Analysis of Significant Factors for Dengue Infection Prognosis Using the Random Forest Classifier

A.Shameem Fathima
Research Scholar,
Department of Computer Science and Engineering
Manonmanium Sundaranar University
Tamilnadu , India

D.Manimeglai
Head of the Department,
Department of Information Technology
National Engineering College
Tamilnadu , India

***Abstract*—Random forests have emerged as a versatile and highly accurate classification and regression methodology, requiring little tuning and providing interpretable outputs. Here, we briefly explore the possibility of applying this ensemble supervised machine learning technique to predict the vulnerability for complex disease - Dengue which is often baffled with chikungunya viral fever. This study presents a new-fangled approach to determine the significant prognosis factors in dengue patients. Random forests is used to visualize and determine the significant factors that can differentiate between the dengue patients and the healthy subjects and for constructing a dengue disease survivability prediction model during the boosting process to improve accuracy and stability and to reduce over fitting problems. The presented methodology may be incorporated in a variety of applications such as risk management, tailored health communication and decision support systems in healthcare**

Keywords—Data Mining; Dengue Virus; Machine learning; Random Forest

I. INTRODUCTION

Dengue is a rigorous fever spread by the nibble of an infected mosquito Aedes aegypti . Chikungunya is a crippling viral disease transmitted to humans by infected mosquitoes [1]. It is also an arbovirus that shares the same vector with dengue virus. The disease shares some clinical signs with dengue, and can be misdiagnosed in areas where dengue is common. Thus, in dengue-endemic region, chikungunya is also a significant cause of viral fever causing outbreaks associated with severe morbidity. As these reemerging tropical viral diseases have been increasing in the past several years, several research studies have contributed to investigate factors in diseases [2]. The vital aspects of clinical informatics and public health informatics may be essential to improve the ability to bring basic research findings and evaluate the efficiency of interventions across communities which continues to be beyond the reach of scientists and health professionals.

Presently, highly developed techniques in the fields of data mining, a new stream of methodologies, have come into reality; they provide processes for discovering useful patterns or models from large datasets [3]. One of the most common widely used techniques in data mining is classification. It is used to extract models describing important data classes and to predict the outcome in unseen data at the single point of time [4]. Therefore, in order to aid medical practitioners, predict the accurate outcomes, data mining is needed to process voluminous data available from previously solved cases and to imply the possible treatments based on analyzing the abnormal values of some significant attributes.

Generally intelligent techniques used in dengue fever analysis are fuzzy theory [5], decision trees [6], and Bayesian classifier [7]. Recently Random Forest technique has happened to be an attractive ensemble method in machine learning. As a result, several research studies have successfully applied the algorithm to solve classification problems in object detection, including face recognition, video sequences and signal processing systems [8]. The dataset is collected from various laboratories and hospitals in Tamil Nadu. The main contribution is to provide some experimental insights about the behavior of the variable importance index based on random forests. The performance of the random forests is investigated to generate better perfection models in Dengue survivability. The 10-crossfold validation method, confusion matrix, accuracy, sensitivity, specificity and ROC curve are used to evaluate the dengue virus survivability prediction models.

The remainder of this paper is organized as follows section II introduces the basic concepts of Random Forest. Section III presents the methodologies and experimental design used in this paper. Experiment results and discussions are presented in section IV. The conclusion and outline of future work are given in section V.

II. BASIC CONCEPTS OF RANDOM FOREST

Random Forests [RF] is essentially a data mining package based fundamentally on regression tree analysis [9]. RF tries to perform regression on the specified variables to provide the suitable model. RF uses bootstrapping to produce random trees and it has its own cross validation techniques to validate the model for prediction / classification. Being one of the ensemble learning techniques, Random Forest has been proven to be especially accepted and dominant techniques in the pattern recognition and machine learning for high-dimensional classification [10] and skewed problems [9]. These studies used RF to construct a collection of individual decision tree classifiers which utilized the classification and regression trees (CART) algorithms [11]. The RF model building procedure is essentially the same as a normal classification tree, but with randomness introduced. The procedure is as follows:

1) For the whole set of training data points (predictors and their corresponding response), RF.

2) Each tree when terminal nodes are reached is saved and RF repeats the process. The user specifies how many times this process is repeated (how many trees to grow).

Once the total number of trees is grown the model (or forest) can be saved for subsequent loading in R. RF also supplies the variable importance for each of the predictors in the training data.

Not only is there often a large number of records in the database, but there can also be a large number of fields (attributes, variables); so, the dimensionality of the problem is high. A high-dimensional data set creates problems in terms of increasing the size of the search space for model induction in a combinatorial explosive manner. In addition, it increases the chances that a data mining algorithm will find spurious patterns that are not valid in general. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables. As a classifier, random forest performs an implicit feature selection, using a small subset of "strong variables" for the classification only [12], leading to its superior performance on high dimensional data. The outcome of this implicit feature selection of the random forest can be visualized by the "Gini importance" [9], and can be used as a general indicator of feature relevance. This feature importance score provides a relative ranking of the features, and is – technically – a by-product in the training of the random forest classifier

III. Methodologies and Experiment Design

This paper, focusing on random forests, the increasingly used statistical method for classification and regression problems introduced by Leo Breiman in 2001, proposes to investigate two classical issues of variable selection. The first one is to find important variables for interpretation and the second one is more restrictive and tries to design a good cost-conscious prediction model. In this section, the viral data preparation used in this experiment is first described. Then the performance evaluation methods including accuracy, sensitivity, specificity and Receiver Operating Characteristic (ROC) curve is presented.

A. Dataset

The Dengue survivability data and viral particles in samples of patients clinically suspected for having dengue fever were obtained from several hospitals, King Institute of preventive Medicine and laboratory diagnostic centers in Tamil Nadu, India. The data includes patient information that was diagnosed with dengue during the year 2009-2011. Clinical presentation was recorded from the patients at different stages those during included in the study. The arboviral survivability data consist of nearly 5000 instances and 29 attributes. These variables are widely used in our hospitals for the diagnosis and monitoring of dengue patients .The whole dataset if divided into two classes, 'Dengue positive' in which the patients are suspected for having dengue fever and also on real time PCR result proves to be Dengue positive and Dengue negative – class in which the patients are suspected for having dengue fever but on real time PCR result proves to be negative. All this raw data does not necessarily equates to having useful information; on the contrary, it could lead to an information overflow rather than insight. What doctors need is high-quality support for making decisions. Data mining techniques can be used to extract useful knowledge from clinical data, to provide evidence for and thus support medical decision making Symptoms for chikungunya and dengue are almost identical - high fever, headache, eye ache, joint pain, rashes and lethargy. These viral diseases are characterized by an abrupt onset of fever frequently accompanied by joint pain. Other common signs and symptoms include muscle pain, headache, nausea, fatigue and rash. The joint pain is often very debilitating, but usually lasts for a few days or may be prolonged to weeks. Symptoms appear between 4 and 7 days after the patient has been bitten by the infected mosquito and these include:

- High fever (40°C/ 104°F)
- Joint pain (lower back, ankle, knees, wrists or phalanges)
- Joint swelling
- Rash
- Headache
- Muscle pain
- Nausea
- Fatigue

B. Evaluation methods

For the success of any data mining project, the data and especially the number of attributes play an important role. The more attributes are used, the higher the probability becomes that strong predictors are identified, and non-linearity and multivariate relationship can occur that intelligent techniques can exploit. If number of attributes increases, the density of the data set in pattern space drops exponentially and complexity of models can grow linearly or worse [13]. Complex models (i.e. a large number of parameters) have a higher chance of over fitting to the training data and will not perform well on new data (low generalization), so attribute selection is important. In this experiment, evaluation methods including basic performance measures and ROC curve are applied.

These evaluation methods are based on the confusion matrix. The confusion matrix is a visualization tool commonly used to present performances of classifiers in classification tasks [3]. It is used to show the associations between real class attributes and that of predicted classes. The intensity of effectiveness of the classification model is calculated with the number of correct and incorrect classifications in each possible value of the variables being classified in the confusion matrix [14] (see Fig. 1).

		Predicted Class	
		Dengue Positive	Dengue Negative
Outcome	Dengue Positive	TP	FN
	Dengue Negative	FP	TN

Fig. 1. The Confusion Matrix

The confusion matrix is used to compute true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), as represented in Fig. 1.

C. Performance Measures

There are three commonly used performance measurements including accuracy, sensitivity and specificity [3]. The accuracy of classifiers is the percentage of correctness of outcome among the test sets exploited in this study as defined in (1). The sensitivity is referred as the true positive rate, and the specificity as the true negative rate. Both sensitivity and specificity used for measuring the factors that affect the performance are presented in (2) and (3), respectively.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad \ldots\ldots\ldots\ldots \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{(TP+FP)} \quad \ldots\ldots\ldots\ldots \quad (2)$$

$$\text{Specificity} = \frac{TN}{(TN+FN)} \quad \ldots\ldots\ldots\ldots \quad (3)$$

The risk rate of the corresponding integrated risk factor associated with each prediction method is reported. It is computed as the ratio of the probability of developing disease among those predicted susceptible to the probability of developing disease among those predicted non-susceptible.

D. Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic curve graphically interprets the performance of the algorithm implemented. It is used as an evaluation criterion for the predictive performance of the classification or the data mining algorithms [15]. ROC curve is a two-dimension graph in which the true positive rate (TPR) (4) is plotted on the Y axis and the false positive rate (FPR) (5) is plotted on the X axis. *TPR* is the true positive value which is the number of correct predictions. *FPR* is the false positive value which is the number of incorrect predictions.

$$TPR = \frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots \quad (4)$$

$$FPR = \frac{FP}{TN+FP} \quad \ldots\ldots\ldots\ldots \quad (5)$$

ROC analysis offers more robust evaluation of the relative prediction performance of the models than the tradition comparison of relative error, such as error rate [16].

IV. Results and Discussion

All analyses were carried out using R - a free, cross-platform, open-source statistical analysis language and program. It is also an alternative to expensive commercial statistics software such as SPSS. Packages extend the functionality of R by enabling additional visual capabilities, statistical methods, and discipline-specific functions [17]. The recommended R distribution includes a number of packages in its library. These are collections of functions and data [18]. The base package, the stats package, the datasets package and several other packages, are automatically attached at the beginning of a session**.** Both of the random Forest package, ROCR package, party package and rpart package [17] [18] is frequently used.

For biological research applications, interpretability of results is a key factor in selecting a particular machine learning method. For the experiment results, we are interested in the percentage of correctly classified instances of the algorithm (accuracy percentage) and the number of rules or size of trees produced by the classifiers. For the experimental setup, all the original datasets are entered in to excel sheet and saved as csv file format and imported as input to the R software for analysis. Next, the identified classification technique is implemented and tested on the viral dataset.One part of the data is used to create the classifier, the other part is held out to test the performance of the model on cases that have not been used for training. A more sophisticated internal validation method is cross validation. This procedure will result in a more accurate estimate of the model performance.

For RF analysis, RF classification tree methods (number of trees =500; number of variables tried at each split =5) is used. To measure the importance of predictor variables, the mean decrease in accuracy and Gini index at each node were used. Fig. 2 illustrates the 29 most important variables of each measure. Mean Decrease in Accuracy exploits the margin, defined as the average of (% of votes for true class in the untouched OOB data) - (% of votes for the correct class in the variable-permuted OOB data) over all trees. In other words, the larger the size of the margin, the more important the predictor is. Gini importance is calculated for each variable using the Gini impurity criterion of the resulting subsets of the data at each decision node where the variable was used. Gini impurity is based on the squared probabilities of cases and controls in the two resultant subsets after a split is made using a variable. By definition, the impurity in the resulting subsets must be less than in the parent subset. The Gini index for a given variable is the sum over all trees of the decrease in Gini impurity after each split that involved that variable.

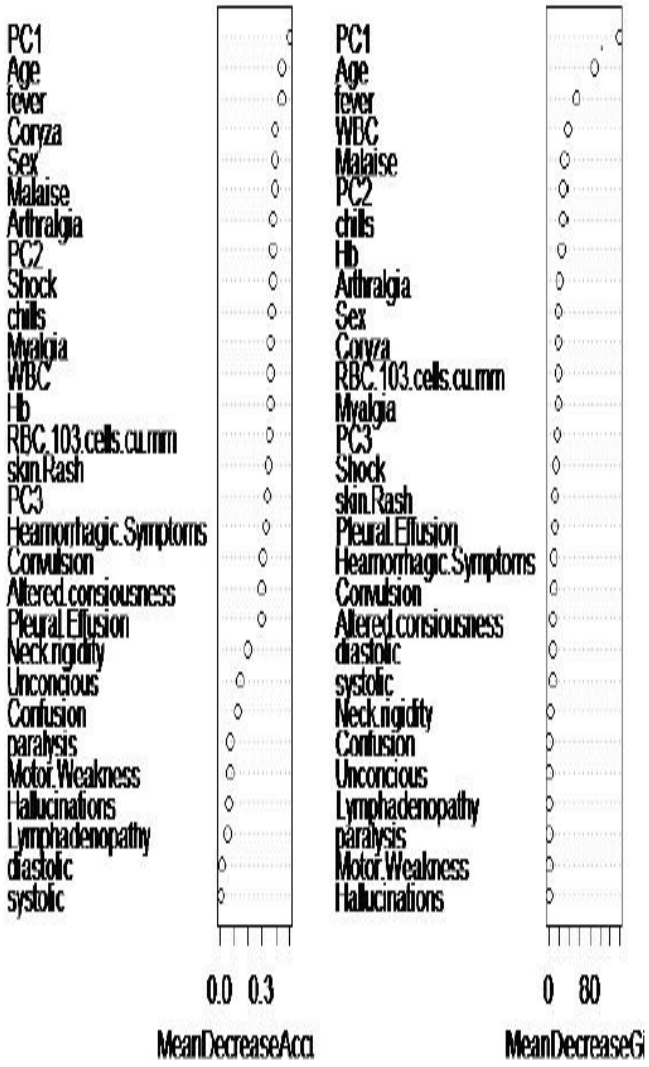

Fig. 2. Variable Importance Plots with Top 29 Variables caption

Left panel contains the 29 most important variables for predicting case control status descending by Mean Decrease Accuracy (average of (% of votes for true class in the untouched OOB data) - (% of votes for the correct class in the variable-permuted OOB data) over all trees). Right panel contains the 29 most important variables descending by Mean Decrease Gini Index (adding up the Gini decrease for each individual variable over all trees). Both the accuracy measure and the Gini index detected the variables which had significant p-values less than 0.0001 for the Fisher's exact test within the 29 most important variables.

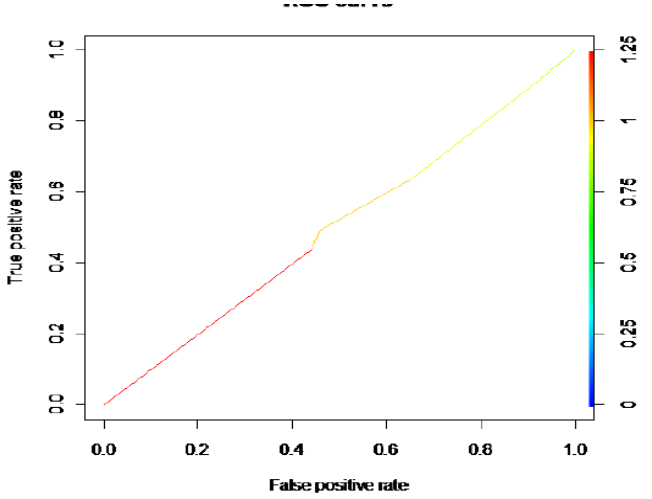

Fig. 3. ROC Curve

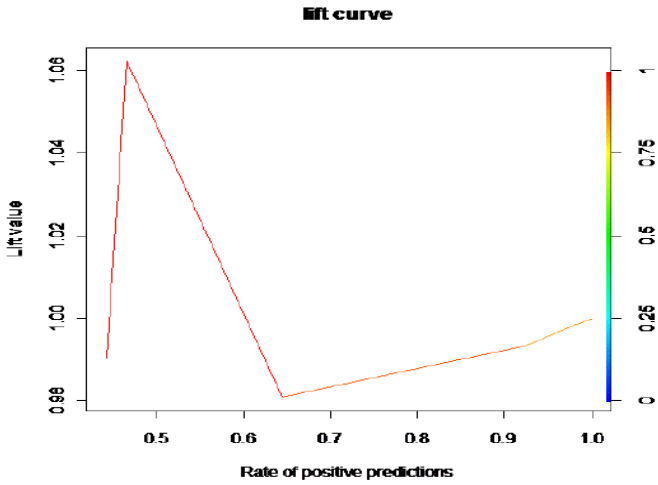

Fig. 4. LIFT CURVE

A predictive model is created using cforest (Breiman's random forests) from the package *party*, to evaluate the predictive model on a separate set of data, and then the performance using ROC curves and a lift chart is plotted. These charts are useful for evaluating model performance in data mining and machine learning. The performance of the model applied to the evaluation set is plotted as an ROC curve and lift chart as seen in Fig 3 and 4 respectively.

Permutation importance, on the other hand, is a reliable measure of variable importance for uncorrelated predictors when sub-sampling without replacement — instead of bootstrap sampling — and unbiased trees are used in the construction of the forest [19]. To meet this aim, conditional permutation is performed in which the importance measure is able to reveal the fake correlation between the response variable and other predictor variables. The results of conditional permutation scheme are shown in Fig.5.

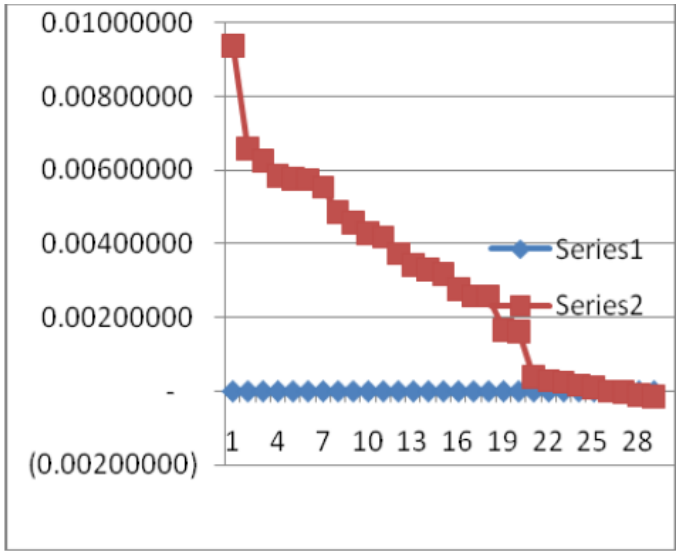

Fig. 5. Plot showing the conditional Importance of each variable

By inferring from these results the most important attributes are identified in the order of Platelet count 1, Malaise, Coryza, Myalgia, Platelet count, Chills, Arthralagia, White blood cells count , Fever .These results are compared with the instantaneous study of the viral diseases by the doctors and virologists reported by the World Health Organization .The report proves that the Patients with dengue had significantly lower platelet, white blood cell (WBC) and Signs of rash and indicators of liver damage, in combination with other variables such as age, myalgia , WBC count, and platelet counts [20]. The findings of this study suggest that several clinical and laboratory measures could potentially distinguish patients with dengue from those with other viral disease. Low platelet count and decreases in WBC and neutrophils were independently associated with the presence of dengue [21] [22] [23].The performance measures obtained by the implemented technique is tabulated and shown below in table 1

TABLE I. PERFORMANCE OF THE SINGLE CLASSIFIER ON THE DATA

RF- Performance measures	
Sensitivity	.9404
Specificity	0.9219
Accuracy	0.9234
Risk Rate	0.519
TPR	0.51
FPR	0.99

As shown from the results, using RF as a base learning algorithm ability of prediction is reduced and the present study highlighted important clinical observations of dengue viruses, to rule out the present confusion and may help to establish a diagnostic algorithm to distinguish dengue from other viral patients. The study also guides in early detection of the viral diseases so that appropriate management may be undertaken to reduce the long-lasting consequences in health. Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data.

V. CONCLUSION AND FUTURE WORK

Identification of the influential clinical symptoms and laboratory features that help in the diagnosis of dengue fever (DF) in early phase of the illness would aid in designing effective public health management and virological surveillance strategies. Keeping this as our main objective, we develop in this paper a new computational intelligence-based methodology that predicts the diagnosis in real time, minimizing the number of false positives and false negatives.Given its performance, random forest and variable selection using random forest should probably become part of the "standard tool-box" of methods for the analysis of dengue data. The proposed method can be used for variable selection fulfilling the objectives above. Screen plots can be used to recover the important variables that are related to the diagnosis of Dengue, with-out being adversely affected by collinear ties; the proposed method is capable of extracting patterns, but with-out the cooperation and feedback from the medical practitioner, these results would be useless. Besides, this method is not aimed at replacing the medical practitioner and researchers, but rather to complement their invaluable efforts to save more human lives. As for further work, the plan is to investigate the diversity of the number of classifiers such as linear discriminant analysis, logistic regression and support vector machines in this aspect. Another possibility to investigate is using the RF algorithm in larger data sets with scores of attributes. Finally, a comparison of the classifiers ensemble would be of interest.

ACKNOWLEDGMENT

Thanks to the Virology department staff at King Institute of Preventive Medicine, Chennai, Tamilnadu, doctors and microbiologists who provided us with a cosmic amount of viral data needed for our research study and validated our results.

REFERENCES

[1] Chakkaravarthy, V.M., S. Vincent and T. Ambrose,1011. Novel Ap proach of Geographic Information Systems on Recent outbreaks of Chikungunya in Tamil Nadu, India. J. Env.Sci. Tech.4(4):387-394 *(references)*

[2] T. Srinivasan, A. Chandrasekhar, J. Seshadri and J. B. S. Jonathan,―Knowledge discovery in clinical databases with neural network evidence combination,‖ in Proc. International Conference on Intelligent Sensing and Information, 2005, pp. 512-517.

[3] J. Han and M. Kamber, Data mining: concepts and techniques. 2nd.ed. San Francisco: Morgan Kaufmann, Elsevier Science, 2006.

[4] M. T. Skevofilakas, K. S. Nikita, P. H. Templaleksis, K. N. Bir bas, I.G. Kaklamanos and G. N. Bonatsos, ―A decision support system for breast cancer treatment based on data mining tech-nologies and clinical practice guidelines,‖ in IEEE-EMBS the Twenty-SeventhAnnual International Conference on Medicine and Biology Society,2005, pp. 2429-2432.

[5] Parido, A., and P. Bonelli. A new approach to fuzzy classifier systems.In Proceedings of the Fifth International Conference on GeneticAlgorithms. pp. 223–230. 1993.

[6] Hassanien, A.E. Classification and feature selection of breast cancer data based on decision tree algorithm. International Journal of Studies in Informatics and Control Journal, 12(1), 33– 39.2003.

[7] Cheeseman, P., and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U.M. Fayyad, G. Piatetsky–Shapiro P. Smyth and R.Uthurusamy (Eds.), Advances in Knowledge Discovery and DataMining. AAAI Press/MIT Press.1996

[8] M. Zhou and H. Wei, ―Face Verification Using GaborWavelets And AdaBoost,‖ in the Eighteenth International Conference on PatternRecognition, Hong Kong, 2006, pp. 404-407.

[9] L. Breiman, ―Random Forests,‖ J. Machine Learning vol. 45, pp. 5– 32, 2001.

[10] N. Meinshausen, ― Quantile Regression Forests,‖ J. Machine Learning Research, vol. 7, pp. 983–999, 2006.

[11] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and regression trees. Wadsworth: Belmont, 1984

[12] Breiman L. Technical Report 670. Technical report, Department of Statistics, University of California, Berkeley, USA; 2004. Consistency for a simple model of random forests.

[13] Bishop, C.M. (1995), Neural Networks for Pattern Recognition, Oxford University Press, Oxford, UK.

[14] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees and A. Zana si, Discovering data mining from concept to implementation. Upper Saddle River, N.J.: Prentice Hall, 1998.

[15] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled and D. Roth,―Generalization bounds for the area under the ROC curve,‖ J. Machine Learning Research, vol. 6, pp. 393-425, 2005.

[16] R. O. Duda, D. G. Stork and P. E. Hart, Pattern classification. 2nd ed. New York: Wiley, 2001.

[17] Maindonald, J. H., 2001. Using r for data analysis and graphics, http://wwwmaths.anu.edu.au/~johnm/r/usingR.pdf.

[18] R Core Development Team. *An Introduction to* R.http://cran.r-project.org

[19] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn.Bias in random forest variable importance measures:Illustrations,sources and a solution. BMC Bioinformatics, 8:25, 2007b.

[20] Ageep AK, Malik AA, Elkarsani MS. Clinical presentations and laboratory findings in suspected cases of dengue virus. *Saudi Med J.* 2006;27:1711–1713

[21] Dengue hemorrhagic fever: diagnosis, treatment, prevention and control. 2nd edition. Geneva: World Health Organization 1997.

[22] A. Hapfelmeier and K. Ulm –A new Variable selection approach Using RamdomForests Computational Statistics & Data Analysis, 2013, vol. 60, issue C, pages 50-69

[23] Ranjit S, Kissoon N, Gandhi D, Dayal A, Rajeshwari N, Kamath SR. Early differentiation between dengue and septic shock by comparison of admission hemodynamic, clinical, and laboratory variables: a pilot study. *Pediatr Emerg Care*. 2007;23:368– 375. [PubMed]

[24] (2013). UCI Machine Learning Repository. Available:http://archive.ics.uci.edu/ml/

[25] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," International Journal on Computer Science and Engineering (IJCSE), vol. 2, pp. 250-255, 2010