particular disease that is usually present in a community is referred to as the baseline or endemic level of the disease." From this definition we can say that if a disease is common then only variation from the normal position (norm) should be investigated. Endemic mainly refers to the constant presence of a disease in a population [2]. The classification of percentage of people infected (PPI) is based on this because if a disease is usually present in a community no matter how high it might be with regards to other area, if it is below endemic level it cannot be called an epidemic and the chance of an epidemic in that area is very low. Now epidemic is defined as, "epidemic refers to an increase, often sudden, in the number of cases of a disease above what is normally expected in that population in that area" [2].

First, the dataset is partitioned into two with regards to the mean of the dataset. Then again two means of partitioned dataset is calculated. These three points (means) act as the peak of the three PPI class labels, i.e., low, medium, and high. After that the dataset is again partitioned using the mean value which is the peak of PPI class label high and mean is calculated. The new point is labeled as very high. Then a competitive neural network approach is used exhaustively to get an insight of our data set. This gives us in which season diarrhea infection is more frequent. After that our Fuzzy

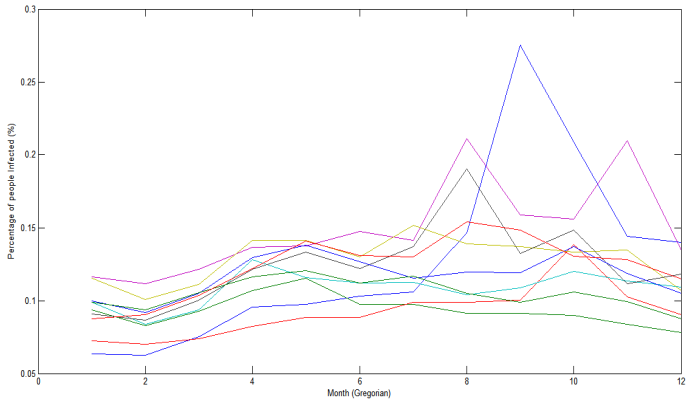

Figure 3. Infection Rate

Inference System is made from the insight gained from the neural network. In the FIS, the rules reflect the definition of epidemic. Basically for each season the endemic rate is calculated and that is put up against the current PPI rate to get the epidemic chance. Lastly we defuzzify the whole process to get a crisp value in the percentage.

A. Division of Months(Seasons)

To make the month's membership function we considered seasons. Bangladesh is known for having six seasons. Most of the common diseases occur due to seasonal change or disturbance in the flow of nature. Therefore, we have taken six seasons each constituting of three consecutive months as seen in Table I. Here we used the Gaussian Membership function. This is used due to its close resemblance to natural distribution. The use of Gaussian Membership function provided us with the flexibility to separate the end points of the months. The standard variance is chosen to be 30 to maintain a wide perspective.

TABLE I. CLASSIFICATION OF SEASONS

Seasons	Months
Spring	February, March, April
Summer	April, May, June
Rainy	June, July, August
Early Fall	August, September, October
Late Fall	October, November, December
Winter	December, January, February

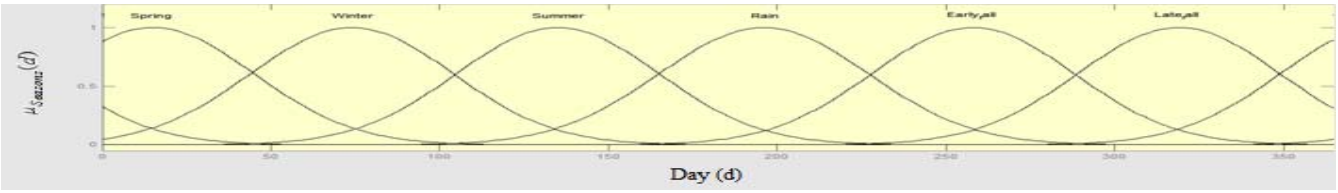

Figure 4. Seasons Membership Function

$$\mu_{Seasons}(x)=\begin{cases} e^{\frac{-(x-135)^2}{2(30)^2}}, 1\le x\le 365 \text{ (x is summer)} \\ e^{\frac{-(x-196)^2}{2(30)^2}}, 1\le x\le 365 \text{ (x is rain)} \\ e^{\frac{-(x-258)^2}{2(30)^2}}, 1\le x\le 365 \text{ (x is earlyfall)} \\ e^{\frac{-(x+44.91)^2}{2(30)^2}}, 1\le x<150 \text{ (x is latefall)} \\ e^{\frac{-(x-319)^2}{2(30)^2}}, 150\le x\le 365 \text{ (x is latefall)} \\ e^{\frac{-(x-74)^2}{2(30)^2}}, 1\le x<150 \text{ (x is winter)} \\ e^{\frac{-(x-439.3)^2}{2(30)^2}}, 150\le x\le 365 \text{ (x is winter)} \\ e^{\frac{-(x-15)^2}{2(30)^2}}, 1\le x<150 \text{ (x is spring)} \\ e^{\frac{-(x-379.1)^2}{2(30)^2}}, 150\le x\le 365 \text{ (x is spring)} \end{cases} \quad (1)$$

B. Classification of Percentage Infected

Percentage of people infected (PPI) in each month with respect to year was divided into four parts: low, medium, high, very high. For the membership function we used trapezoidal function for low and very high. Triangular function was used to define medium and high value. The parameters were decided by observing the endemic and epidemic level of Bangladesh form 1999 to 2007.

$$\mu_{PPI}(x)=\begin{cases} 1, 0\le x\le 0.09661 \text{ (x is low)} \\ \frac{0.1145-x}{0.01789}, 0.09661\le x\le 0.1145 \text{ (x is low)} \\ \frac{x-0.09661}{0.01789}, 0.09661\le x\le 0.1145 \text{ (x is medium)} \\ \frac{0.1349-x}{0.0204}, 0.1145\le x\le 0.1349 \text{ (x is medium)} \\ \frac{x-0.1145}{0.0204}, 0.1145\le x\le 0.1349 \text{ (x is high)} \\ \frac{0.1508-x}{0.0159}, 0.1349\le x\le 0.1508 \text{ (x is high)} \\ \frac{x-0.1349}{0.0159}, 0.1349\le x\le 0.1508 \text{ (x is very high)} \\ 1, 0.1508\le x\le 1 \text{ (x is very high)} \end{cases} \quad (2)$$

C. Competitive Neural Network (CNN)

In order to work with our Fuzzy Inference System, at first we had to identify endemic and epidemic level of diarrhea in Bangladesh. To do so we acquired the data set and pre-processed it as described in the data acquisition part. Then after normalizing the dataset we ran an unsupervised single layer exhaustive competitive neural network to classify the dataset into three clusters. The input of the CNN is the preprocessed data set. Six prototype vectors are used as output which represents a season. Given three months were considered as a single season and the dataset having record of diarrhea infected people of 8 years each prototype vector becomes a column matrix of size 24X1. The weight of the CNN is generated randomly with a size 6X24. Following the Kohonen Rule, for the competitive network, the winning neuron has an output of 1 and the other neurons have an output of 0. Therefore, only the winning neurons incoming weights are updated each time [2]. With a learning rate of .5 we ran the network and generate three different cluster points from our randomly generated weight. The above process is repeated 1000 times to classify the cluster points more accurately and to avoid noise and limitations of neural clustering. The cluster point that has a high average PPI rate is classified as high and subsequently others as medium and low. The result of the process stated above is given on Table II.

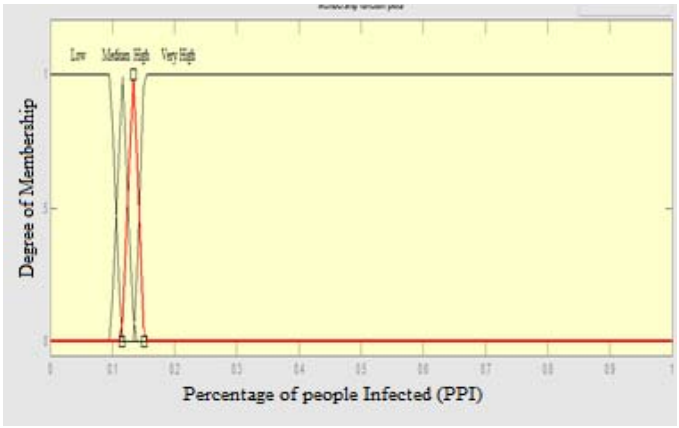

Figure 5. Infection Rate Membership Function

TABLE II. SEASONAL ENDEMIC LEVEL

Iteration	1st	2nd	3rd	4th	5th	6th
Season	Summer	Early Fall	Rain	Late Fall	Spring	Winter
Endemic level	High		Medium		Low	

D. Fuzzy Inference Algorithm

By using the insight gained from the exhaustive use of competitive neural network we now build the consequent for our algorithm. As we are trying to identify a nationwide epidemic the community in this case is constant. The only variables left are time and percentage of people infected, in this case Seasons and PPI respectively. From seasons and PPI we acquire 24 antecedent for which consequent is selected. For selecting consequent we make use of the insight gained from Table II. Taken into account the concept of endemic, we can deduct that, the average PPI of that season is that season's endemic level and so the chance of epidemic is low. PPI which are slightly higher than endemic level has a moderate chance of epidemic and other higher PPI is classified as epidemic.

Based on the concept described above, twenty unique fuzzy rules are generated. For the membership function of our consequent or epidemic chance we have made three divisions: low, medium, high. Again we have used trapezoidal function for low and high and triangular function for the medium chance of epidemic. Now for any given month and population and infected rate of that month we can show the chance of diarrhea epidemic. It is important to note that here we have used Mamdani Min for the fuzzy algorithm. And centroid function was used to get the crisp value of our epidemic chance.

TABLE II. FUZZY RULES

	Percentage of People Infected (PPI)				
Season	**EC**	**PPI_{Low}**	**PPI_{Medium}**	**PPI_{High}**	**PPI_{VHigh}**
	Spring	Medium	High	High	High
	Summer	Low	Low	Medium	High
	Rainy	Low	Medium	High	High
	Early Fall	Low	Low	Medium	High
	Late Fall	Low	Medium	High	High
	Winter	Medium	High	High	High

$$\mu_{EC}(x)=\begin{cases} 1 \ ,0\le x\le 30 (x \text{ is low}) \\ \frac{50-x}{20},30\le x\le 50 (x \text{ is low}) \\ \frac{x-30}{20},30\le x\le 50 (x \text{ is medium}) \\ \frac{70-x}{20},50\le x\le 70 \ (x \text{ is medium}) \\ \frac{x-50}{30},50\le x\le 80 \ (x \text{ is high}) \\ 1 \ ,80\le x\le 100 (x \text{ is high}) \end{cases} \quad (3)$$

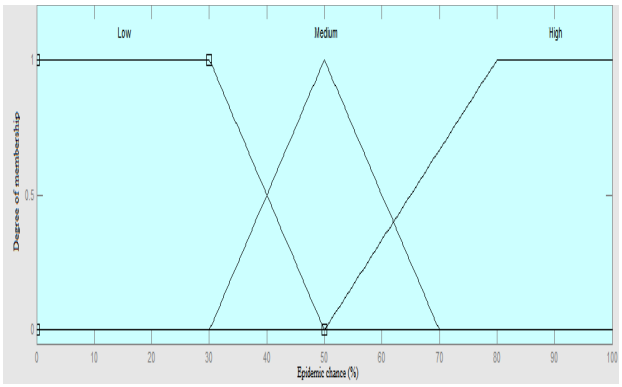

Figure 6. Epidemic Membership Function

E. Defuzzification

Defuzzification is the process of converting a fuzzy value to a crisp value. Center of gravity also known as Centroid Defuzzification method is used to identify the chance of epidemic.

$$output^{*}=\frac{\int \mu_{output}(EC)*EC\ dEC}{\int \mu_{output}(EC)\ dEC} \quad (4)$$

Where μ_{output} is the aggregated membership function obtained from the Fuzzy Inference Algorithm as described before.

V. Result Analysis

Here Fig. 7 presents a view of how the PPI effects the epidemic chance. We can see the PPI rate of 0.3 or above will result in a more than 80% chance of epidemic occurrence. Fig 8. depicts how the chance of diarrhea epidemic changes from season to season. The horizontal axis holds the value from 0 to 365 (days of a year). As the days are divided into seasons, from here we can understand in which month diarrhea infection rate is high. Fig. 9 takes account both of these into consideration and produces the surface graph. From this we

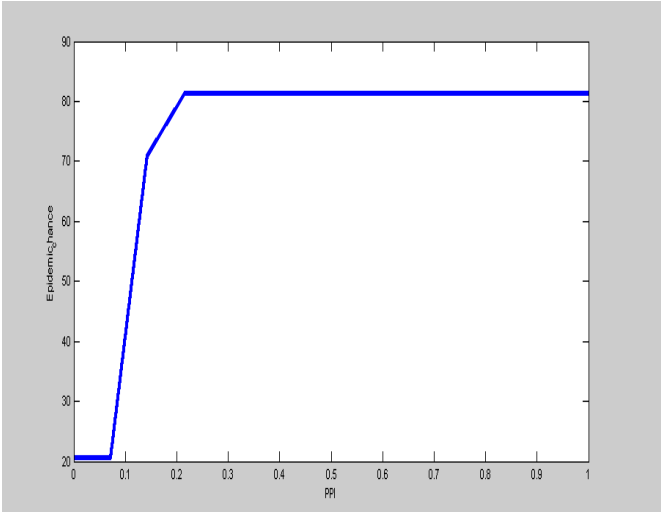

Figure 8. PPI vs. Epidemic Chance

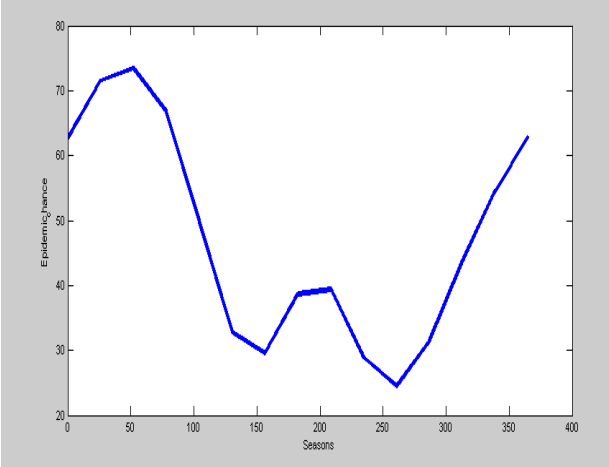

Figure 9. Season vs. Epidemic Chance

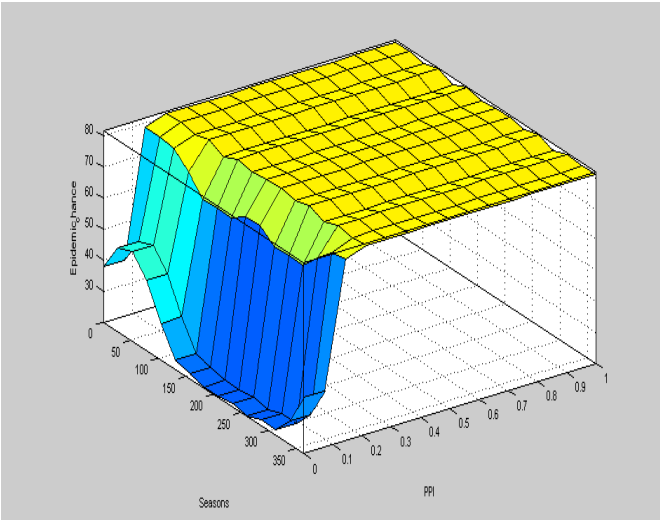

Figure 10. Season-PPI vs. Epidemic Chance

we can see how different PPI in different seasons causes epidemic according to our FIS.

Here we will see how our FIS works for the years: 1982, 1998, 2004, 2006, and 2007. Apart from 1982 all the other years' number of infected people can be found in Fig. 1. In 1982, due to a cholera outbreak during September 173,460 persons were affected with water-borne diseases [3]. Using this we get the PPI value which is 0.20. The population of Bangladesh in 1982 was 85,156,400 [1]. One more thing to note here is that, only the data up to 2007 is used to train this model, the data of the other two years are used only for the testing purpose. The results are:

- In September, 1982 as we can in Fig. 11 there is an 81.7% chance of epidemic occurrence. This is a very high chance according to our model. And we can certainly say that there was an epidemic.
- Again in September of 1998 there was an epidemic occurrence in Bangladesh [3] and our model also shows an 81.7% chance of epidemic in Fig. 12.
- Also there was a diarrhea epidemic in July 2004, according to our data set and our model in Fig. 13 also indicates that.
- The PPI value of July, 2007 is 0.1539 according to Fig. 15. And for this we get 80.7% chance of an epidemic occurrence as illustrated in Fig. 14. But in July, 2006 there were no epidemic and from out FIS we get the same result as shown in Fig. 14.

For our data set in the end we got the confusion matrix as shown in Table IV. It depicts the number of epidemic occurrences. This table gives us an accuracy of 92.5%. The epidemic threshold was chosen to be 70%.

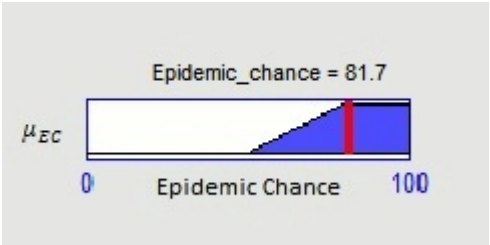

Figure 11. Epidemic Chance in September, 1982

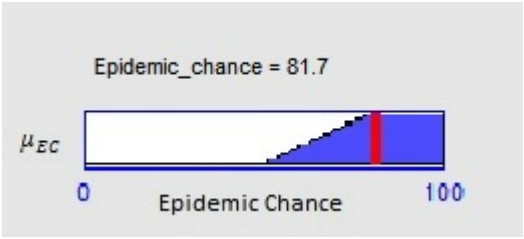

Figure 12. Epidemic Chance in September, 1998

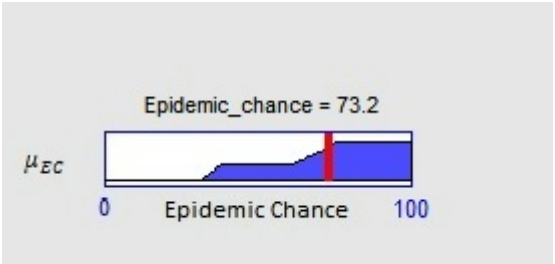

Figure 13. Epidemic Chance in July, 2004

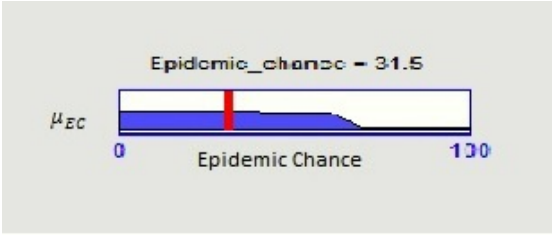

Figure 14. Epidemic Chance in July, 2006

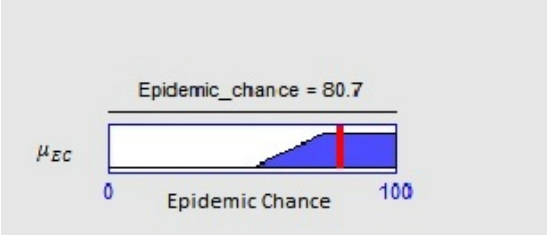

Figure 15. Epidemic Chance in July, 2007

TABLE II. Confusion Matrix

		Predicted Class	
		Yes	No
Actual Class	Yes	15	2
	No	7	96

VI. Conclusion & Future Work

Here we have identified the epidemic in whole Bangladesh. This is a really broad area to consider. In the future we hope to narrow down our research by including another parameter in our algorithm which is place. In this country we can classify it in the divisions or in districts and number death by disease. That will provide us with a more specific and better result.

ACKNOWLEDGMENT

The data set used in this paper to gain the insight using CNN was provided by IEDCR.

References

[1] *Bangladesh-population.* (n.d.). Retrieved from http://www.indexmundi.com/facts/bangladesh/population on 22 April, 2014.

[2] Centers for Disease Control and Prevention. (2012, May 18). Epidemic disease occurrence. *Principles of Epidemiology in Public Health Practice*. Retrieved from http://www.cdc.gov/osels/scientific_edu/ss1978/lesson1/section11.html on 22 April, 2014.

[3] Center for Research on the Epidemiology of Disasters, *EM-DAT: The OFDA/CRED international disaster database* Retrieved from Université catholique de Louvain, Brussels, Belgium website: http://www.emdat.be on April 22, 2014.

[4] I.M. Longini, , M. Yunus, K. Zaman, A.K. Siddique, R.B. Sack, & A. Nizam, "Epidemic and endemic cholera trends over a 33-year period in Bangladesh". *The Journal of Infectious Diseases, 186,* 246-251, 2002, doi: 10.1086/341206.

[5] M. T. Hagan, H. B. Demuth, M. H. Beale, *Neural Network Design.* New York, NY: PWS Pub. Company, 1996.

[6] C.S Rupasinghe, D.S. Gamage, C. De Alwis, M.R.M. Mufthas, R. Dabarera, "Using adaptive fuzzy systems for controlling dengue epidemic in Sri Lanka," Proceedings of *5th International Conference on Information and Automation for Sustainability (ICIAFs), 2010*, pp.459,462. doi: 10.1109/ICIAFS.2010.5715705.

[7] B.D. Stefano, H. Fuks,, & A.T. Lawniczak, "Application of fuzzy logic in CA/LGCA models as a way of dealing with imprecise and vague data." Paper presented at the 13th IEEE Canadian Conference on Electrical and Computer Engineering: IEEE CCECE'2000, May, 2000, Halifax, Nova Scotia, Canada.

[8] C. Torres, S. Barguil, M. Melgarejo, & A. Olarte, "Fuzzy model identification of dengue epidemic in Colombia based on multiresolution analysis", Artificial intelligence in medicine, 60(1), 41-51, 2014. Elsevier B.V. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24388398, retrieved on 22 April, 2014.

[9] L.H. Tsoukalas, & R.E. Uhrig, Fuzzy and neural approaches in engineering, John Wiley & Sons, Inc, 1997.