Clustering and Detection of Liver Disease in Indian Patient Using Machine Learning Algorithms

Manoj Kumar D P [1], Dr.AnandaBabu J [2] , Dr.Raviprakash M L[3], Manjunatha B N[4]

[1]Assistant Professor, Department of Computer Science & Engineering, Kalpataru Institute of Technology, Tiptur-572 201, India.
[2] Associate Professor Department of Information Science & Engineering, Malnad College of Engineering, Hassan-573 202, India.
[3]Associate Professor Department of Computer Science & Engineering, Kalpataru Institute of Technology, Tiptur-572 201, India.
[4] Assistant Professor, Department of Computer Science & Engineering, R L Jalappa Institute of Technology, Doddaballapur-561 203, India.

Abstract –**In Present Days, Machine Learning plays an important role in field of disease classification and prediction regarding to various organs like heart ,kidney ,liver ,stomach etc..to predict automated disease detection using various algorithms,i.e., Naïve Bayes, K-means, and Support Vector Machine. The study concentrates on liver disease-related health care data set and used for comparative performance measurement of the three techniques mentioned above. K-means is used for performing clustering on the training dataset and Naïve-Bayes, Support Vector Machine to predict the test cases using training dataset. Results describe Correct Classifications, Misclassifications, accuracy performance metrics to compare their prediction accuracy. Results derive that SVM classifier provides better accuracy of 81% than Naïve-Bayes.**

Keywords –Machine Learning, Naïve-Bayes,K-means,Support Vector Machine, Liver dataset, Clustering.

1. INTRODUCTION

The liver plays an important role in many bodily functions from protein production and bloodclotting to cholesterol, glucose (sugar), and iron metabolism. It has a range of functions,including removing toxins from the body, and is crucial to survival. The loss of those functions can cause significant damage to the body. When liver is infected with a virus injured by chemicals, or under attack from own immune system, the basic danger is the same that liver will become so damaged that it can no longer work to keep a person alive.

The work focuses on three different machine learning techniques, i.e., Naïve Bayes, Kmeans, and Support Vector Machine, propagation to compare their prediction accuracy and computational complexity. The study concentrates on liver disease-related health care data set and used for comparative performance measurement of the three techniques. The utilization of medicinal datasets has pulled in the consideration of specialists around the world. Machine Learning methods have been broadly utilized as a part of creating choice emotionally supportive networks for ailments forecast through an arrangement of therapeutic datasets. Grouping systems have been broadly utilized as a part of the restorative field for exact order than an individual classifier. Liver malady is a sort of harm to or illness of the liver.

The Liver Disease Detection Problem includes modeling past liver disease patient data with the knowledge of the ones that has liver disease or not. This model is then used to identify whether a new person is prone to get liver disease or not. Our aim here is to detect 100% of the liver disease prediction.

To solve the problems facing physicians in diagnosis of liver diseases. Experience has shown that many patients suffering from liver disorder die daily as a result of misdiagnosis of the diseases. Early prediction of liver disease is very important to save human life and take proper steps to control the disease. This research work explores the early prediction of liver disease using deep learning techniques. The liver disease dataset which is select for this study is consisting of attributes like total bilirubin, direct bilirubin, age, gender, total proteins, albumin and globulin ratio.

Following are the objectives performed by the system:

- The application must provide user interface for doctors input object of the prescription.
- The application should have the capability for preprocessing of the given input.
- The System should be capable to detect the chances of lever disease using past patient data.

2. RELATED WORK

Ashwani Kumar., [1] "Categorization of Liver Disease Using Classification Techniques" presents an approach with the great algorithms to differentiate healthy liver patient and unhealthy liver patient with exact values. They used RandomForest, REP as the classifier to predict the patient liver information with performance metric 80% accuracy, it is considered as best classifier to characterize liver disease in patients.

ChandrasegarThirumalai., [2] "Cost Optimization using Normal Linear Regression Method for Breast Cancer Type I Skin" finished up with the various machine learning algorithms like simple linear regression algorithms and calculating correlations between attributes using Persons Correlation strategy, attributes having values greater than 0.55 are taken into consideration. Using linear regression algorithms, we have obtained the linear equations to state the relations amoung the attributes.

Harsha Pakhale.,[3] "Development of an Efficient Classifier for Classification of Liver Patient with Feature Selection"has interpretated that single model doesnot fulfill the exactness for each input data. For different liver patient attributes, we characterize different models to build up an healthy classifier to predict the liver diseases. Algorithms like CART, RF predicted the liver diseases with an accuracy of 75.34% compared to individuals and other ensemble machine learning algorithms.

Dr. S. Vijayarani., [4] "Liver Disease Prediction using SVM and Naïve Bayes Algorithms" presented that machine learning algorithms like SVM and Naïve Bayes are used as classifier for prediction of Liver disease, with interpretations stating that SVM classifier is considered as best classifier with classification report, which includes accuracy, recall, precision and fi-score results in better accuracy compared to GaussianNB method.

Anju Gulia., [5] "Liver Patient Classification Using Intelligent Techniques" presents intelligence techniques for prediction of liver diseases. Various Classification algorithms like J48, MLP,SVM, RF are used in the classification purpose. Developed Phase-wise classification by collecting the data from UCI repository and extracting the various features like direct bilirubin, total bilirubin and proteins etc..SVM algorithm is considered as better performance classifier with various feature extraction sets.It predicts with an accuracy of 71.869%.

Kalyan Nagaraj., [6] "NeuroSVM: A Graphical User Interface for Identification of Liver Patients" Illustrates an Neuro-SVM cross breed derived from SVM with ANN(artificial neural network) as backpropagation method was produced for dataset taken for liver diseases to predict the organic liver healthy or not.

JankisharanPahareeya., [7] , "Liver Patient Classification using Intelligence Techniques" proposes an Random Forest algorithm over software testing examining 200% training and testing datasets with prediction accuracy of 81.67%.

Bendi Venkata Ramana.,[8] "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis" has proposed an Multi-Layer perceptron with irregular subset attributes choosen from UCI liver dataset repository. Experimentation suggested that having multiple times of epochs determine better accuracy for liver dataset.

Kotsiantis. S.B., [9] "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis" advises todevelop the exactness of Naïve Bayes Classifier demonstration. used an discretization strategy to derive better performance metrics in less time.

3. PROPOSED WORK

3.1 OVERVIEW

The research proposes various machine learning techniques like k-means for clustering the entire dataset , which comprises of various attributes likeglobulin ratio, albumin,proteins, age, gender, total and direct bilirubin etc...Also, there are modeling strategies and results based on certain data which is splitted into training data and testing data.Depending on the dataset, we perform two functions:

1.Clustering: Divide the dataset into group of clusters to perform various asymptotic functions like mean, variance and standard deviation.

2.Detection: Choose the training and testing data, split the train and test dataset using TRAIN_TEST_SPLIT() function and apply the classification and prediction algorithms like Naïve-Bayes and SVM to predict whether the patient has healthy liver or not. Results include performance metrics like accuracy, correct classifications and misclassification data,

System Architecture design-identifies the overall structure for the Liver disease as shown in below diagram.

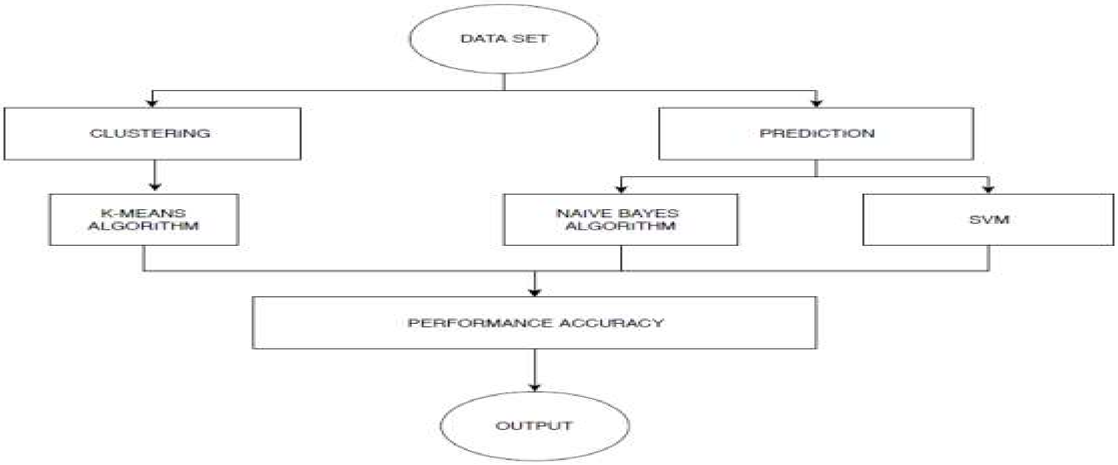


Figure 1: Overall Architecture

The Various UML diagrams like Dataflow diagram, Usecase Diagram, Sequence diagram are shown below.

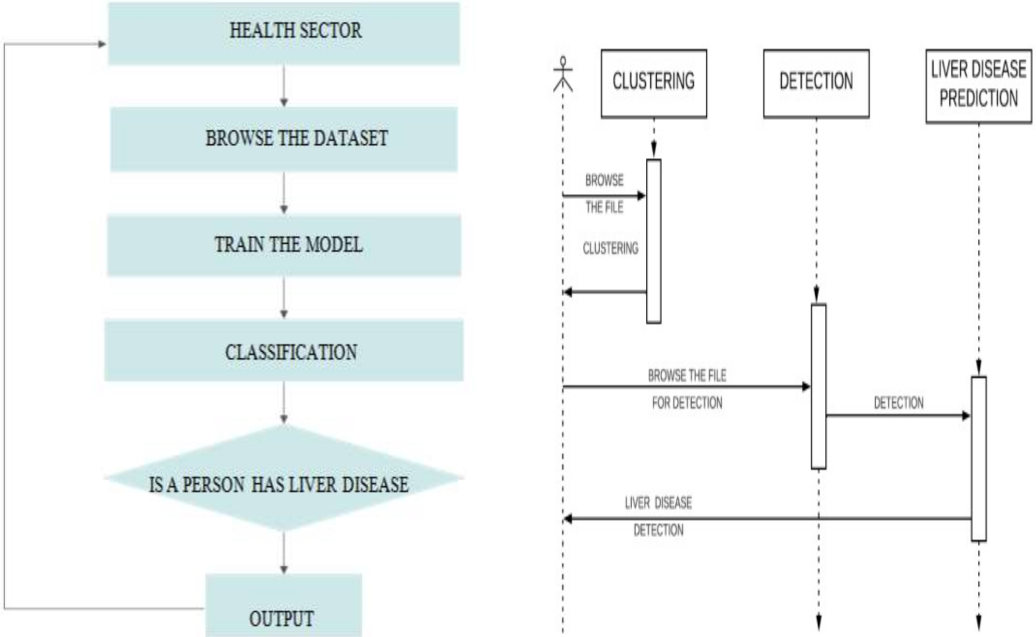


Figure 2: Data flow diagram and Sequence diagram

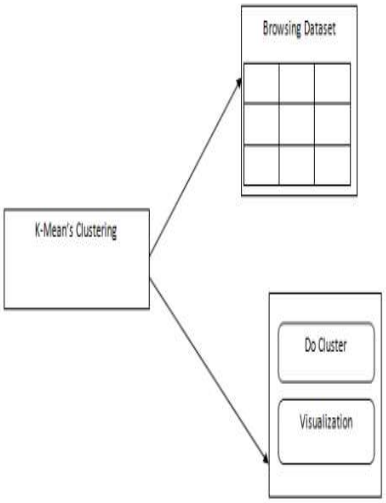
3.2 K-MEANS CLUSTERING

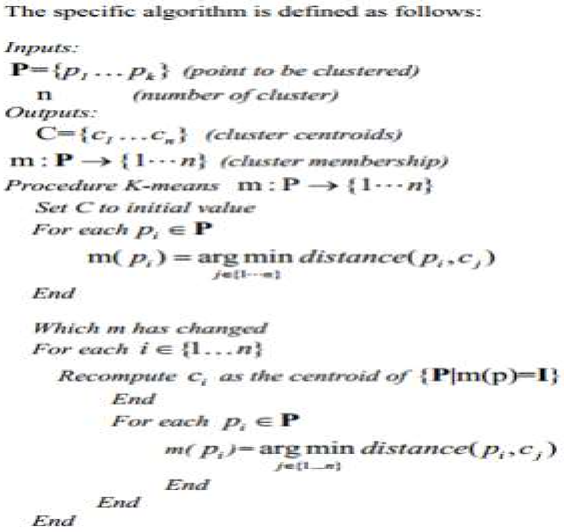
K-means is a clustering algorithm which divides the dataset into multiple clusters. “K” stands for number of divisions or clusters. We use multiple distance measures techniques like Euclidean distance and Manhattan distance function**.**

The sample Dataset is initially divided into K multiple clusters and the observations are randomly assigned to the clusters. For each sample:

- Find the personnel details of customer to the centroide of the cluster.
- IF the data considered is closer to its own cluster THEN Accept ELSE reject
- Repeat steps 1 and 2 until no same observations are moved from one cluster to another
- When step 3 stops, the clusters are stable and each sample is assigned to any of K clusters which results in the lowest possible distance to the centriode of the cluster

A Person or Medical expert wants to analyze the data in order to know which personis prone to get liver disease or not. There is an enormous amount of data which one can actually retrieved from the healthsector. The data can actually be used to detect the chance of liver disease. The proposed system categorizes chance of getting liver diseae based on some personal information. Detection is then carried out to know the liver disease. The system takes all these information first as the input and points out person who are prone to get liver disease. A training dataset is being prepared beforehand based upon the person previous history. The system uses the concept of k-means clustering to cluster the person based upon the similarities or the patterns they share among each other naïve bayes and SVM used to classify and detect the chance of getting liver disease with the training dataset and testing dataset and after classifying it will detect the liver disease.




The specific algorithm is defined as follows:

```
Inputs:
 $\mathbf{P}=\{p_1 \cdots p_k\}$ (point to be clustered)
    n          (number of cluster)
Outputs:
    $C=\{c_1 \dots c_n\}$ (cluster centroids)
 $m:\mathbf{P}\rightarrow\{1\cdots n\}$ (cluster membership)
Procedure K-means $m:\mathbf{P}\rightarrow\{1\cdots n\}$
    Set C to initial value
    For each $p_i \in \mathbf{P}$
        $m(p_i)=\arg\min_{j\in\{1\cdots n\}} distance(p_i,c_j)$
    End
    Which m has changed
    For each $i \in \{1\dots n\}$
        Recompute $c_i$ as the centroid of $\{\mathbf{P}|m(p)=I\}$
            End
            For each $p_i \in \mathbf{P}$
                $m(p_i)=\arg\min_{j\in\{1\_n\}} distance(p_i,c_j)$
                End
          End
    End
```

Figure 3: Clustering

3.3 NAIVES-BAYES CLASSIFIER:

Classification techniques are maximum ideal for predicting or describing facts sets with binary or nominal classes. They are less effective for ordinal classes because they do now not consider the implicit order among the classes. Other kinds of relationships, which includes the subclass–super class relationships among classes also are unnoticed. The remainder of this bankruptcy focuses best on binary or nominal magnificence labels.

A class technique is a scientific method to constructing type fashions from an enter information set. Examples encompass choice tree classifiers, rule-based totally classifiers, neural networks, aid vector machines, and Naive Bayes classifiers. Each approach employs a mastering set of rules to pick out a model that quality fits the connection among the attribute set and sophistication label of the enter records. The version generated by using a studying algorithm ought to each suit the input records properly and efficiently are expecting the magnificence labels of facts it has never visible earlier than. Therefore, a key goal of the gaining knowledge of set of rules is to construct fashions with correct generalization capability, models that accurately expect the elegance labels of previously unknown facts.

The tree has 3 styles of nodes:

• A root node that has no approaching edges and at least zero active edges

• Internal node, each of which has precisely one approaching edge and at least two active edges.

• Leaf or terminal node, each of which has precisely one approaching edge and no friendly edges.

3.4 SUPPORT VECTOR MACHINE:

Support Vector Machine (SVM) is a machine learning algorithm for classification and prediction purposes. In this algorithm, we plot each data item as a point in n-dimensional space with each value as an particular coordinate plotted as a hyper-plane.

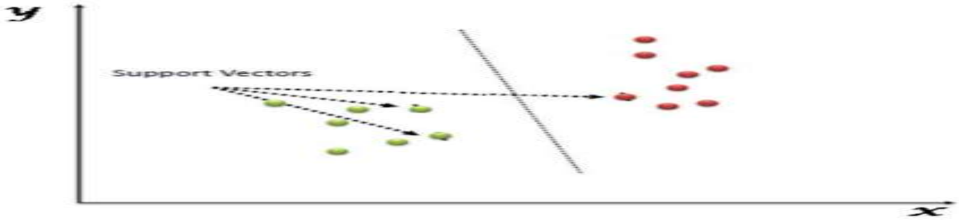


Figure 4: SVM Model

An SVM creates hyperplanes that have the largest margin in a high-dimensional space to separate given data into classes. The margin between the 2 classes represents the longest distance between closest data points of those classes.

STEP1: Select the feature sets from different classes of data

STEP2: Calculate the intersection points of each class of feature and plot, repeat for all the features of data.

STEP3: Remove the features which are intersecting and data of all the classes.

STEP4: Plot the hyper planes for the remaining points.

STEP5: Calculate the distance of the hyper planes in different class of objects.

STEP6: Select the hyper plane which is consistent for each class of data

4. RESULTS AND DISCUSSION

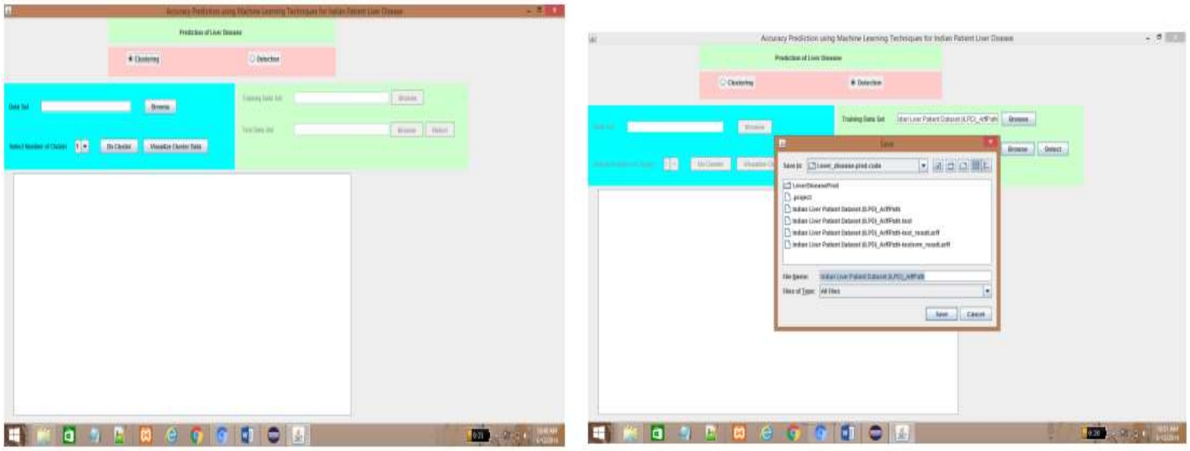


Figure 5: Front Page and Dataset Selection Page.

Front page containes two options that is Clustering and Detection. User can select the option based on their requirements either Clustering or Detection. User can browse both training and testing data set. The training data and testing data are stored in .arff extension file. In Detection user has to browse both training and testing data set. In Clustering user has to browse only training data.

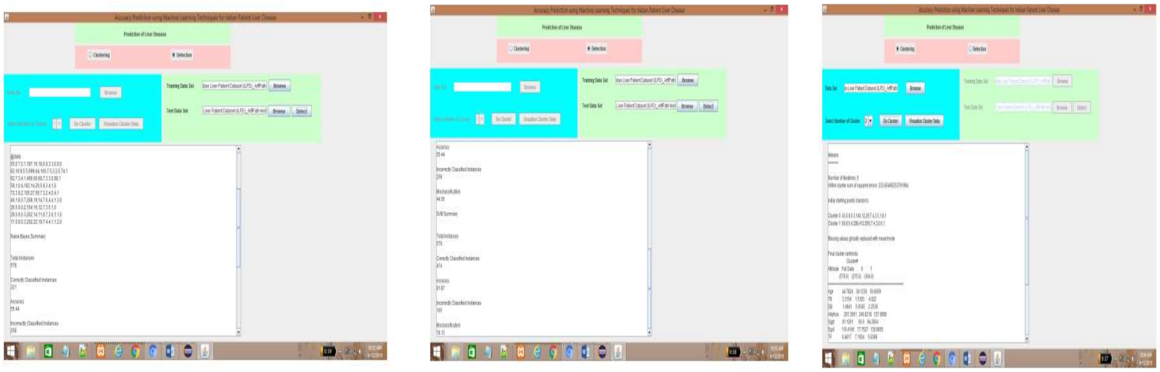


Figure 6: Summary of Naïve Bayes ,SVM and K-Means

User can view the summary of Naïve Bayes. It contains Total instances, Correctly classified instances, Accuracy, Incorrect classified instances, Miss classifications

In summary of K-means algorithm it will contain number of iterations, within cluster sum of squared errors, initial starting point, final cluster centroid. Finally it produces the percentage of clustered instances.

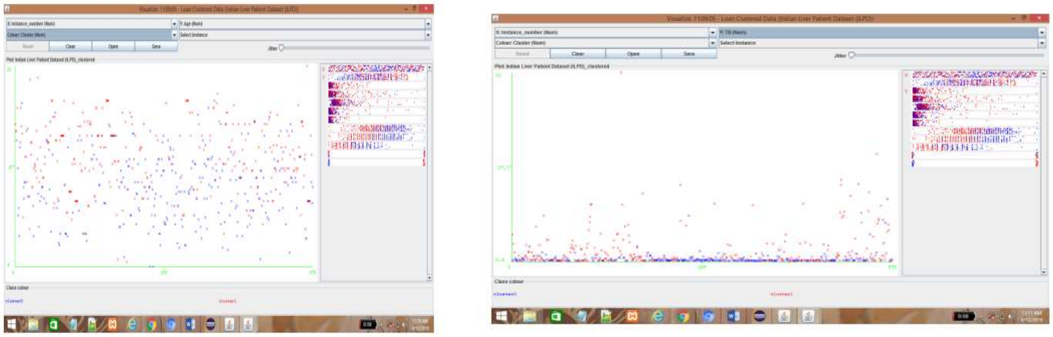


Figure 7: Visualization of Cluster with respect to Age and TB

5. CONCLUSION

Classification is a major machine learning technique which is used in healthcare services for automated disease detection. This work uses K-means clustering algorithm to partition the training dataset into different clusters and Naïve-Bayes and SVM for prediction of liver disease. We compare the two above mentioned algorithms, concludes that SVM is best algorithm with better accuracy than Naïve-Bayes classification. On the other hand, while comparing the execution time, the Naïve Bayes classifier needs minimum execution time.

REFERENCES:

[1] Ashwani Kumar, Neelam Sahu, "Categorization of Liver Disease Using Classification Techniques", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 5 Issue V, May 2017, IC Value: 45.98 ISSN: 2321-9653.

[2] ChandrasegarThirumalai, IEEE Member, Rashad Manzoor, "Cost Optimization using Normal Linear Regression Method for Breast Cancer Type I Skin", International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.

[3] Harsha Pakhale, Deepak Kumar Xaxa, "Development of an Efficient Classifier for Classification of Liver Patient with Feature Selection", International Journal of Computer Science and Information Technologies, Vol. 7 (3), 2016, 1541-1544.

[4] Dr. S. Vijayarani, Mr.S.Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015.

[5] Anju Gulia, Dr. RajanVohra , Praveen Rani "Liver Patient Classification Using Intelligent Techniques", International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5110-5115.

[6] Kalyan Nagaraj and Amulyashree Sridhar, "NeuroSVM: A Graphical User Interface for Identification of Liver Patients", IJCSIT. 5(6): 8280-8284 (2014).

[7] JankisharanPahareeya, Rajan Vohra, Jagdish Makhijani , Sanjay Patsariya, "Liver Patient Classification using Intelligence Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014.

[8] Bendi Venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B, “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis”, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011 page no 101-114.

[9] Kotsiantis. S.B, “Increasing the Classification Accuracy of Simple Bayesian Classifier”, AIMSA, pp.198- 207, 2004.

[10] V.Kirubha, S.Manju Priya, “Survey on Data Mining Algorithms in Disease Prediction”, International Journal of Computer Trends and Technology (IJCTT) – Volume 38 Number 3 - August 2016.

[11] Bendi Venkata Ramana, Prof. M. S. Prasad Babu and Prof. N. B. Venkateswarlu, “Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis, International Journal of Computer Science Issues, ISSN: 1694 -0784, May 2012.

[12] A.S.Aneeshkumarand C.JothiVenkateswaran, “Estimating the Surveillance of Liver Disorder using Classification Algorithms”, International Journal of Computer Applications (0975 – 8887), Volume 57– No.6, November 2012.

[13] S. Karthik A, Priyadarishini, J. Anuradha and B. K. Tripathy (2011). “Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types”, Advances in Applied Science Research.

[14] Esraa M Hashem, Mai S Mabrouk (2014). “A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis”, American Journal of Intelligent Systems. 4(1): 9-14.