International Journal of Scientific & Engineering Research, Volume 6, Issue 1, January-2015
ISSN 2229-5518

Machine learning to Classify the Re-Emerging Arboviral Diseases

A.ShameemFathima, Dr.D.Manimegalai

Abstract— Recent advances in computing and developments in technology have facilitated the routine collection and storage of medical data that can be used to support medical decisions.In most cases however, there is a need for the collected data to be analysed in order for a medical decision to be drawn, whether this involves diagnosis, prediction, course of treatment,or signal and image analysis. Intelligent machine learning methods such as neural computing and support vector machines can be shown to be suitable approaches to such complex tasks. This paper presents a study of the use of intelligent methods for medical decision making that aims to investigate and demonstrate their potential in such an application.The medical application presented in this work is the classification of the Re-Emering Arboviral Disease such as Dengue and Chickungunya.The development of such a system is of great importance since – according tothe World Health Organization they affect much humans than any other infectiousdisease and its now on the increase. The proposed combinatorial methods have been applied to publicly available dengue virus that is re-emerging throughout the tropical world, causing frequent recurrent epidemics. As the number of such cases increases, the availability of information regarding these diseases is important in order to help experts in taking proper measures. The initial clinical manifestation of dengue is countered with other febrile states confounding both clinical management and disease surveillance. In this paper, we discuss the application of machine learning techniques that differentiate dengue from other febrile illnesses in the primary care setting and predict severe arboviral disease among population. By applying sound analysis to a very detailed dataset from one of the largest outbreaks that affected South India in recent times, we not only describe the spread of dengue virus with high detail but also identify the best possible classifier for our Arboviral infected dataset.

Index Terms— Machine learning, Support Vector Machines, Naïve Bayes classifier, Randomforest, Neural networks, Dengue

1 Introduction

Datamining roots are traced back along three family lines which are Classical Statistics, Artificial Intelligence and Machine learning [1]. The main focus of this paper is on Machine learning, a process capable of independently acquiring data and Integrating that data to generate useful knowledge. Classification is used to assign items to a discrete group or class based on a specific set of features. Classification algorithms are a core component of statistical learning / machine learning. The classification problem may be encountered in different domains, such as "disease diagnosis". Disease diagnosis usually depends on many symptoms and results of medical exams that demonstrate the presence or absence of the disease. Thus, disease diagnosis can be described as a classification problem which is the aim of this study.

The concept of machine learning [2] is implemented by way of computing software system that act as human being who learns from experience, analyses the observations made and self-improves providing increased efficiency and effectiveness. The processes by which most modern predictive models are developed are adaptive, using numerical methods that adjust model parameters based upon analysis of data and model performance. The machine learning process generates a predictive model through mechanical analysis of data and introspection; model parameters are determined without human involvement. Machine learning techniques are often used for financial analysis and decision-making tasks such as accurate forecasting, classification of risk, medical diagnostic tasks, and data mining.[3] For example, a task in the health care field is to determine, given a set of observed symptoms, whether or not a person has a disease. These detection tasks are binary classification problems.However, implementing and comparing different machine learning techniques to choose the best approach can be challenging Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions.[4] The training dataset includes input data and response values. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset. A test dataset is often used to validate the model. Using larger training datasets often yield models with higher predictive power that can generalize well for new datasets. Supervised learning includes two categories of algorithms: Classification: for categorical response values, where the data can be separated into specific “classes” and Regression: for continuous-response values. Common classification algorithms include: Support vector machines (SVM), Naïve Bayes classifier, Neural networks. This aim of this paper is to implement and evaluate against different machine learning techniques such as SVM, Naïve Bayes classifier, and Random forest , that are an ensemble learning method for classification and Neural Networks.

The overall purpose of this study is to investigate clinical parameters in dengue infection that would be useful in distinguishing dengue fever from other febrile illnesses and in predicting the extent of disease severity in an early stage of infection. Our goal is to conduct research in the identification of one of the Arbovirus - Dengue viruses. We have been interested in learning whether or not specific data for the Arboviral-dengue viruses are fittingly classifiable by using the theory of the machine learning namely using the SVM, Naïve Bayes classifier, Neural networks and Randomforest. The problem is to find algorithms suitable to apply in order to discover relationships between data attributes and make predictions that could be useful for decision support. The diagnosis and treatment of dengue is guided by the symptoms and findings that the patient presents, and cannot depend on laboratory confir-

International Journal of Scientific & Engineering Research Volume 6, Issue ƚϭ,)ÜÕÌıl Ỹưk
ISSN 2229-5518

mation, since routine tests cannot confirm dengue with the speed required for patients in critical condition. The real potential of data mining is only fully realised when a hybrid of the techniques is implemented. Not all algorithms do equally well in all areas and it thus makes sense to implement a variety of algorithms .This is one of the central aim of this study to sythesise a number of different machine learning techniques such as SVM, Naïve Bayes classifier, Randomforestand Neural networks. The above discussion again outlines the need to achieve hybrid learning of classification techniques used for the disease diagnosis and it is the development of this methodology to which we now progress using MATLAB R2013a.

The present aim is not to evaluate the complexities of each algorithm in light of the data but instead to test their efficacy in classifying the data in the context of clinical Ar-boviral diagnosis.Algorithms are treated in a black-box fashion.This paper is organized as follows: Section 1 provides an introduction to the topic of the research describing the problem that is discussed. Section 2 describes the source of the diagnostic viral data under studyand its depiction. Section 3 explains the machine learning techniques implemented and Section 4 gives the experimental setup along with results and the finally the overall conclusions have been conveyed along with perspectives of future work in section 5.

2 SOURCE OF THE INVESTIGATIVE VIRAL DATA UNDER STUDY

2.1 Dataset

The medical dataset we are classifying includes 4950 real records of patients suffering from viral infection from several hospitals, King Institute of preventive Medicine and laboratory diagnostic centers in Tamil Nadu, India. The data were collected and reported in investigation form of viral infection of patients from the year 2009 to the year 2011. Clinical presentation was recorded from the patients at different stages that are included in the study. The entire dataset is put in one file having many records. Each record corresponds to most relevant information of one patient. Totally there are 26 attributes (symptoms) and one class attribute. Datasets that are available for dengue describe information about the patients suffering with dengue disease and without dengue disease along with their symptoms. Blood samples were collected from patients, who were clinically diagnosed to have acute viral infection, at the time of hospital admission and the time of discharge .The total of 4950 viral infected patient records with Twenty six attributes is used in the experiment. We took into account cases marked as "probable" and "confirmed" and did not include cases labeled "discarded."In our study of interest, Twenty six independent variables factors/symptoms causes for infection (Age, Gender, Fever, Chills, Coryza,Systolic pressure, Diastolic pressure, Shock, Myalgia, Malaise, Arthralgia , Hallucinations, Confusion, Altered consciousness, Unconscious, Convulsion, Neck rigidity, Motor weakness, Paralysis, Lymphadenopathy, Skin rash, Hemorrhagic symptoms, Pleural effusion, Hemoglobin, Red Blood Count and Platelet count at admission) are the scaled numeric variables and report is the categorical variable. The data undergoes an extensive data preprocessing task such as data selection, cleaning, reduction, and discretization.

In this research work ,the data is loaded from the CSV-file, by the use of the interactive tools into dataset arrays as the Dataset arrays make it easier to work with data of different datatypes to be stored as part of the same matrix. The data is segregated into response and predictors. Cross validation is almost an inherent part of machine learning. Cross validation may be used to compare the performance of different predictive modeling techniques. In this work, we use holdout validation. Other techniques including k-fold and leave-one-out cross validation are also available. The data is partitioned into training set and test set. The training set will be used to calibrate/train the model parameters. The trained model is then used to make a prediction on the test set. Predicted values will be compared with actual data to compute the confusion matrix. Confusion matrix is one way to visualize the performance of a machine learning technique. Our focus for this research work is to build machine learning models using clinical viraldata for prediction of dengue diagnostic tasks and to understand what factors influence this outcome.

3 IMPLEMENTED MACHINE LEARNING TECHNIQUES

Machine learning algorithms use training data to construct models of the relationships between a set of input attributes and a target attribute.[5] These algorithms often are used for either regression or classification. The implementation of SVM, Naïve Bayes classifier, neural networks and Randomforest are explained in the coming section. The Steps involved in our in Supervised Learning [6] are defined below:

- ➢ Prepare Data -All supervised learning methods start with an input data matrix, usually called X here. Each row of X represents one observation. Each column of X represents one variable, or predictor
- ➢ Choose an Algorithm -There are tradeoffs between several characteristics of algorithms, such as:
- ✓ Speed of training
- ✓ Memory usage
- ✓ Predictive accuracy on new data
- ✓ Transparency or interpretability, meaning how easily you can understand the reasons an algorithm makes its predictions
- ➢ Fit a Model - The fitting function you use depends on the algorithm you choose.
- ➢ Choose a Validation Method - The three main methods to examine the accuracy of the resulting fitted model are:

- *Ms.Shameem Fathima is currently pursuing Ph.d in Computerscience and Engineering at Manonmanium Sundaranar University, India.Her area of Interest is Datamining and Machine learning.*
- *Dr.D.Manimegalai is currently the Head of Department of Information Technology,National Engineering College,India. She had her BE & ME from Government College of Technology, Coimbatore and PhD from Manonmaniam Sundaranar University, Tirunelveli .She had been a brilliant supervisor to many academaciand throughout the country.She has modest number of publications in leading journals,international and national conferences .*

International Journal of Scientific & Engineering Research Volume 6, Issue 4, April-2015
ISSN 2229-5518

- ✓ Examine the resubstitution error.
- ✓ Examine the cross-validation error.
- ✓ Examine the out-of-bag error for bagged decision trees.

➢ Examine Fit and Update Until Satisfied- After validating the model, if needed ,change it for better accuracy, better speed, or to use less memory.

- ✓ Change fitting parameters to try to get a more accurate model.
- ✓ Change fitting parameters to try to get a smaller model.
- ✓ Try a different algorithm.

➢ Use Fitted Model for Predictions -To predict classification or regression response for most fitted models, the predict method is used.

3.1 Support Vector Machines

The SVM methodology comes from the application of statistical learning theory to separating hyper planes for binary classification problems [7]. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points.The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab.SVMs are function-based classifiers, which can be expressed in the standard form of quadratic optimization programming, which can be solved easily. Two key elements in the implementation of SVM are the techniques of mathematical programming and kernel functions. Kernels can also be constructed to incorporate domain knowledge. This so-called „kernel trick" gives the SVM great flexibility. [8]

As with any supervised learning model, we first train a support vector machine, and then cross validate the classifier. Using the trained machine we classify (predict) new data. In addition, to obtain satisfactory predictive accuracy, we use various SVM kernel functions, and the parameters of the kernel functions are tuned. an SVM classifier using fitcsvm is used to train, and optionally cross validate. The inputs are:

X — Matrix of predictor data, where each row is one observation, and each column is one predictor.

Y — Array of class labels with each row corresponding to the value of the corresponding row in X. Y can be a character array, categorical, logical or numeric vector, or vector cell array of strings. Column vector with each row corresponding to the value of the corresponding row in X. Y can be a categorical or character array, logical or numeric vector, or cell array of strings.

KernelFunction — The default value is 'linear' for two-class learning, which separates the data by a hyperplane. The value 'rbf' is the default for one-class learning, and uses a Gaussian radial basis function. An important step to successfully train an SVM classifier is to choose an appropriate kernel function.

Standardize — Flag indicating whether the software should standardize the predictors before training the classifier.

ClassNames — Distinguishes between the negative and positive classes, or specifies which classes to include in the data.

The resulting, trained model (SVMModel) contains the optimized parameters from the SVM algorithm, enabling you to classify new data. Classification of the new data is done using the predict function. The resulting vector, label, represents the classification of each row in X. score is an n-by-2 matrix of soft scores. Each row corresponds to a row in X, which is a new observation. The first column contains the scores for the observations being classified in the negative class, and the second column contains the scores observations being classified in the positive class.

3.2 Naive Bayes Classification

The naive Bayes classifier is designed for use when predictors are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid.[9] It classifies data in two steps:
Training step: Using the training data, the method estimates the parameters of a probability distribution, assuming predictors are conditionally independent given the class.
Prediction step: For any unseen test data, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test data according the largest posterior probability.

The class-conditional independence assumption greatly simplifies the training step since you can estimate the one-dimensional class-conditional density for each predictor individually. While the class-conditional independence between predictors is not true in general, research shows that this optimistic assumption works well in practice. [10] This assumption of class-conditional independence of the predictors allows the naive Bayes classifier to estimate the parameters required for accurate classification while using less training data than many other classifiers. This makes it particularly effective for data sets containing many predictors. The training step in naive Bayes classification is based on estimating P(X|Y), the probability or probability density of predictors X given class Y. [11]The naive Bayes classification model ClassificationNaiveBayes and training function fitcnb provide support for normal (Gaussian), kernel, multinomial, and multivariate, multinomial predictor conditional distributions. To specify distributions for the predictors, use the DistributionNames name-value pair argument of fitcnb. You can specify one type of distribution for all predictors by supplying the string corresponding to the distribution name,[12] or specify different distributions for the predictors by supplying a length D cell array of strings, where D is the number of predictors (that is, the number of columns of X).

3.3 RandomForest

A random forest is basically an ensemble of decision trees. Each tree classifies (often linearly) the dataset using a subset of variables. [13] The number of trees in the forest and the number of variables in the subset are hyper-parameters and must be chosen a-priori. The number of trees is of the order of hundreds, while the subset of variables is quite small compared with the total number of variables.Random forests also provide a natural way of assessing the importance of input variables (predictors). This is achieved by removing one variable

International Journal of Scientific & Engineering Research Volume 6, Issue [illegible]
ISSN 2229-5518

at a time and assessing whether the out-of-bag error changes or not. If it does, the variable is important for the decision.

MATLAB package has two approaches for calculating variable importance: The first is "predictorImportance": The second is permutation method. predictorImportance computes estimates of predictor importance for tree by summing changes in the mean squared error (MSE) due to splits on every predictor and dividing the sum by the number of branch nodes. If the tree is grown without surrogate splits, this sum is taken over best splits found at each branch node. If the tree is grown with surrogate splits, this sum is taken over all splits at each branch node including surrogate splits. imp has one element for each input predictor in the data used to train this tree. At each node, MSE is estimated as node error weighted by the node probability. Variable importance associated with this split is computed as the difference between MSE for the parent node and the total MSE for the two children [14].

Feature importance measures the increase in prediction error if the values of that variable are permuted across the out-of-bag observations. This measure is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble.[15] The Class TreeBagger creates ensemble of bagged decision trees. TreeBagger generates in-bag samples by oversampling classes with large misclassification costs and undersampling classes with small misclassification costs. Consequently, out-of-bag samples have fewer observations from classes with large misclassification costs and more observations from classes with small misclassification costs. If you train a classification ensemble using a small data set and a highly skewed cost matrix, then the number of out-of-bag observations per class might be very low. Therefore, the estimated out-of-bag error might have a large variance and might be difficult to interpret. The same phenomenon can occur for classes with large prior probabilities.The Steps involved in our in Supervised Learning - classification to create an ensemble are defined below:

i. Put Predictor Data in a Matrix
ii. Prepare the Response Data
iii. Choose an Applicable Ensemble Method
iv. Set the Number of Ensemble Members
v. Prepare the Weak Learners
vi. Call fitensemble

3.4 Neural Networks

‘Neural network’ (NN), is a mathematical model or computational model based on biological neural networks. Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. A a neural network can be trained to perform a particular function by adjusting the values of the connections (weights) between elements.[16]

Typically, neural networks are adjusted, or trained, so that a particular input leads to a specific target output.In practical applications, when the exact model is not known, a neural network can be used for example of a ‘black-box’ technique. By no means, however, should the neural network be seen as the ultimate solution for problems with undefined or only partially defined models. The main reason is that it gives no additional information about the physical relationships and thus it will give no physical insight into the process. On the other hand, neural network have the ability to learn the relationship between the input and the output. The network usually consists of an input layer, some hidden layers, and an output layer.

The neural network system consist of three parts; training, testing and validation. 80% of the data were used in training, while the remaining 20% were used in testing. Feed forward back propagation algorithm with Levenberg-Marquardt (trainlm) training was used in this system because; it has the fastest convergence, very accurate training and lower mean squared error (MSE).Function tansig and purelin were used as the activation function for the input and output respectively. In supervised learning the output from the neural network is compared with a set of targets, the error signal is used to update the weights in the neural network.

Neural networks are good at fitting functions. To define a fitting problem for the toolbox, arrange a set of Q input vectors as columns in a matrix. Then, arrange another set of Q target vectors (the correct output vectors for each of the input vectors) into a second matrix. The neural network toolbox makes it easier to use neural networks in matlab. The toolbox consists of a set of functions and structures that handle neural networks. The toolbox is based on the network object. This object contains information about everything that concern the neural network, e.g. the number and structure of its layers, the conectivity between the layers, etc. In oue experiments, newff is used to create a feed-forward backpropagation network to allow an easy construction. The Steps involved in this task are defined below:[17]

- ✓ Define one sample: inputs and outputs
- ✓ Define and custom network
- ✓ Define topology and transfer function
- ✓ Configure network
- ✓ Train net and calculate neuron output

Each time a neural network is trained, can result in a different solution due to different initial weight and bias values and different divisions of data into training, validation, and test sets. As a result, different neural networks trained on the same problem can give different outputs for the same input. To ensure that a neural network of good accuracy has been found, retrain several times.[18]

It is very difficult to know which training algorithm will be the fastest for a given problem. It depends on many factors, including the complexity of the problem, the number of data points in the training set, the number of weights and biases in the network, .The training algorithms that are used for this study are Levenberg-Marquardt (LM) , Bayesian Regularization (BR), Scaled Conjugate Gradient (SCG) and the Resilient Backpropagation (RP) .These algorithms are available in the Neural Network Toolbox software and they use gradient- or Jacobian-based methods.[19]

The programming was written into Matlab’s M-File to obtain the values for true positive, false negative, true negative and false positive when the thresholdvalue is varies from 0 to 1. Next, the values for true positive rate, false positive rate,

International Journal of Scientific & Engineering Research Volume 6, Issue
ISSN 2229-5518

accuracy, precision, sensitivity, specificity, and 1-specificity were calculated.

4 Experimental Results and Discussion

This section is dedicated to the comparison between the four methods of diagnosis namely Support Vector machine, Random Forest, Naïve Bayes classifiers and neural networks.. For the experimental setup, all the original medical datasets are entered in to excel sheet and saved as csv file format. Next, all the identified algorithms are tested to the medical dataset with the option of using 10-fold cross-validation. We report sensitivity, specificity, and accuracy of the prediction 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 Complex models (i.e. a large number of parameters) have a higher chance of over fit-ting 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 are applied. These evaluation methods are based on the confusion matrix, a visualization tool commonly used to present performances of classifiers in classification tasks. 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 (see Fig. 1) and the description of the confusion matrix involved in our anlaysis is shown in Fig.2

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

Fig: 1 The Confusion Matrix

Measure	Meaning of the measure
TP	No symptoms of disease
FN	No symptoms of disease but disease exists
FP	Symptoms of disease exists but disease does not exist
TN	Symptoms of disease exists and the disease exist

Fig: 2 Description of the Confusion Matrix

There are three commonly used performance measurements including accuracy, sensitivity and specificity. [20]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.

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

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

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

The first experiment using the an SVM classifier is done using fitcsvm that is used to train, and optionally cross validate The data set is read and divide at random the dataset in two subsets, one with about 80% of the instances to training, and 20% of instances to testing.The parameters are chosen and the model is trained and then tested.

The Second experiment is done using the Naïve Bayes . The naive Bayes classification model ClassificationNaiveBayes and training function fitcnb provide support for normal (Gaussian), kernel, multinomial, and multivariate, multinomial predictor conditional distributions.

The third experiment is done using the random Forest and we have informative views of the data and evaluate the results of classification intuitively. Here we focus on the use of random forests for classification tasks, for predicting the disease status from a set of selected risk factors for predicting whether a patient has dengue or not.

The fourth experiment is done using the neural networks for the training algorithms Levenberg-Marquardt , Bayesian Regularization , Scaled Conjugate Gradient and the Resilient Backpropagation .Table 1 shows the Evaluation Measure of the machine learning techniques implemented on the original datasets.

Methods	Evaluation Measure			
	TP	TN	FP	FN
RF	843	94	54	0
SVM	843	113	35	0
NB	843	0	148	0
LM	843	111	37	0
BR	843	126	22	0
Scg	843	106	42	0
Rp	843	115	33	0

Table: 1 Evaluation Measure of the machine learning techniques implemented on the original datasets

International Journal of Scientific & Engineering Research Volume 6, Issue
ISSN 2229-5518

The fifth experiment is used to find features' importance in Random Forest implemented in Matlab using TreeBagger.The important predictors identified from this technique are : are Platelet count at admission, Coryza , RBC , Myalgia , Gender, Hemorrhagic symptoms, Pleural effusion, Malaise, Confusion, Arthralgia, Fever, Neck rigidity, Shock, Chills, Diastolic pressure, Systolic pressure, Skin rash, Convulsion.

A new dataset with only the significant features is obtained on which classification analysis is again done using svm, naives bayes classifier, RandomForest and Neural Networks.Again all the classification techniques are analysed using the new dataset wich contains only the important predictor variables.All the results are evaluated and compared. Table 2 shows the Evaluation Measure of the machine learning techniques implemented on the the important predictor variables.

Methods	Evaluation Measure			
	TP	TN	FP	FN
RF	843	92	56	0
SVM	843	111	37	0
NB	843	0	148	0
LM	843	112	36	0
BR	843	110	38	0
Scg	843	106	42	0
Rp	843	104	44	0

Table: 2 Evaluation Measure of the machine learning techniques implemented on the datasets of important predictor variables

The three commonly used performance measurements including accuracy, sensitivity and specificity are tabulated below.Table 3 shows the Performance measurements of the original dataset. Table 4 shows the performance measurements of the machine learning techniques implemented on the dataset which involves only the predictor variables

Methods	Performance measurements		
	Sensitivity	Specificity	Accuracy
RF	1	0.635135	0.94551
SVM	1	0.763514	0.964682
NB	1	0	0.850656
Lm	1	0.75	0.962664
Br	1	0.851351	0.9778
Scg	1	0.716216	0.957619
Rp	1	0.777027	0.9667

Table: 3 Performance Measure of the machine learning techniques implemented on the original datasets

Methods	Performance measurements		
	Sensitivity	Specificity	Accuracy
RF	1	0.621622	0.943491
SVM	1	0.75	0.962664
NB	1	0	0.850656
Lm	1	0.756757	0.963673
Br	1	0.743243	0.961655
Scg	1	0.716216	0.957619
Rp	1	0.702703	0.9556

Table: 4 Performance Measure of the machine learning techniques implemented on the datasets of important predictor variables

Our experiments show that in medical domains various classifiers perform roughly the same. The best accuracy is achieved by Bayesian Regularization (BR), **0.9778** when implemented on the original datasets and by Levenberg-Marquardt(LM) **0.963673**, when implemented on the datasets of important predictor variables. The good performance of SVM and random forest indicate that they can possibly be adjusted to improve specific adhoc methods for prediction of susceptibility to complex diseases such as Dengue. Naïve's Bayes shows the worst performance when applied on both the datasets. We show that random forest variable importance measures are a sensible means for variable selection in many classification tasks in bioinformatics and related scientific fields. The usage of both random forest algorithms and their variable importance measures in the system for statistical computing is illustrated and documented thoroughly in this application. Inspired by this approach, future research on variable importance measures for variable selection with random forests aims at providing further means of statistical inference, that can be used to guide the decision on which and how many predictor variables to select in a certain problem.The physicians found that the combination of classifiers was the appropriate way of improving the reliability and comprehensibility of diagnostic systems. The combination should be done in an appropriate way and the reliability of each classifier on the given new case should be taken into account, clearly demonstrate.

5 CONCLUSION AND FUTURE WORK

The historical development of development of machine learning and its applications in medical diagnosis shows that from simple and straightforward to use algorithms, systems and methodology have emerged that enable advanced and sophisticated data analysis. In the future, intelligent data analysis will play even a more important role, due to the huge amount of information produced and stored by modern technology. Current machine learning algorithms provide tools that can significantly help medical practitioners to reveal interesting relationships in their data. Regarding the future role of machine learning in medical diagnosis, our views are as follows:
Machine learning based diagnostic programs will be used as any other instrument available to physicians: as just another source of possibly useful information that helps to improve diagnostic accuracy. The final responsibility and judgement

International Journal of Scientific & Engineering Research Volume 6, Issue
ISSN 2229-5518

whether to accept or reject this information will, as usual, remain with the physician.The authors plan to worwork with hybrid learning of the neural networks and Genetic algorithms as their future work.
as usual, remain with the physician.The authors plan to work with hybrid learning of the neural networks and Genetic algorithms as their future work.

ACKNOWLEDGMENT

A heartfelt gratitude is expressed by the authors of this paper to the department of virology, king institute of preventive medicine and research, chennai, India, for releasing the viral data for research and education

REFERENCES

[1] I.Bratko, and M.Kubat (eds.): Machine Learning, Data Mining and Knowledge Discovery: Methods and Appli-cations, John Wiley & Sons, 1998.

[2] Kukar M. and Gro·selj C., Machine learning in stepwise diagnostic process, Proc. Joint European Conf. onArtifi¯cial Intelligence in Medicine and Medical Decision Making, pp.315-325, Aalborg, Denmark, 1999.

[3] Kononenko I., Bratko I., Kukar M., Application of machine learning to medical diagnosis. In R.S.Michalski,I.Bratko, and M.Kubat (eds.): Machine Learning, Data Mining and Knowledge Discovery: Methods and Appli-cations, John Wiley & Sons, 1998.

[4] Cestnik B., Estimating Probabilities: A Crucial Task in Machine Learning, Proc. European Conf. on Artificial Intelligence, Stockholm, August, 1990, pp. 147-149.

[5] Michalski R.S., Bratko I., and Kubat M. (eds.): Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications, John Wiley & Sons.

[6] Michie D., Spiegelhalter D.J., Taylor C.C (eds.) *Machine learning, neural and statistical classification*, Ellis Horwood, 1994.

[7] An Introduction To Support Vector Machines and other Kernel Based Learning Methods Cristianini, N. and Shawe-Taylor, Cambridge University Press, 2000.

[8] A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin April 15, 2010 http://www.csie.ntu.edu.tw/~cjlin

[9] Comparison of Fuzzy Diagnosis with K-Nearest Neighbor and Naïve Bayes Classifiers in Disease Diagnosos. Asaad Mahdi, Ahmad Razali, AliAlWakil, brain. broad research in artificial intelligence and neuroscience, vol 2, no 2 (2011)

[10] The impact of datamining Techniques on Medical daignostics, Data Science Journal, Volume 5, 19 October 2006

[11] 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

[12] Mitchell, T. (1997) Machine Learning, McGraw Hill.

[13] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32. doi:10.1023/A:1010933404324

[14] Altmann A, Tolosi L, Sander O, Lengauer T (2010). "Permutation importance:a corrected feature importance measure". Bioinformatics. doi:10.1093/bioinformatics/btq134

[15] Tolosi L, Lengauer T (2011). "Classification with correlated features: unreliability of feature ranking and solutions.". Bioinformatics. doi:10.1093/bioinformatics/btr300.

[16] H.B.Burke, D.B.Rosen, P.H.Goodman, "Comparing artificial neural-networks to other statistical methods for medical outcome prediction," *IEEEJournal*, 1994, pp.2213-2216

[17] O.H.Beahrs, D.E.Henson, R.V.P.Hutter, B.J.Kennedy, "Manual for staging of cancer," 4th Edition, Philadelphia: JB Lippincott, 1992

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

[19] 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.

[20] Michalski, R. and Stepp, R. (1983), "Learning From Observation: Conceptual Clustering," In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors,Machine Learning: An Artificial Intelligence Approach, vol. 1, pp. 331-363,TIOGA Publishing Co.