ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*
*Journal DOI: 10.21474/IJAR01*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**RESEARCH ARTICLE**

# Framework for Diagnosing Hepatitis Disease using Classification Algorithms.

**S. Pushpalatha[1] and Dr. Jadesh Pandya[2].**

1. MCA Department, S. K. Patel Institute of Computer Studies, Gandhinagar, Gujarat, India.
2. BISAG, Gandhinagar, Gujarat, India.

## *Manuscript Info*

***Manuscript History:***

Received: 12 May 2016
Final Accepted: 19 June 2016
Published Online: July 2016

***Key words:***
Accuracy, Classification technique, Diagnosis, Hepatitis, Neural Network, Navie Bayes, Support Vector Machine, Sensitivity and Specificity.

***Corresponding Author***

**S. Pushpalatha.**

## *Abstract*

Hepatitis is a liver disease which affects major population in all age group. Diagnosing the disease is the challenging task for many public health physicians. In this study, we propose the performance and usage of classification techniques like Neural Network, Naive Bayes and Support Vector Machine to predict the accuracy in diagnosing hepatitis disease. Hepatitis dataset is taken from the Indian Liver Patient Dataset to train the proposed model. The performance of the proposed classification techniques is compared and evaluated based on the sensitivity and specificity to obtain the accuracy of the model. Based on the evaluation, Neural Network obtained improved accuracy rate compared to the other techniques thereby, minimizing the time duration in diagnosing of the hepatitis disease with reduced possible errors.



## Introduction:-

Viral Hepatitis is one of the most common infectious diseases, which affects majority of the population in all age group. Every year 1.5 million deaths are caused due to this infectious disease [12]. Viral hepatitis is an inflammation and damage caused in the liver which is categorized into 5 different types called HAV, HBV, HCV, HDV and HEV [13]. For all these types the target organ is liver which has specific symptoms.

Hepatitis diagnosis is made by the routine blood testing. Beside clinical test, machine learning and pattern recognition are widely used for the early diagnosis of the hepatitis disease. Physician mostly takes decision by comparing the result value of the previous patient. Hence it becomes difficult for the patient to take accurate decision in the diagnosis process. The automated diagnosis process can be achieved by the classification algorithms [9].

The rest of the paper is organized as follows: Section 2 covers the related work. Section 3 mentions the different classification techniques used in this work. The dataset description and methodology is discussed in Section 4 and 5. Analysis and result are given in the section 6. Finally, in section 7 the conclusion of the research is mentioned.

## Related Work:-

Yılmaz Kayaa et al. [1] implemented a new hybrid medical decision support system based on rough set (RS) and extreme learning machine for the diagnosis of hepatitis disease. Javed Salimi Sartakhti et al. [2] presented a novel machine learning method using hybridized Support Vector machine and simulated annealing for hepatitis diagnosis. They used a stochastic method for difficult optimization problems.

Duygu et al. [9] proposed an intelligent hepatitis diagnosis system using Principle Component Analysis and Least Square Support Vector Machine Classifier. G. Sathya Devi [13] proposed the application of CART algorithm in Hepatitis Disease Diagnosis using decision trees C4.5 algorithm, ID3 algorithm and CART algorithms.

A.H.Roslina et al. [10] Implemented a prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method. Enas M. F. El Houby [4] proposed a framework for prediction of HCV patient's response to the treatment of HCV from clinical information. Fadl Mutaher et al. [7] presented the comparative analysis in the prognostic of hepatitis data using Rough set technique over Multi- layer Neural Network using back-propagation algorithm.

Hui-Ling et. Al.[5] developed a new medical diagnostic method using local fisher discriminant analysis and support vector machines for hepatitis diagnosis problem. Ihsan Omur Bucak et.al.[6] implemented the diagnosis of liver disease by using CMAC neural network approach to shorten the medical diagnosis process and help the physician in the complex cases.

Rong-Ho Lin[14] presented an Intelligent model for liver disease diagnosis using classification and Regression tree and case-based reasoning.

## Classification Algorithms:-

Classification is used to classify data into predefined class labels. To classify data, a classification algorithm creates a classification model consisting of classification rules. Classification can be used to diagnose hepatitis and prognosis based on symptoms and health conditions [9]. In this there are two steps process consisting of training and testing. The training set is used to builds a classification model by analyzing training data containing class labels. The second step is testing. It examines a classifier using testing data for accuracy in which the test data contains the class labels or its ability to classify unknown objects for prediction. There are many classification algorithms like Naive Bayes, Decision Tree, J48, Neural network, etc. The techniques used in this research are as follows,

**Neural Network:-**

Neural Network (NN) is an adaptive system that learns from examples using interconnected processing nodes. Artificial neural networks serve as general purpose mechanisms for training a machine by examples[3]. Neural networks are classified as artificial intelligence because of their ability to learn and its basis in biological activities of the human brain. They are modeled after the human brain, which are perceived as highly connected network of neurons termed as nodes. It has three parts (layers): an input layer, a hidden layer and the output layer. The number of input, hidden, and the output nodes is referred to as the neural network topology or the network architecture.

**Naïve Bayes:-**

The naïve Bayes (NB) classifier is based on Bayes rule of conditional probability[8]. It assumes that all attributes of the dataset are independent of each other given the context of the class. The assumption of the conditional probability may be expressed as

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m \mid \theta) = \prod_{i=1}^{m} p(X_i = x_i \mid \theta) \quad (1)$$

Naïve Bayes model has shown itself to be more consistently robust to violation of the conditional independence assumption. Naïve Bayes uses a single scan of the data set to estimate the components.

**Support Vector Machine:-**

Support Vector Machines (SVM) is an algorithm that attempt to find a linear separate (hyper plane) between the data points of two classes in multi-dimensional space [10]. SVM are well suited to dealing with interactions among features and redundant features. Viewing the input data as two sets of vectors in an n-dimensional space, SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two datasets. To calculate the margin, two parallel hyper-planes are constructed, one on each side of the separating hyperplane, which is "pushed up against" the two dataset.

## Dataset Description:-

For this research dataset are collected from Indian Liver Patient Dataset [15].

Table I. Dataset description

| Parameters | Normal Range |
|---|---|
| Age | 0 – 100 |
| Gender | Male or Female |
| Total bilirubin | 0.3 - 1.0 mg/dl |
| Direct Bilirubin | 0 - 0.2 mg/dl |
| ALP:<br>Alkaline Phosphatase | Male: 25 - 120 U/L<br>Female: 25 - 90 U/L |
| ALT: alanine transaminase<br>(SGPT) | Men: 10 - 50 U/ L<br>Female: 5 -38 U/L |
| AST: aspartate transaminase (SGOT) | Men: 8 -40 U / L<br>Female: 6 -34 U/L |
| Total protein | 5.8 - 8 gm/dl |
| Albumin | 3.5 - 5 gm/dl |
| Ration Albumin:Globuine | 0.7 - 1.4 |
| Class | 0, 1 |

## Methodology:-

The methodology of this research is performed by first collecting the dataset and the collected dataset is initially pre-processed to find the normality of the dataset by using scaling method. Then to find the linearity of the dataset generalized multi-linear mixed model is used. In the next phase model processing is done by splitting the dataset into training and test dataset. It is then applied in different classification algorithms to obtain the accurate results. In the final phase evaluation is done by cross-validation technique to obtain maximum accuracy within an optimal time Select the model which gives best accuracy rate, the obtained model is then evaluated with different evaluation parameters. Fig. 1. Represents the framework for diagnosing hepatitis disease.

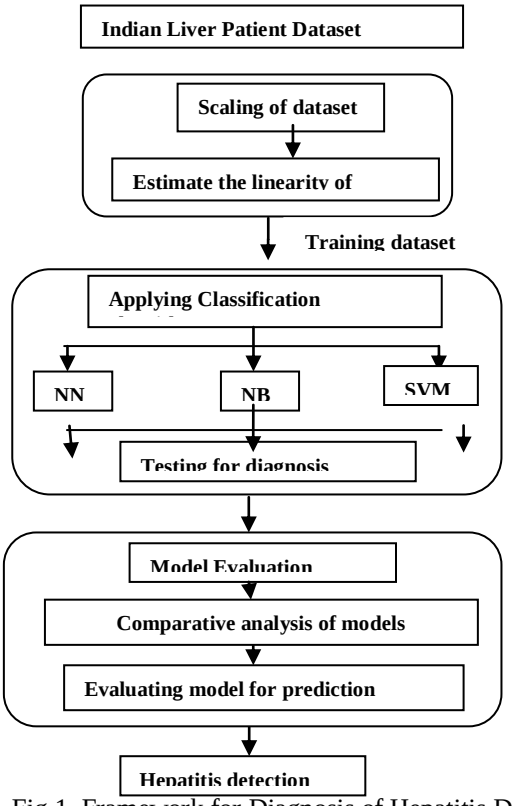


Fig 1. Framework for Diagnosis of Hepatitis Disease

Following steps are followed in the proposed work, which are represented below,

**Step –1: Pre-Processing**

The dataset used for the experiment are first normalized using scaling method (2).

$$Y_i = \frac{(Y_i - Y_{min})}{(Y_{max} - Y_{min})} + 1 \quad (2)$$

From the obtained result calculate the p-value using (3), if the p-value is less than 0.05 reject the hypothesis.

$$x = (m/n\text{-}P) / SqRt[p(1\text{-}P)/n] \quad (3)$$

**Step – 2: Estimating the linearity of the dataset**

In the next step calculate the linearity of the model with different dataset. The generalized multi-linear mixed model for normally distributed vector elements with random effect using binomial type are used to calculate the predicated result using (4).

$$\gamma_i = glm(E(y|\gamma)) = Xi1 + \sum(\beta X12 * \beta X13 * \ldots * \beta X1n) \quad (4)$$

where, Yi = linear predicted, γ = vector of data elements, β = intercept weight of the dataset elements.

**Step -3: Finding the accuracy**

Accuracy is calculated using Standard Error (SE), R-Squared value ($r^2$) and Akaike Information (AIC), which is represent in the formula (5), (6), (7).

$$SE = sqrt[\ p * (\ 1 - p\ ) / n\ ] \quad (5)$$

where p is the sample proportion, n is the sample size

$$r^2 = 1 - \frac{SS\ Error}{SS\ Total} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (6)$$

where SS Total is Total Sum of Square and SS Error is Error Sum of Square

$$AIC = 2k - 2 \times \ln(L) \quad (7)$$

where $k$ is the number of model parameters. And $\ln(L)$ is the log-likelihood function for the statistical model. Lesser the value more the likelihood.

**Step – 4:Appling in classification algorithms**

Read the pre-processed dataset and split the dataset as 50 -50% as training and test set. Apply the training set dataset in the NeuralNetwork, Navie Bayes and Support Vector Machine

**Step – 5: Performance measure**

The performance of the algorithm is examined by evaluating the sensitivity, specificity analysis, precision and classification accuracy using (8),(9),(10), (11).

Sensitivity = t-positive / positive (8)

Specificity = t-negative / negative (9)

Precision = t-positive / (t-positive + f-positive) (10)

Accuracy = sensitivity * [positive / (positive + negative) + specificity * negative / (positive + negative) (11)

Table II. Normalized dataset

| Variable | Min | Max | p-Value |
|---|---|---|---|
| Age | 0 | 99 | 0.1036 |
| Total Bilirubin | 0.4 | 75 | 0.4795 |
| Direct Bilirubin | 0.1 | 19.7 | 0.9744 |
| Alk. Phosphate | 63 | 2110 | 0.6397 |
| SGPT | 10 | 2000 | 0.1542 |
| SGOT | 10 | 4929 | 0.01942 |
| Total Protein | 2.7 | 9.6 | 0.99219 |
| Albumin | 0.9 | 5.5 | 0.1971 |
| Ration Albumin:Globuine | 0.3 | 2.8 | 0.2665 |
| Class | 0 | 1 | 0.5446 |

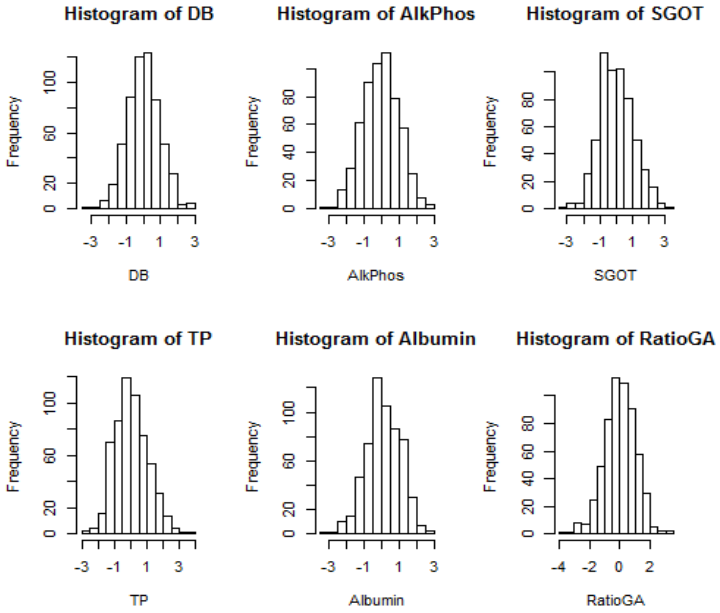


Fig 2. Histogrm representaion of normalized data

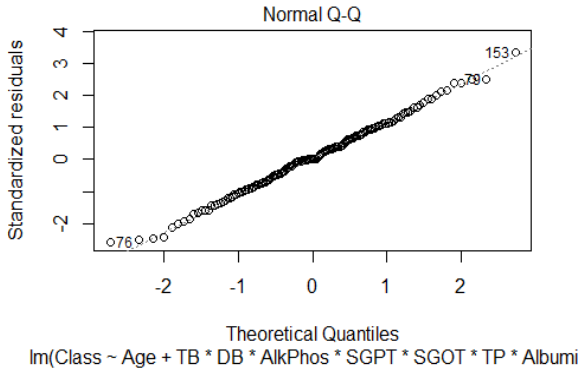


Fig 3.NormalQQ of Hepatitis dataset

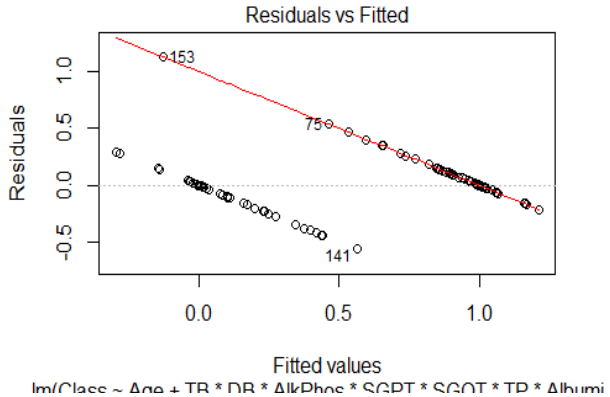


Fig 4. Residual vs. Fitted for Hepatitis dataset

Table III. Results obtained for the hepatitis dataset

| **Algorithms →** | **Neural Network** | **Naive Bayes** | **SVM** |
|---|---|---|---|
| Accuracy (%) | 98.07 | 82.58 | 84.52 |
| Kappa | 0.9214 | 0.4974 | 3853 |
| Sensitivity | 96.87 | 65.62 | 31.25 |
| Specificity | 98.24 | 86.99 | 98.37 |
| Positively Predicted Value | 92.12 | 90.68 | 83.33 |
| Negatively Predicted value | 99.18 | 56.76 | 84.61 |
| Prevalence | 0.24 | 0.2065 | 0.206 |
| Detection Rate | 0.21 | 0.1355 | 0.064 |
| Detection Prevalence | 0.2112 | 0.2387 | 0.077 |
| Balanced Accuracy | 97.56 | 76.31 | 64.81 |
| System (sec) | 0.2 | 0 | 0.2 |
| Elapsed (sec) | 0.85 | 0.01 | 0.01 |

## Analysis and results:-

The hepatitis disease dataset of 155 patients are used in this research. The dataset consists of 11 attributes and a class variable with two possible values which are shown in Table – 1. R language is used to perform the experimental work for this research. The dataset is normalized using scaling method and p-value is calculated the results are displayed in the Table – 2. The normalized dataset are represented in the histogram in Fig. 2. Fig. 3 shows the normalized Q-Q which is getting fitted in the normal line and the Fig. 4. Represents the residual Vs fitted values for the hepatitis dataset.

The linearity of dataset are calculated using generalized multi-linear mixed model for normally distributed vector elements with random effect using binomial type are used to calculate the predicated result. The accuracy generalized multi-linear mixed model is estimated using Standard Error, R-Squared value and Akaike Information.
Then the model is processed by splitting the dataset as 50 -50% as training and test set. Apply the training set dataset in the neural network, Naive Bayes and Support Vector Machine.
The performance of the algorithm is represented in Table – 3, which shows accuracy, kappa, sensitivity, specificity, positively and negatively predicted value, prevalence, detection rate and elapsed rate. Fig. 5. Shows the proposed algorithm results.

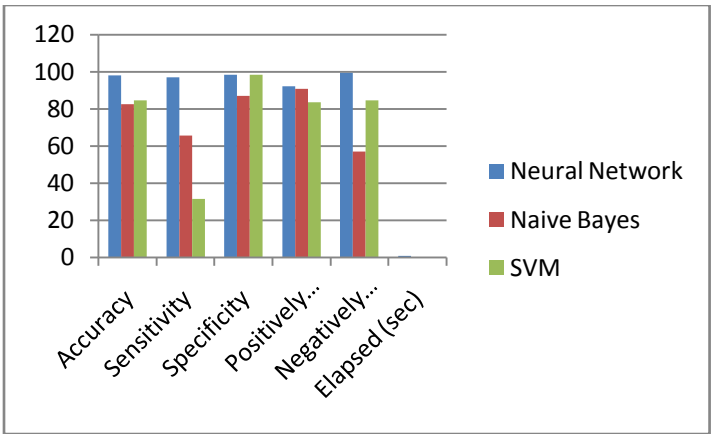


Fig 5. Proposed algoithm results

## Conclusion:-

This research is conducted for designing the framework for Diagnosing Hepatitis using three different classification algorithms including Neural Network, Naive Bayes and Support Vector Machine on Hepatitis data sets. Among these three algorithms neural network gives better results compared to other two algorithms. The proposed work can still enhanced by considering many other algorithms to achieve better results.

## References:-

1. Yılmaz Kaya, Murat Uyar, A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. 2013 Elsevier, 3429–3438
2. Javad Salimi Sartakht, J. S. (2011). Hepatitis disease diagnosis using a novel hybrid method. Elsevier 2011, 570-579.
3. YugalKumar and G. Sahoo, "Prediction of different types of liver disease using rule based classification model", Technology and Health Care (2013), 417 – 432
4. Enas M. F. El Houby, "A Framework for Prediction of Response to HCV Theraphy Using Different Data Mining Technique", Hindawi Publishing Corporation, Volume 2014, 181056.
5. Hui-Ling Chen, Da-You Liu, Bo Yang, Jie Liu, Gang Wang. A new Hybrid method based on local fisher discriminant analysis and support vector machine for hepatitis disease diagnosis, Elsevier, 2011, 11796-11803.
6. Ihsan Omur Bucak, Semra Baki, Diagnosis of liver disease by using CMAC neural network approach. 2010 Elsevier, 6157–6164 Rong-Ho Lin, An intelligent model for liver disease diagnosis, 2009, Elsevier, 53-62.
7. Fadl Mutaher Ba-Alwi, H. M. (Volume 4, Issue 8, August-2013). Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach. International Journal of Scientific & Engineering Research, 680-685.
8. S. B. Kotsiantis, Increasing the Classification Accuracy of Simple Bayesian Classifier, AIMSA, pp. 198-207, 2004
9. Duygu ȘCalisir, Esin Dogantekin, A new intelligent hepatitis diagnosis system: PCA–LSSVM, 2011 Elsevier 2011, 10705–10708.
10. H. Roslina et.al. "Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method", IEEE 2010, pg. no. 2209 – 2211
11. W.M. Lee, Hepatitis B virus infection, N. Engl. J. Med. 337 (1997) 1733.
12. J. Cohen, The scientific challenge of hepatitis C, Science 285 (1999) 26.
13. G.Sathyadevi, Application of CART Algorithm in Hepatitis Disease Diagnosis, 2011 IEEE, 1283-1287
14. Rong-Ho Lin, An intelligent model for liver disease diagnosis, 2009, Elsevier, 53-62.
15. http://networkrepository.com/Indian-liver-patients.php.