

Clinical Decision Support Model for Prevailing Diseases to Improve Human Life Survivability

Archana L. Rane

Department of MCA, K. K. Wagh Institute of Engineering Education and Research
Panchvati, Nashik - 422 003, India, E-mail: rane_nehete_archana@yahoo.com

Abstract— Constantly increasing amount of heterogeneous prevailing disease patient data can redefines medical research and clinical practice for human life survival. Computational intelligent techniques help to translate them into knowledge base that is applicable in health-care. Prediction of such diseases at early stages is biggest challenge for doctors in the country. Previous studies on prevailing diseases focus on individual diseases rather than many with similar symptom. Few of these models have constraints in finding good parameters with high accuracy. The proposed clinical decision support system in this paper models the patient's diseases state statically from his heterogeneous data to reveal the correct diagnosis by formalizing the hypothesis based on test results and symptoms of the patient before recommending treatments for the prevailing diseases. Its goal is to assist clinician in diagnosing the patient by analyzing his available data and relevant information. The proposed model designed using data mining techniques such as neural network, decision tree, statistical method, Naive Bayes, classifier and clustering pattern analysis for improving human life survivability. Several clinical data-set are used to evaluate and demonstrate the proposed model for early prediction of prevailing disease.

Keywords—Prevailing diseases, neural networks, Fuzzy logic, Decision trees, Genetic algorithms, and Statistical methods, Classifier and clustering pattern analysis

I. INTRODUCTION (Heading 1)

Prevailing diseases spread generally due to moist state of the climate and sudden transition from heat to cold or vice versa in the environment. They are attributed as vicissitudes, concurring circumstance, product fever, dysentery or liver diseases. Fever and dysentery exists in the largest proportion. The diseases do not necessarily represent the total disease burden experienced by the local population. These diseases account for more than 17% of all infection diseases causing more one million deaths annually. The risk to an individual traveler varies considerably by the specific location, visit duration, type of activities, type of accommodations, time of year, and other factors. Many of these diseases are preventable through informed protective measure if detected at early stages. It may be difficult for clinician to distinguish between these prevailing diseases due to medical symptom similarity at early stages. Large heterogeneous data set is produced for each patient with extensive monitoring and use of multiple measurement technologies. It is difficult for clinician to interpret and utilize the data set along with symptoms for effective action at early stages. Diseases like Colds-Flu-Gripe (CFG), Dengue (De), Malaria (M), Cholera (CI), Leptospirosis (L), Chikungunya (CG), Chicken pox (CP), Diarrhoea (Di)

have similar characteristics at early stage which makes clinician job difficult in diagnosing life-threatening disease. The Colds, flu is found everywhere in the world which are viral type infection and it hits the human digestive system and chest [3]. Dengue is a type of viral diseases transmitted by Aedes mosquitoes. Mosquitoes are considered one of the most dangerous creatures on the planet because of their ability to spread deadly diseases. The symptoms start about 4-7 days after being bitten. The dengue mosquito prefers to feed human at any time during the day, however, especially indoors, in shady areas, or when it is overcast. Like most viruses, there is no specific treatment. Doctors recommend acetaminophen, plenty of fluids and rest for dengue and hospitalization for hemorrhagic fever. Cholera is caused by bacteria, *Vibrio cholera*. It's transmitted from person to person by direct contact (often via healthy people carrying the disease) or consuming contaminated food and water. Endemic cholera is primarily a pediatric disease, although adult morbidity and mortality are significant, especially during epidemics. The lethality of cholera is due to the physiological consequences of rapid and profound dehydration. Malaria is caused by parasites, primarily *Plasmodium falciparum* or *Plasmodium vivax*. Female Anopheles mosquitoes pick up the parasites by feeding on infected humans. The parasite develops in a mosquito's body for 10 to 18 days, and then is passed on when the mosquito injects saliva while feeding. Once in the human body, malaria parasites migrate to the liver, where they grow and multiply. Eventually the parasites move into the blood stream to continue developing in red blood cells. As they multiply and are released, they destroy the blood cells. Leptospirosis is caused by bacteria of the genus *Leptospira*. If the disease is not treated, the patient could develop kidney damage, meningitis (inflammation of the membrane around the brain and spinal cord), liver failure, and respiratory distress; however, it is rare that death occurs. Humans become infected through contact with water, food, or soil containing urine from infected animals. It is easily treatable with antibiotics. Chikungunya is caused by a virus that is spread to people through the bite of infected mosquitoes [10]. Like Dengue, it is transmitted by Aedes mosquitoes, especially *Aedes aegypti* and *Aedes albopictus*. The incubation period is usually 3-7 days. There is currently no vaccine to prevent Chikungunya. Chicken pox is a highly contagious disease caused by primary infection with varicella zoster virus (VZV). It usually starts with a vesicular skin rash mainly on the body and head rather than on the limbs. The rash develops into itchy, raw pockmarks, which mostly heal without scarring. On examination, the observer typically finds skin lesions at various stages of healing, and ulcers in the oral cavity and tonsil areas. Chicken pox is an airborne disease

which spreads easily through coughing or sneezing by ill individuals or through direct contact with secretions from the rash. Diarrhoea is the condition of having at least three loose or liquid bowel movements each day. It often lasts for a few days and can result in dehydration due to fluid loss. The most common cause is an infection of the intestines due to a virus or bacteria or parasite, or a condition known as gastroenteritis. These infections are often acquired from food or water that has been contaminated by stool, or directly from another person who is infected. It may be divided into three types: short duration watery Diarrhoea, short duration bloody Diarrhoea, and if it lasts for more than two weeks, persistent Diarrhoea. The short duration watery diarrhea may be due to an infection

by cholera. If blood is present it is also known as dysentery. These diseases have many common symptoms as shown in Table I at early stage which makes clinician difficult to treat in few cases. The symptoms of these diseases are completely non-linear in nature. Decision support system may help in proper prediction of these diseases. Non-linear computational intelligent techniques can be used to design such systems. Data mining is successful and fastest growing fields in the computer industry to provide solution for finding useful information from large data in various fields like bio-informatics, pharmaceuticals, banking, retail, sports and entertainment etc.[2].

TABLE I. SYMPTOMS OF DISEASES AT EARLY STAGES

Disease	CFG	De	M	CI	L	CG	CP	Di
Viral, bacterial infection	√	√	√	√	√	√	√	√
Fever	√	√	√		√	√	√	
Headache	√	√	√		√	√		
Effect on eyes P=pain, Y=yellowish, R/Y=red/yellow	P	P		Y	R/Y			
Bone, muscle and joint pain	√	√		√	√	√	√	√
cough	√						√	
Bleeding		√						√
Rash		√			√	√	√	
Loss of appetite		√	√	√	√	√	√	√
Nausea/vomiting			√	√	√		√	√
Chills				√	√	√		
Dysentery			√	√	√			√
Recovery time (Days)	3-5	7-14	5-8	5-8	7-8	3-7	10-21	3-5
Criticle Symptoms	Sore Throat, Hit digestion	Decreasing platelets count, Plasmas leakage, Shock evidence, Weak pulse	Coma	Liver damage, Decreasing platelets count	Kidney/liver failure, Meningitis, Respiratory distress	Swelling		Risk of dehydration

A. Related Work

Data mining is an interdisciplinary sub-field of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [1]. The legacy system is supposed to use to store the large data which is tedious to maintain. It is challenge to understand, integrate and put on the various methods to utilize and discover the knowledgeable data from the large data. Researchers are starved for discovery of knowledge from this morass of data to predict future trends and to make better decisions in science, industry, and markets. Data mining systems are composed of data pre-processing, knowledge discovery models, and a data-concept description. The aim to use data mining technique is usually enables one to collect, store, access, process and ultimately describe and visualize data sets. Lot of work has been done on classification of dengue data by using different data mining techniques [4]. In [4], a hybrid model is designed that uses genetic algorithm for the determination of weight in neural network model. This model predicts dengue accurately by adjusting parameters to achieve optimal prediction performance. The dengue notification system discussed in [5] notifies the patient whether they are infected with dengue fever

or not by using a fuzzy logic approach. The application of machine learning techniques discussed in [6] makes a distinction between dengue and other feverish illnesses in the primary care setting and predicts severe arboviral disease among population.

B. The Proposed System

We have designed and developed clinical decision support system for prevailing diseases to improve effective action taken by the clinician by providing real-time decision support based on non-linear data mining techniques such as neural networks, fuzzy logic, decision trees, genetic algorithms, and statistical methods, classifier and clustering pattern analysis. Prediction of diseases like Colds-Flu-Gripe, Dengue, Malaria, Cholera, Leptospirosis, chikungunya, Chicken pox, Diarrhoea which has similar characteristics at early stage is the main intention of this work. The proposed model assists the clinician in diagnosing the patient by analyzing his available data and relevant information about life-threatening prevailing disease. The contributions are as follows: (1) Collection of patient data set from expert doctors on these diseases available in Nasik. (2)Dividing the collected data set into training set for the data mining technique training and testing set for testing the performance of the corresponding technique. (3) Application of

various data mining techniques like neural networks, fuzzy logic, decision trees, genetic algorithms, and statistical methods, classifier and clustering pattern analysis (4) Analyzing the result for various parameters.

II. CLINICAL DECISION SUPPORT SYSTEM

The proposed support system for prevailing diseases is designed using data mining techniques such as neural networks, fuzzy logic, decision trees, genetic algorithms, and statistical methods, classifier and clustering pattern analysis. Total 316 patient data sets are collected from expert doctors. Each patient data set has 34 attributes collected from discharge summaries of patients that include patient demographic details, admission reasons, discharge details, lab tests outcome, and complaints on reporting, course of treatment. The attributes includes age, gender, infection, viral, bacterial, degenerative, class of people, climate, hygienic level, type of disease, coughing, fussiness, temperature, high temperature, duration of fever, platelets, platelets count, headache, pain behind eyes, bone/muscle/joint pain, seriousness, sudden drop in blood pressure, bleeding, organ damage, rash, treatment period, transform from person to person, loss of appetite/taste, feeling of nausea/vomiting, breathing difficulties, pain in the chest, dysentery, and chills. Data mining technique are applied on the collected attribute of a patient to support the decision for clinician in prediction of survivability of the human life at early stages. The collected patient data set is divided into 2 categories to measure the accuracy of the model as follows. (1) The training data set which contains 190 records used to train the corresponding data mining technique. (2)The testing data set which contains 126 records used to test the performance of the corresponding data mining technique using open source software, WEKA [9]. The selected classification algorithms[8] are Decision tree, Nave Bayes, multilayered perception, K-nearest neighbor and Support vector machine. The attributes are ranked based on priority using ranking algorithm available in Weka tool.

We used 10 fold cross validation in experiment. The data set is divided into 10 parts. Nine parts of data are used as training data and remaining part is used as test data.

A. Experimental Data

For the experimental setup, the collected medical data-sets are entered in to arff file format required in WEKA. All the identified algorithms are tested on the medical dataset. In this study, patients are over 11 years old. For patients, who are younger than 11 years old will be considered as children patients. Too young patients are not able to explain their symptoms to the doctors. The information related to symptoms of these patients is mainly dependant on their parents or guardians. Therefore, only adult patients will be considered in the experiment.

The data mining techniques such as neural networks, fuzzy logic, decision trees, genetic algorithms, statistical methods, classifier and clustering using WEKA are investigated for its usage before this study. WEKA is a well-known machine learning software. The number of parameters relating to training data set is analyzed through many experiments to find

the optimal result like multilayer perceptron algorithm (MLP), J48 algorithm, k-means algorithm, Nave Bayes classifier, Knearest neighbor and Support vector machine.

III. RESULT DISCUSSION

Recently Receiver Operating Characteristic (ROC) curves[11] are described as another way of performance measurement of computational intelligence diagnosis system with one output. It quantifies the accuracy of a computational intelligence diagnosis system by comparing the decisions or classifications. The result of ROC curve is not sensitive to the probability distribution of training and testing or to decision bias [7]. ROC curve is plot of true positive ratio (also called as sensitivity) vs false positive ratio (computed as (1-specificity)) which is computed using four decisions in the Contingency matrix. Figure 1 shows ROC curve for Dengue using multilayer neural network.

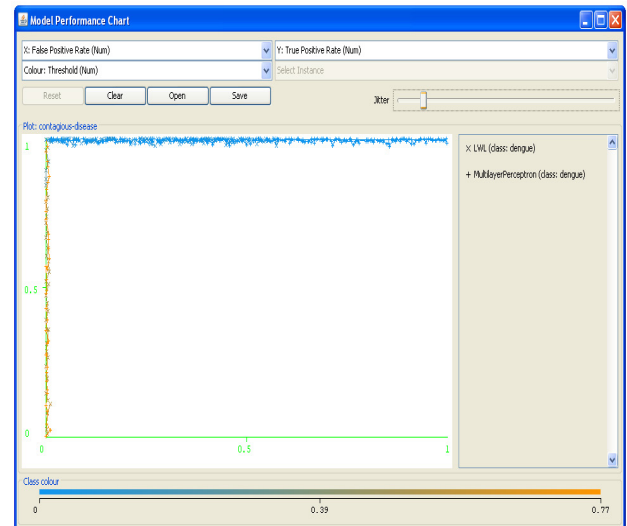


Figure 1 ROC curve obtained using WEKA for Dengue

The obtained ROC curve indicates the each output response is close to 1 for most of the test sample. We can conclude that the performance is perfect as all predictions falls towards true positive ratio.

When the computational intelligence system has multiple output classes, the confusion matrix is used as performance metric which contains information about actual and predicted result about the classification system. An $n \times n$ metric is constructed for a system comprising n -classes. Row reflects standard classification whereas column reflects classifications made y the computational intelligence system. Positions along main diagonal of the matrix indicates the correctly classified instances in test set while remaining positions indicates misclassified number of instances of row class in column class. Performance of such systems is commonly evaluated using the data in the matrix as shown in Table II.

The notations used in this matrix are a for Colds Flu and Gripe, b for chickenpox, c for dengue, d for diarrhea, e for Chikungunya, f for Malaria, g for Leptospirosis, and h for Cholera. The value of the matrix has two indices X_{ij} indicates the number of sample data set which has of the disease i as well

as disease j where i = a, ..., h and j = a, ..., h. The row represents the actual diagnosis of the disease and column represents predicted diagnosis of the disease on the patient. For Example (1) aa indicates actual and predicted disease both are the Colds Flu and Gripe and the prediction is correct. (2) ab indicates actual is Colds Flu and Gripe but predicted is chickenpox and prediction is wrong. A classifier is chosen by selecting percentage split. The default ratio is 66% for training and 34% for testing. This experiment uses the ratio 60:40. The right pane in Figure 2 shows the results for training and testing.

TABLE II. STRUCTURE OF CONFUSION MATRIX

		Prediction Results							
		a	b	c	d	e	f	g	h
Diagnostic Results	a	aa	ab	ac	ad	ae	af	ag	ah
	b	ba	bb	bc	bd	be	bf	bg	bh
	c	ca	cb	cc	cd	ce	cf	cg	ch
	d	da	db	dc	dd	de	df	dg	dh
	e	ea	eb	ec	ed	ee	ef	eg	eh
	f	fa	fb	fc	fd	fe	ff	fg	fh
	g	ga	gb	gc	gd	ge	gf	gg	gh
	h	ha	hb	hc	hd	he	hf	hg	hh

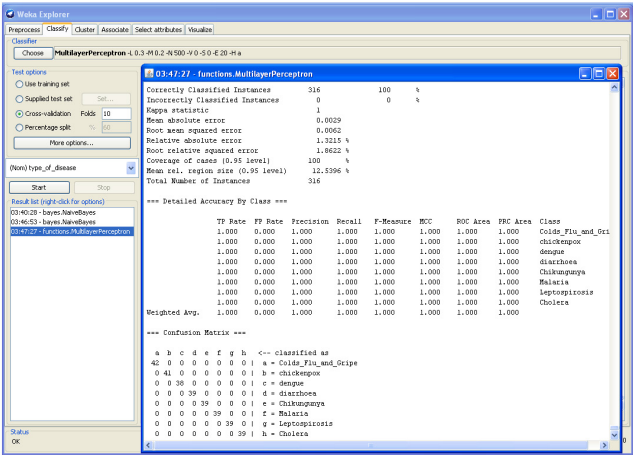


Figure 1Results of trainingusing MLP in WEKA

TABLE III. STRUCTURE OF CONFUSION MATRIX USING SVM

		Prediction Results							
		a	b	c	d	e	f	g	h
Diagnostic Results	a	12	0	0	0	0	0	0	0
	b	0	19	0	0	0	0	0	0
	c	0	0	11	0	0	0	0	0
	d	0	0	0	16	0	0	0	0
	e	0	0	0	0	19	0	0	0
	f	0	0	0	0	0	11	0	0
	g	0	0	0	0	0	0	17	0
	h	0	0	0	0	0	0	0	20

It also indicates other performance parameters like confusion matrix, the number of correctly classified and misclassified samples. The generated confusion matrix for SVM, Naïve Bays, decision tree, ANN, KNN data mining technique from WEKA is as shown in Tables III, IV, V, VI respectively. Figure 2 shows the results of training the

multilayer perceptron using WEKA for one technique. The combined result of all the technique is given in Table VII.

TABLE IV. STRUCTURE OF CONFUSION MATRIX USING NAVIE BAY

		Prediction Results							
		a	b	c	d	e	f	g	h
Diagnostic Results	a	11	0	0	0	0	0	0	0
	b	0	19	0	0	0	0	0	0
	c	0	0	11	0	0	0	0	0
	d	0	0	0	16	0	0	0	0
	e	0	0	0	0	19	0	0	0
	f	0	0	0	0	0	11	0	0
	g	0	0	0	0	0	0	17	0
	h	0	0	0	0	0	0	0	20

TABLE V. STRUCTURE OF CONFUSION MATRIX USING DECISION TREE

		Prediction Results							
		a	b	c	d	e	f	g	h
Diagnostic Results	a	42	0	0	0	0	0	0	0
	b	0	39	0	0	0	0	0	0
	c	0	0	38	0	0	0	0	0
	d	0	0	0	38	0	0	0	0
	e	0	0	0	0	39	0	0	0
	f	0	0	0	0	0	39	0	0
	g	0	0	0	0	0	0	39	0
	h	0	0	0	0	0	0	0	39

TABLE VI. STRUCTURE OF CONFUSION MATRIX USING ANN AND KNN

		Prediction Results							
		a	b	c	d	e	f	g	h
Diagnostic Results	a	42	0	0	0	0	0	0	0
	b	0	41	0	0	0	0	0	0
	c	0	0	38	0	0	0	0	0
	d	0	0	0	39	0	0	0	0
	e	0	0	0	0	39	0	0	0
	f	0	0	0	0	0	39	0	0
	g	0	0	0	0	0	0	39	0
	h	0	0	0	0	0	0	0	39

TABLE VII. PERFORMANCE PARAMETERS

Technique		DT	ANN	KNN	SVM	NB
Algorithm used		J48	MLP	LWL	SMO	Naive Bays
Summary of result analysis	Correctly classified instances	313	316	316	125	124
	Incorrectly classified instances	3	0	0	1	2
	Kappa Statistics	0.989	1	1	0.9909	0.9817
	MAE	0.004	0.0029	0.1381	0.1877	0.0041
	RMSE	0.049	0.0062	0.2376	0.2913	0.0511
	Relative absolute error	1.68	1.3215	63.152	85.4572	1.8825
	Root relative squared error	14.8	1.8622	71.8441	87.59	15.372
	Coverage of cases (%) (0.95 level)	99.05	100	100	100	100
	Mean rel. region size (%) (0.95 level)	13.29	12.5396	44.106	76.8849	13.0952
	Total Number of Instances	316	316	316	126	126

There are three commonly used performance measurements including accuracy, sensitivity and specificity. The accuracy of data mining technique is the percentage of correctness of outcome among the test sets exploited in this study as defined in Equation 1. It is the ratio of correctly classified instances and total number of instances in test set in Table 3. All data mining techniques gives 100% accuracy in most of the cases. Incorrectly classified instances are calculated as total instances considered minus correctly classified instances.

$$Accuracy = \frac{\sum_{i=a}^h X_{ii}}{\sum_{i,j=a}^h X_{ij}} \quad (1)$$

Kappa statistics computed on two set of categorized data as degree of agreement between them and its value varies in interval from zero to one. Higher value indicates stronger agreement between two set of categorized data. Zero value indicates no agreement and one indicate perfect agreement. Higher value of Kappa statistics is expected for outstanding result using data mining technique. The value of kappa statistics using the proposed model is very high as given in the Table VII. ANN and KNN have perfect agreement as its Kappa statistic value is one.

Absolute error is the difference between the predicted value and the actual value i.e. $Error_i = DiagnosisResult_i - PredictedResult_i$ where $DiagnosisResult_i$ is observed value, $PredictedResult_i$ is predicted value. Mean Absolute Error (MAE) is computed for each data mining technique as sum of all absolute value of error divided by number of predicted set N as given in Equation 2.

$$MAE = \frac{1}{N} \sum_{i=1}^N (|Error_i|) \quad (2)$$

MAE in most of the prediction technique is very small and close to zero which is indication of ideal/acceptable performance. The MAE = 0.0029 is obtained using ANN that is minimum among all the techniques for normalized data set. SVM gives maximum RMSE = 0.1877 which shows its inefficiency in prediction of life-threatening disease.

Root Mean Square Error (RMSE) is defined as square root of the difference between observed value and predicted value by computing the average sum of the squared error with ideal performance yielding zero RMSE. It can be computed as shown in equation 3.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Error_i)^2} \quad (3)$$

RMSE in most of the prediction is very small and close to zero which is indication of ideal/acceptable performance.

Occasionally, RMSE is more due to improper description of symptom by patient. The RMSE = 0.0062 is obtained using ANN that is minimum among all the techniques for normalized

data set which can be neglected and can be improved further in future. SVM gives maximum RMSE = 0.2913 which shows its inefficiency in prediction of life-threatening disease.

IV. CONCLUSION

Prevailing diseases spread generally due to moist state of the climate and sudden transition from heat to cold or vice-versa in the environment. Prediction of such diseases at early stages is biggest challenge for doctors in the country due to similarity in symptoms. The proposed clinical decision support system described in this paper is designed using various data mining techniques such as neural network, decision tree, statistical method, Naive Bayes, classifier and clustering pattern analysis for improving human life survivability related to the life threatening diseases like Colds-Flu-Gripe, Dengue, Malaria, Cholera, Leptospirosis, Chikungunya, Chicken pox, Diarrhoea which have similar characteristics at early stage. This is novel approach for clinical decision support considering many diseases at the same time analyzing its performance through many data mining techniques. It is observed from the obtained performance parameters values that ANN outperforms in this regards while SVM gives lowest acceptance result among the techniques chosen but the accuracy of all provided almost same performance.

REFERENCES

- [1] Han and M. Kamber, "Data mining: concepts and techniques," 2nd ed. San Francisco: Morgan Kaufmann, Elsevier Science, 2006.
- [2] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine Learning Tools and Techniques," Morgan Kaufmann, San Francisco, 2005.
- [3] World Health Organisation, Available: <http://www.searo.who.int/en/SectionIOI>
- [4] Nor Azura Husin, Norwati Mustapha, Md. Nasir Sulaiman, and Razali Yaakob, "A Hybrid Model using Genetic Algorithm and Neural Network for Predicting Dengue Outbreak," 4th Conference on Data Mining and Optimization (DMO), September 2012, Langkawi, Malaysia
- [5] Tajul Rosli Bin Razak, Muhammad Hermi Ramli and Rosmawati Abd. Wahab "Dengue Notification System using Fuzzy Logic," International Conference on Computer, Control, Informatics and Its Applications, 2013
- [6] Shameem A. Fathima and Nisar Hundewale, "Comparative Analysis of Machine learning Techniques for classification of Arbovirus," International Conference on Biomedical and Health Informatics, Hong Kong and Shenzhen, China, Jan 2012
- [7] Russell C. Eberhart, and Yuhui Shi, "Computational Intelligence Concepts to Implementations," Morgan Kaufmann of Elsevier, 2007.
- [8] Sofianita Mutalib, Nor Azlin Ali, Shuzlina Abdul Rahman and Azlinah Mohamed, "An Exploratory Study in Classification Methods for Patients Dataset," International Conference on Data Mining and Optimization, 2009, pp. 79-83.
- [9] M. H. E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An update," Spec. Interest Grp. On Knowl. Disc. and DM, vol. 11, June 2009.
- [10] Chakkaravarthy, V.M., S. Vincent and T. Ambrose, "Novel Approach of Geographic Information Systems on Recent outbreaks of Chikungunya," Journal of Environmental Science Tech. vol. 4, no. 4, pp. 387-394.
- [11] <http://www2.cs.uregina.ca/dbd/cs831/notes/ROC/ROC.html>