IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-03, Issue-09 E-ISSN: 2347-2693

Expert System to Predict the Type of Fever Using Data Mining Techniques on Medical Databases

M.V.Jagannatha Reddy[1*] and B.Kavitha[2]

[1*]CSE Dept. Madanapalle Institute of Technology & Science. Madanapalle. Chittoor(dt). A.P. INDIA
[2]Dept. of Computer Science. Govt. Degree College, Srikalahasti. Chittoor(dt). A.P.-INDIA

www.ijcseonline.org

Received: Aug/12/2015 Revised:Aug/28/2015 Accepted:Sep/20/2015 Published: Sep/30/2015

Abstract----By finding the most important medical symptoms and laboratory data helps in building an expert system to predict the dengue fever in early stages. We developed in this project a new expert system to predict the dengue fever in early stages. This methodology consists of three important steps: a) manual missing value imputation method is applied that makes the data consistent. b) An expert doctors opinion is taken for selecting most influential attributes for dengue fever also we done internet survey . c) A neural network model is used for accurate prediction of dengue fever. The expert system is developed using MATLAB 2013. This methodology is seems to be giving good predictive results compared to other techniques.

Key words: Dengue Fever, Expert system, Neural Network, Prediction.

INTRODUCTION

Dengue fever is disease transmitted by mosquitoes[1] and causing sudden high fever and pains in the joints. Also known as break bone fever. The first case is detected in the philiphines in 1953, the disease is identified as one of the most dangerous disease in the humans[2]. Accurate prediction of disease is possible only after several tests of laboratory and clinical symptoms.

A multi variant model was constructed for predicting hemoglobin using predictors i.e., they have used various attributes such as vomiting sensation , weight, sex and other factors. These techniques are used only after two to twelve days from the day of illness.

The world health organization is made classification for identifying affected individual persons based on the laboratory and clinical symptoms. The models developed for the diagnosis of dengue fever is affected by missing values and influential features. This may be due to incorrect data entry or not collected properly at the time of data collection.

In order to ovoid incorrect prediction, we followed procedure as below

1) Missing values are filled manually with appropriate normal values
2) In order to ovoid too many attributes for analysis, we took advice from different expert physicians. So that only most influential attributes are collected.
3) After data preprocessing, we used MATLAB 2013a for accurate prediction of dengue fever using neural networks.

2. RELATED WORK

2.1: - Using Decision Tree:

In this method[3] they said Dengue infection is a disease typically found in hot and sticky region. The doctors need to understand the features on dengue infection in order to correctly categorize the patients, since these patients require different treatment. Their dataset consists of clinical and laboratory data. The data was collected from the first visit of patient to hospital until the date of discharge. They obtained two sources of datasets from different regions of Thailand, which are Srinagarindra Hospital and Songklanagarind Hospital. These datasets consists of more than 400 attributes. They used decision tree as a data mining tool. They propose a set of meaningful attributes from the temporal data. Their experiments are divided into four parts. In all four experiments they use decision trees. The first two experimental results show the useful knowledge to classify dengue infection from Srinagarindra Hospital's dataset and Songklanagarind Hospital's dataset, respectively. Each set of knowledge is tested by different dataset to make sure that the test data was a real unseen data. The third experimental results show the useful knowledge when they integrated two datasets. Another objective of this research is to detect the day of defervescence of fever which is called day0. The day0 date is the critical date of dengue patients that some patients face the fatal condition. Therefore the physicians need to predict day0 in order to treat the patients. They expect to have an intelligent system that can trigger the day0 date of each patient. They set up four experiments. In the first three experiments, they find

knowledge in order to classify type of dengue infection. For forth experiment, they tried to predict the day of effervescence with the data before day0 date. They applied decision tree approach to all experiments. Note that they use sensitivity, specificity and accuracy as performance measures. Their approximate accuracy of all four experiments using decision tree is around 96.5%. In another decision tree algorithm method researchers Tanner, et al and Tarig, et al [5]. They classified 1,200 patients using decision tree approach. They found six important features and they got 84.7 % correctness .

2.2 Genetic Algorithm and ANN :

Predicting the dengue fever can be done by different methods one of the method described by[4] identifying the important medical symptom and laboratory information without a medical expert. In their research effort an intelligent based technique that identifies the diagnosis of dengue fever is proposed. They followed three major steps they are: (1).A method to enter absent input values that contains diverse dataset.(2).Genetic algorithm for attribute selection for sorting out a subset of mainly important symptom that can identify the disease. (3).An artificial neural network method that employs back-propagation method for increasing the accuracy of predicting the dengue fever. The researchers say that this method reduces the number of false prediction and increases the correctness of predicting dengue fever compared to decision trees.

2.3: Self organizing map and multi layer FFNN :

A combination of the self-organizing map and multilayer feed-forward neural networks was employed for the risk prediction of dengue patients in the Tarig's research. They clustered patients into two groups which are low risk and high risk using tree criteria [6]. They used only examples from Day0 until Day2.they got 70% of correctness. Fatimah Ibrahim et al. [7] predicted the day of defervescence of fever (day0) from 252 dengue patients (4 DF and 248 DHF). Researchers used Multi- Layer Perceptrons and got 90% correctness.

2.4: - Multiple Linear Regressions:

In this method [8], they done research to find out whether the atmospheric things can be applied to classify the annual dengue affected persons of Dhaka city of Bangladesh. They got the monthly dengue affected cases and atmospheric data for 2000-2008 from DGHS(directorate general of health services).MDD(meteorological department of Dhaka) of Bangladesh. And data for the period of 2000-2007 is used to build a model using multiple linear regressions. Validation of the model is done using 2001,2003,2005 and 2008. Normal monthly humidity, rainfall, minimum and maximum temperature is used as autonomous variables and amount of dengue belongings report monthly is used as dependent variable. correctness of the model for classifying disease was find through ROC (receiver operative characteristics)curve. Atmospheric attributes like, rainfall, maximum temperature and relative humidity were considerably connected with monthly report dengue cases. The model containing atmospheric data of two lag month explain 61 per cent of difference in values of report dengue belongings and this system was found to classify dengue disease (≥ 200 belongings) through normal accuracy [area in ROC curve = 0.89, 95% CI = (0.89-0.98)]. The classification system has some disadvantage in classifying the monthly cases of dengue blongings.

2.5 –New Fuzzy Association Rule Mining:

In this method [9], they explain a novel prediction technique utilizing Fuzzy Association Rule mining to get relationships among climatic, meteorological, medical, and socio political data from Peru. These associations are in the form of rules. The best set of rules is automatically select and forms a classifier. This classifier is then used to predict future dengue occurrence as either HIGH (found) or LOW (not found) Results: Their automatic technique built three dissimilar fuzzy association rule systems . Using the first two weekly systems, they predicted dengue occurrence three and four weeks in advance, correspondingly. The third prediction is a four week duration, specially four to seven weeks from time of prediction. By means of previously unused test data for the period 4–7 weeks from time of prediction gives a positive predictive value of 0.686, a negative predictive value of 0.976, a sensitivity of 0.615, and a specificity of 0.982.

2.6- Wrapper-Based Attribute Selection and Decision Tree:

In this method [10], they develop a novel method that predict the finding in real time, that minimize the amount of wrong positive and wrong negative values. Their method includes three main steps:
(1) They used for inputting missed values new method that can be used on any data consists of mixed dataset (2) Wrapper based attribute selection techniques are used to obtain a influential features. That could predict the disease (3) Decision Tree model is used to produce rules. They said their predictive models developed are to be more precise than the other methods applied in the finding dengue fever.

2.7-Time Series Poisson Multi Variate Regression Model:

In this method [11], they developed and authenticate a predictive method that can predict dengue belongings and give early caution sign in Singapore. The time series Poisson multivariate regression method is developed by weekly average temperature and collective rainfall for the duration 2000–2010. Climate information is also collected using piecewise linear spline functions. They analyze various lag times among dengue and weather variables to predict the best dengue forecasting period. Auto regression, trend and seasons is measured in the method. They tested the method by guessing the dengue cases for week one of 2011 up to week sixteen of 2012 using weather data only. The best duration for dengue forecast was sixteen weeks. Their method estimated correctly with errors of 0.3 and 0.32 of the standard deviation of positive cases during the model building and justification time, correspondingly. In the conclusion they said differentiate among occurrence and non occurrence to a 96% (CI = 93–98%) in 2004–2010 and 98% (CI = 95%–100%) in 2011. The method forecast the occurrence in 2011 exactly with less than 3% likelihood of false alarm.

They have constructed a weather based dengue fore-casting method that gives caution sixteen weeks in early of dengue disease with more precision. They said that models using temperature and rainfall could be simple and price efficient measures for dengue fore-casting.

2.8-Artificial Neural Network Model Using Humidity, Temperature and Rainfall:

This research [12] is done by generating the patterns for dengue disease using Artificial Neural Networks They collect the real data from Singaporean National Environment Agency (NEA). This data was used to model the manners of dengue patients based on the physical attributes of mean relative humidity, mean relative temperature and total rainfall. The data collected weekly consists of dengue positive cases for a period of six year , January 2001 to April 2007.

The Neural Networks model developed produces the results in 1000 epochs in a few seconds. The network is also used to predict the 2005 occurrence. The correlation coefficient for the year 2005 period is 0.70 . There was a drop in correlation to 0.70 from 0.76. The problem in prediction seen between weeks 34 and 38.

Method: The Artificial Neural Network model with Back Propagation algorithm is used in this research [12]. The dataset consists of 330 weekly measurements i.e., sets of mean relative humidity, mean temperature,total rainfall and the total number of dengue positive cases. Then it is divided into two different data subsets called training dataset and test dataset, respectively. The train data is used in developing the neural network model and test data is used to test the model with train data. The training dataset consists of 104 weeks data for two years. For testing the model remaining 226 weeks data is used. The artificial neural network model used had three layers i.e., input layer, hidden layer and output layer. They used three parameters, they are mean relative humidity, mean relative temperature and total rainfall. The ANNs output is the number of dengue confirmed cases. The fig below shows the design of ANN. They said that network done well in predicting the dengue cases except for the year 2005.

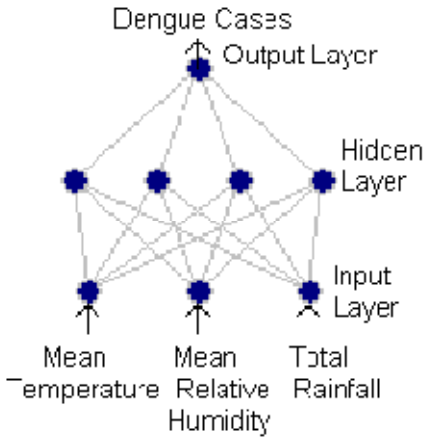


Fig: Artificial neural Network Model

This ANN model gives the correlation coefficients of 0.84 for training dataset and 0.76 for test dataset.

EXISTING SYSTEM

1) In the existing methods for missing values they used automated data mining missing value imputation techniques in all the methods explained above. These techniques may fill approximate or wrong values in many cases. This will affect the final results.

2) In the existing methods for feature selection they used algorithms. This technique also may choose less important attributes. So this makes the processing time increase. Also it may affect final results.

3) By using the above techniques they used Artificial neural networks(ANN) based on Humidity, rainfall and temperature.

PPROPOSED SYSTEM

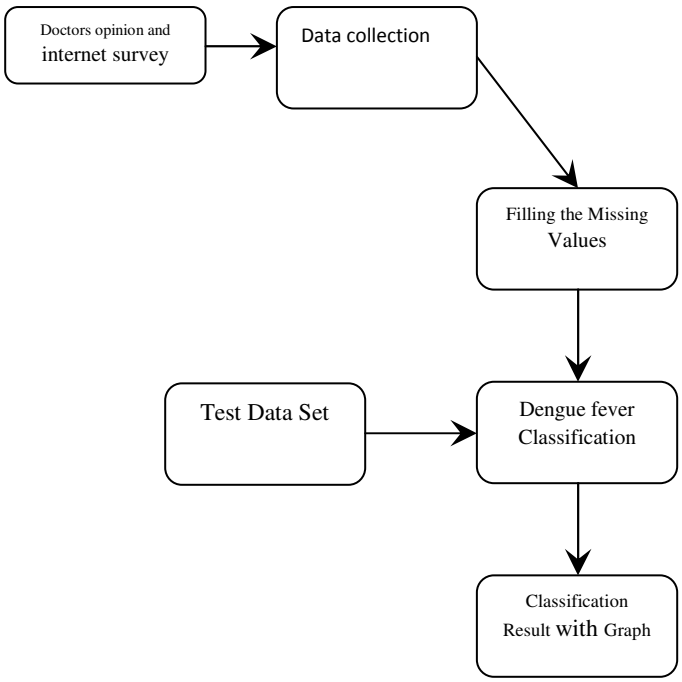
We propose A new expert system for predicting dengue fever. Our methodology consists of three major steps.

1) A manual missing value imputation method is used. This reduces the false value entry. So that our results will improve marginally.

2) For selecting the most influential attributes that predict the dengue fever we took expert doctors opinion and internet survey. This process reduces collecting unnecessary attributes during data collection. This helps in accurate prediction of dengue fever.

3) After preprocessing the data we use neural networks for predicting dengue fever. This will be implemented by using MATLAB 2013a.
So as we are expected this method gave accurate results as explained in the implementation section.

System Architecture:



Implementation:

1. Missing value filling
2. Data Representation
3. Disease prediction

1. Missing Value filling:
We visited to Hyderabad for collecting data in various hospitals. During this period we interacted with some of the doctors to collect opinion and to finalize most influential attributes to predict dengue cases. We collected data through manually and some of the reports from patients. During this period we interacted with patients to collect the information related to clinical data such as headache, vomiting body pains etc.,.
Data collected is tabulated and missing values are filled manually with appropriate normal values. This makes to get the prediction values more accurate.

2. Data Representation
The data collected is of mixed data. That is it includes categorical data and numerical data. So for the data to be used in MATLAB it should numerical data for neural networks. So we converted the categorical data to numeric data set by replacing, if the attribute value is YES then it is replaced by 1. If the attribute value is NO then the value is replaced by 0. Remaining numerical values kept untouched.

The dengue fever dataset collected can be used to predict the new dengue fever case. This project is implemented by using MATLAB 2013a.

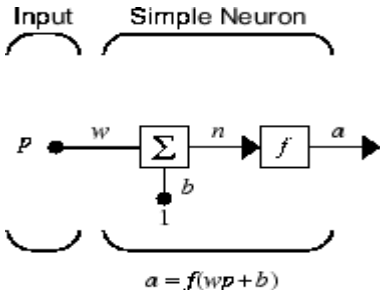
Work flow for Neural Network [13]

The Basic steps for neural network design are,

1. Collecting required data
2. Develop the network
3. Organize the network
4 Then initialize the weights
5 Using known dataset the network is trained
6. Confirm the network using test data
7 finally we can use the network for new dataset

Neuron Model:

Simple neuron is as follows



In the above example there are three different functions takes place. p is the first function, it is scalar input function which is multiplied by scalar weight w which produces wp scalar product. The second function is wp is added to scalar bias b(wp+b) to obtain net input n. At last the net input is passed via transfer function f. This produces the output a.

3. Disease prediction

Algorithm:

Input: df_dataset
Output: performance of neural network
Step 1:Perform data cleaning
Step 2:Do data transformation
Step 3:select dataset to input of NN
Step 4:select target data for network output
Step 5: Train the network
Step 6:Validate and test the network
Step 7: plot the ROC and Confusion curves

Validation and test data:

In this step our dataset consists of 203 samples are divided into three samples training samples, validation samples and testing samples. 143 samples (70%) are used for training the neural network and the remaining 15% each i.e., 30 samples are used for validation and testing purpose.
Training: Training samples are presented to the network during training and the network is adjusted according to its errors.
Validation phase: In this phase samples are used to network generalization, and to halt training when generalization stops improving.
Testing phase: In this phase samples have no effect on training network and so provide an independent measure of neural network performance during and after training.

Train Network:

Train the neural network to classify the inputs according to the targets. Training automatically stop when generalization stops improving as indicated by an increase in the mean square error of the validation samples.

Results:

	Samples	MSE	%E
Training	143	$2.49422e^{-2}$	$2.79720e^{-0}$
Validation	30	$6.52483e^{-2}$	$6.66666e^{-0}$
Testing	30	$2.10199e^{-3}$	0

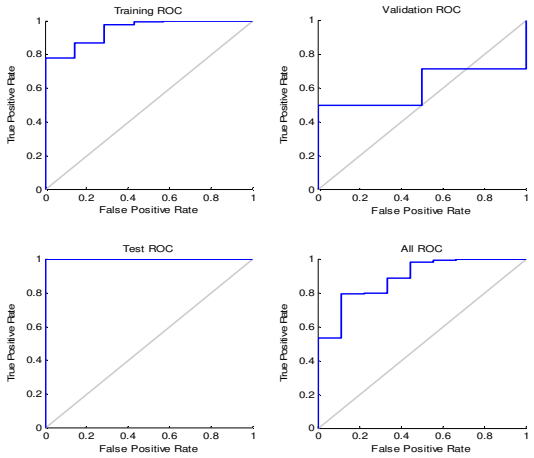
MSE: Mean Square Error is the average squared difference among outputs and targets. If the MSE value is less classifier accuracy is good.
Percentage Error: percentage error indicates the part of datasets that are misclassified. A value of 0(zero) means that no misclassification. 100 indicate highest misclassification.

Test Network Results:

MSE:	$2.75234e^{-2}$
%E:	$2.95566e^{-0}$

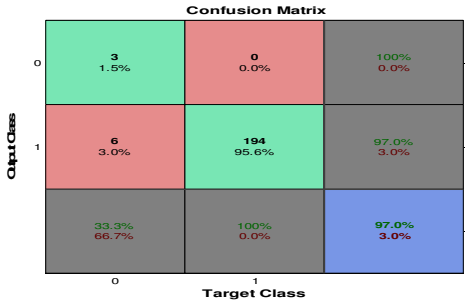
ROC Curves:

In the receiver operating characteristic is a measure used to ensure the quality of classifiers. For each class of a classifier, roc apply threshold values across the interval [0,1] to outputs. For each threshold, two values are calculated, the True Positive Ratio (the number of outputs greater or equal to the threshold, divided by the number of one targets), and the False Positive Ratio (the number of outputs less than the threshold, divided by the number of zero targets).ROC curves is also called relative operating characteristics [14] because it is a discrimination between two characteristics. i.e, True positive rate and false positive rate.

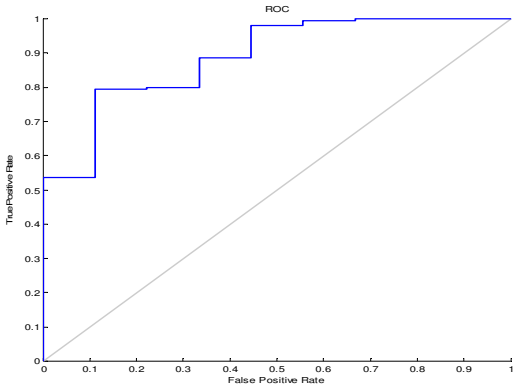


Confusion matrix: Confusion matrix or it is also called error matrix[14]. This matrix gives the performance of the algorithm or model. In this each column represents instance in a predicted class while each row represents instance in real class and vice versa.

Test confusion:



Test ROC:



Conclusion and Future Enhancement

A Neural network model developed using matlab is tested to predict the dengue fever cases. This model is used to generate performance curves, ROC curves, confusion curves for both training and test data these analysis shows that this methodology is good than the other methodologies in diagnosing the dengue fever. This model gives an accuracy of 100.0% in children and adults using both clinical and laboratory features. Based on the performance of the model we conclude and recommend that this neural network model can be used to build an expert system to predict the new dengue cases in the early stages.

In future this model can be extended to predict any type of fever like malaria, typhoid, viral fever etc.based on the clinical and laboratory reports

Acknowledgements:

This project is supported by UGC-SERO Hyderabad under Minor Research Projects. Grant no: MRP-4608/14(SERO/UGC) for the year 2013-14.

REFERENCES

[1] D.J. Gubler, "Dengue and dengue hemorrhagic fever," *Clin. Microbiol. Rev.*, vol. 11, pp. 480–496, 1998.

[2] T. P. Monath, "Dengue: The risk to developed and Developing countries," *Proc. Nat . Acad. Sci. USA*, vol. 91, no. 7, pp. 2395–2400, 1994.

[3] Thitiprayoonwongse, Prapat Suriaphol and Nuanwan Soonthornphisaj Data Mining of Dengue Infection Using Decision Tree Daranee Latest Advances in Information Science and Applications, ISBN: 978-1-61804-092-3

[4] Revathi N. Prof.S.J.K.Jagadeesh Kumar Genetic lgorithm Optimization And Neural Network For The Diagnosis of Disease. International Journal of Computer Applications & Information Technology Vol. II, Issue I,. January 2013 (ISSN: 2278-7720)

[5] L. Tanner, M. Schreiber, J.G. Low, A. Ong, T. Tolfvenstam, Y.L. Lai, L.C. Ng, Y.S. Leo, L. Thi uong, S.G. Vasudevan, C.P. Simmons, M.L. Hibberd and E.E. Ooi, Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness,PLoS Neglected Tropical Disease, Vol.2, 2008.

[6] T. Faisal, F. Ibrahim and M.N. Taib, A noninvasive intelligent approach for predicting the risk in dengue patients, Expert Systems with Application,Vol.37, No.3, 2010, pp. 2175-2181.

[7] F. Ibrahim, M. N Taib, W. A. B. Wan Abas, C. G. Chan and S. Sulaiman, A novel dengue fever (DF) and denguehaemorrhagic fever (DHF) analysis using artificial neural network (ANN), Computer Methods and Programs in Biomedicine, No.79, 2005, pp. 273-281.

[8] Md. Nazmul Karim, Saif Ullah Munshi*, Nazneen Anwar & Md. Shah Alam**. "Climatic factors influencing dengue Cases in Dhaka city: a model for dengue prediction". Indian J Med Res 136, July 2012, pp 32-39.

[9] Anna L Buczak*, Phillip T Koshute, Steven M Babin, Brian H Feighner and Sheryl H Lewis Buczak et al "A data- Driven epidemiological prediction method for dengue outbreaks using local and remote sensing data." BMC Medical Informatics and Decision Making 2012, 12:124

[10] Vadrevu Sree Hari Rao, Senior Member,IEEE, and Mallenahalli Naresh Kumar "A New Intelligence-Based Approach for Computer-Aided Diagnosis of Dengue Fever".IEEE Transactions

on information Technology in biomedicine. Vol. 16, no. 1, January- 2012.

[11] Yien Ling Hii1*, Huaiping Zhu2, Nawi Ng1, Lee Ching Ng3, Joacim Rocklo¨ v1Francis Mutuku, DVBNTD/CWRU/Emory University, Kenya.Forecast of Dengue Incidence Using Temperature andRainfall. Received May 6, 2012; Accepted October 2 2012; Published November 29, 2012. PLOS Neglected Tropical Diseases www.plosntds.org. November 2012 | Volume 6 | Issue 11 e1908

[12] B. Gultekin Cetiner a, Murat Sari b and Hani M. Aburas c. Recognition of dengue disease patterns using artificial neural networks. 5th International Advanced Technologies Symposium (IATS'09), May 13-15, 2009, Karabuk, Turkey.

[13] http://www.mathworks.in/help/pdf_doc/nnet/

[14] https://en.wikipedia.org

AUTHORS PROFILE

M.V. Jagannath Reddy received his B E in Electronics & Communication Engg. and M.Tech. in Computer Science Engg. He is pursuing Ph.D in CS from Rayalaseema University, Kurnool, India. He has got 18 years of teaching and industrial experience. He served as the Head, Dept of CSE & IT at MITS, Madanapalle, during the year 2009-2010. His areas of interests include Data Mining and Data warehousing, Intelligent Systems, DBMS. He is a member of ISTE, IAENG and IACSIT. He has published more than 15 papers in International journals and conferences. Some of his publications appear in IEEE Explorer and IJCSIT digital libraries. .

Dr. B. Kavitha received the MCA from Sri Padmavathi Mahila University, Tirupat. And Ph.D in Computer Science from Sri Padmavathi Mahila University, Tirupati. She has more than 14 years of teaching experience. Presently she is working has Lecturer in Computer Applications, Government degree college(Men), Srikalahasti. Her areas of interest are Fuzzy logic in data bases, Software Engineering, Data warehousing and mining. She published more than 15 papers in international journals and conferences. Some of her publications appear in IEEE Explorer and IJCSIT digital libraries.