IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Evaluation Of Data Mining Techniques For Hepatitis Disease Classification Using Weka Algorithms Zeror, Multilayer Perceptron And Randomforest

Sushilkumar Rameshpant Kalmegh

Professor
Department of Computer Science, Sant Gadge Baba Amravati University,
Amravati (M.S.) 444 602, India

Abstract: In this atudy, a data mining technique for Hepatitis disease classification utilizing Weka is reviewed and evaluated. Weka is used to propose the data mining algorithms ZeroR, Multilayer Perceptron, and RandomForest for disease categorization. Weka is used to evaluate the algorithms. Any disease can be categorized using these algorithms. The accuracy of these algorithms is compared.

Index Terms **- Classification, Hepatitis, Multilayer Perceptron, Random Forest, Weka, ZeroR**

I. INTRODUCTION

There is an unavoidable decline in the percentage of data that people comprehend as its amount rises. Information is concealed throughout all of this data. Data mining involves electronic data storage and computer-assisted, or at least enhanced, search. This isn't exactly new either. The notion that patterns in data may be automatically found, recognized, verified, and utilized for prediction has long been a concept employed by statisticians, communication engineers, and economists. The startling rise in possibilities to identify patterns in data is novel. One subject that requires learning in a practical, rather than theoretical, sense is data mining. In order to assist explain the data, we are interested in methods for identifying and characterizing structural patterns in the data.

As the medical Internet Things are changing in the medical area; the amount of medical data is growing quickly, as is its diversity. Disease classification is the process of analyzing and classifying data pertaining to different diseases using machine learning techniques. The objective is to create models that, given input features, can reliably predict or categorize whether a certain disease will manifest or not. Classification of diseases aids in logical information organization. By classifying illnesses, we establish a methodical framework for comprehending and discussing them. It gives medical workers a common language to communicate and exchange knowledge about illnesses.

Data on hepatitis disease was used in this particular research work. Weka Classifier was used to compare ZeroR, Multilayer Perceptron, and Random Forest with test modes Cross-Validation 5 FOLDS and 10 FOLDS. There are six sections to this study. The introduction is covered in the first section, which is followed by the literature needed to analyze the techniques used. System design comes in third, followed by the analytical datasets. The performance analysis comes in fifth, followed by the conclusion.

II. LITERATURE SURVEY

Electronic health records (EHRs) are collected, integrated, and analysed to create appropriate treatment plans and treatment strategies for patients. Electronic medical records make it easier to diagnose, diagnose and treat diseases. However, existing EHR models are critical. Many problems such as big data, privacy and private information limit the development of centralized medical information. A new method and algorithm [1] has been proposed to manage the mining of medical data distributed to different areas (hospitals and clinics) using rules in the organization. This medical information cannot be transferred to another site. Therefore, the global computing need needs to be divided into local computing to accommodate information distribution in the network. The ability to decompose computation must be broad enough to handle different distributions and different participants in each instance of the computing world. Each data source is represented by an agent. The agent then deals with international names or exchanges some small orders with other agents or goes anywhere and does the usual work that can be done in any local location. The goal is to perform international operations with minimal communication or travel by participants in the network, in a manner that preserves the confidentiality and security of local information. The rule of thumb only applies to heart disease prediction using actual heart disease data. This actual data resides in different medical facilities and cannot be moved to a central location.

Recently, privacy-preserving methods for horizontal and vertical distributed systems are being developed based on agent-requiring star topologies [2]. Each site distributes information through an un-trusted third-party intermediary. This tool is responsible for creating integrated forecast models. However, these models have high communication costs due to frequent data exchange with central agents. Therefore, a new predictive model [1] was developed to collect data from multiple referral sources that do not transmit confidential patient data. Therefore, the patient's privacy is protected. A decentralized algorithm has been added to my organization rules based on trust rules and weighted rules based on big local data. Instead of creating specific rules for each local user, the integrated model generalizes the extracted rules to distributed storage using independent storage.

Each medical canter’s information is based on a different local area (hospital). Figure 1 shows an EHR model with integrity (multiple local areas such as pharmacies, hospitals, and clinics) and a hybrid model. Each care centre is located in a different area. Moving data from one local location to another is not a good solution for many reasons, including data ownership, privacy, communication overhead, and big data. But building decision-making models from medical records requires big data.

Information is used only through the extraction and decision making of many important confidential information in many healthcare applications [3]. The study in [4] provides a comprehensive look at the data mining process in the medical field by providing descriptive information about the information contained in data, data storage, business intelligence, and data mining. Data mining tools use patients' original medical records, test history, and personal information. Therefore, if symptoms are detected, prevention can be done and doctors can reduce the effect or prevent the disease. The use of data mining in healthcare has many advantages, such as detecting health insurance fraud, predicting and diagnosing diseases, managing hospital bedbugs, and creating smart treatments based on patient-appropriate recommended medications [5].

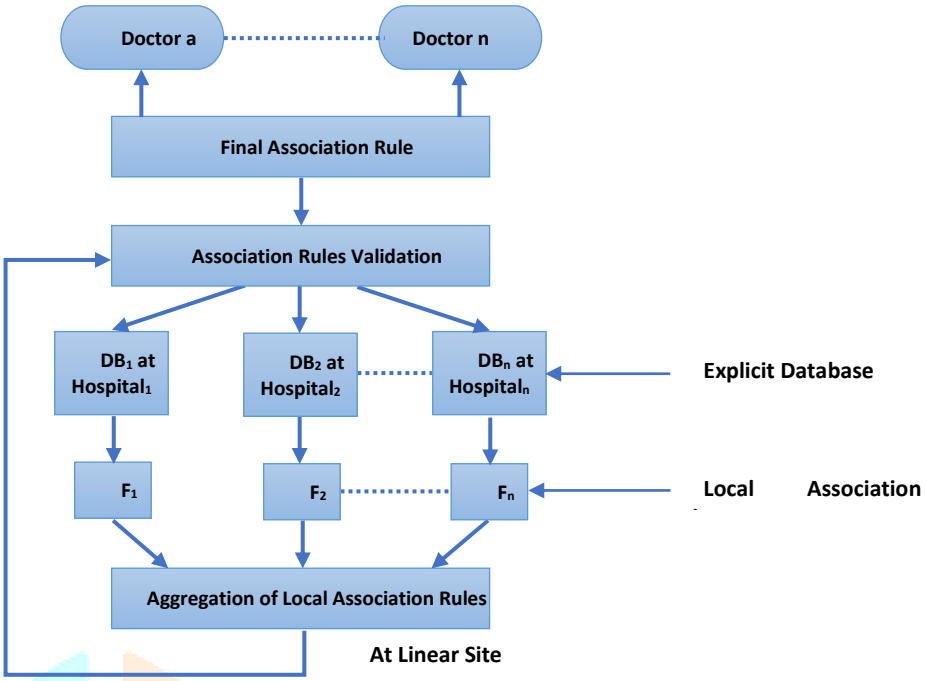

Figure 1: EHRs Model [1]

Various studies have been conducted to use machine learning techniques for disease diagnosis and prediction. Jiaxin et al. proposed an autonomous diagnosis system that uses an extreme learning machine on patient serum index data to predict the fibrosis stage and inflammatory Chronic hepatitis C activity grade [6]. A hybrid model for hepatitis prediction was put forth by Sushrutha et al. [7]. They've created a hybrid of genetic search multilayer perceptron technology and algorithm. The study [8] uses PCA-MLP to examine the effects of applying various folds for cross validation on missing values. Imputation technique looked into various ANN models for hepatitis prediction [9].

1.1 Classification

Classification can also mean categorization, which is the process of identifying, differentiating, and comprehending concepts and items. A classifier is an algorithm that does classification, particularly in a concrete implementation. There are instances when the term "classifier" also refers to the mathematical function that a classification algorithm uses to assign input data to a category. According to machine learning jargon, categorization is an example of supervised learning, or learning that takes place in the presence of a training set of accurately identified observations. Clustering, also called cluster analysis, is the corresponding unsupervised process that entails classifying data according to a measure of intrinsic similarity.

A data mining approach called classification provides a detailed manual for figuring out the result of a fresh data instance. The tree it generates is just that-a tree in which every node denotes a location where a choice must be taken in light of the input, and one must proceed to the next node and so on until one reaches a leaf that indicates the expected result. It may sound complicated, but it's actually very simple.

The question of whether categorization techniques that do not use a statistical model qualify as "statistical" is also up for debate. The term "classification" may be used differently in other domains. For instance, in community ecology, it typically refers to cluster analysis, which is a form of unsupervised learning, as opposed to supervised learning.

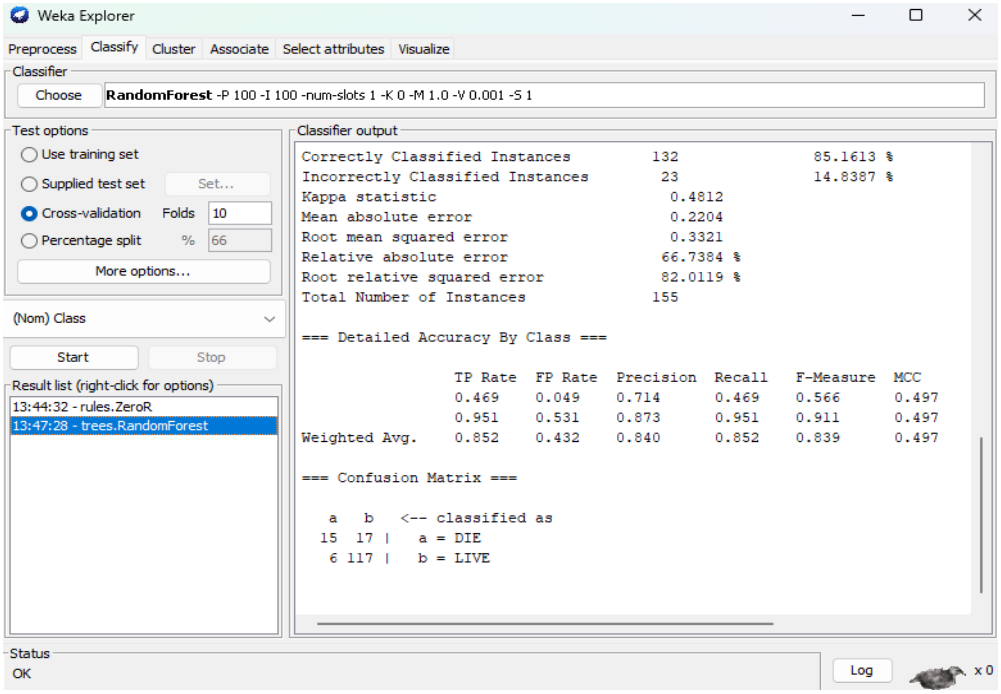

Figure 2: Processing of arff file by RandomForest Classifier on Test Mode Cross-Validation 10 Folds

1.2 ZeroR Classifiers

The ZeroR is the simplest method for classification, based on the target and ignoring all different predictions. The ZeroR classifier only predicts most groups (classes). Although ZeroR is not predictive, it is useful in determining performance as a benchmark for other classification methods. Creates frequency tables of targets and selects frequency values. Since ZeroR does not use any of these, there is nothing to predict contributions to the model. ZeroR can only predict most of the classes correctly. As mentioned before, ZeroR is only suitable for determining the performance of alternative distributions.

1.3 Multilayer Perceptron Classifiers

Multilayer perceptrons fall into the category of feed forward algorithms in that the components are combined with the initial weight in the weight and are affected by activation just like perceptrons. But the difference is that each combination line spreads to the next layer. Each layer provides the next layer with calculated results, which is an internal representation of the data. This path goes from hidden layer to output layer.

But there is more. If the algorithm just calculates the weight in each neuron propagates the result to the output layer and stops there, it will not learn the weights that reduce the cost. If the algorithm only counts one iteration, there will be no real learning. This is where back propagation comes into play. Back propagation is a learning method that allows multiple sensor layers to adjust weights in the network to reduce processing cost, as shown in Figure 3. There is a complex requirement for back propagation to work properly. Functions that combine inputs and weights in neurons (such as numerical weights) and threshold functions (such as ReLU) must be different. These functions require bounded derivatives because gradient descent is the most common optimization used in multilayer perceptrons.

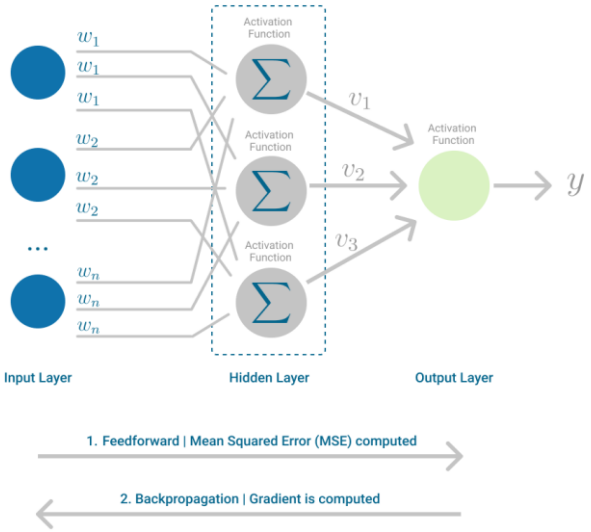

Figure 3: Multilayer Perceptron the Feed foward and Back propagation Steps.

At each iteration, the gradient of the mean square error is calculated for all input and output pairs after being sent to all weight layers. The weight of the first hidden layer is then adjusted according to the gradient value to reveal it again. This is how the weights are transmitted to the beginning of the neural network. One iteration of gradient descent is represented mathematically as:

$$\Delta_{\omega}(t) = -\varepsilon \frac{dE}{d\omega(t)} + \alpha \Delta_{\omega}(t-1)$$

$\Delta_{\omega}(t)$: Gradient Current Iteration
ε: Bias
E: Error
$\omega(t)$: Weight Vector
α: *Learning Rate*
$\Delta_{\omega}(t-1)$: Gradient Previous Iteration

This process continues until the gradient of each input-output pair converges; this means that the new calculated gradient does not change more than the specified convergence threshold compared to the previous iteration.

1.4 Random Forest Classifiers

Random Forest is a supervised machine learning algorithm widely used in classification and regression problems. It creates a decision tree of different models and performs majority voting on the distribution and mean during regression. One of the most important features of the random forest algorithm is that it can process data with continuous variables (such as regression) and categorical variables (such as distribution); Shows better for classification problems.

The algorithm is detailed below as shown in Figure 4.

- **Step 1:** In a random forest, n random data are taken from a data set containing k data.
- **Step 2:** Create a separate decision tree for each model.
- **Step 3:** Every tree decides to reproduce.
- **Step 4:** Final results are determined by the majority of votes or the distribution average and returned accordingly.

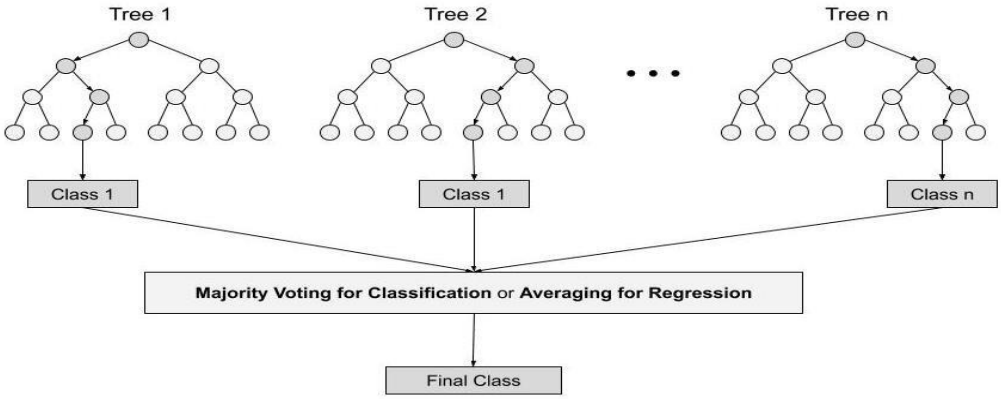

Figure 4: Random Forest Algorithm.

Hyper parameters are used in random forest to improve the performance of the model and estimate its strength or speed up the model. Applying hyper parameters increases predictive power:

- **n_estimators**: Number of trees generated by the algorithm before averaging the prediction.
- **max_features:** Random forest determines the maximum number of features for the distribution of nodes.
- **mini_sample_leaf:** Determine the minimum number of leaves to separate one from the inside.

Following hyper-parameters surges the speed of the random forest.

- **n_jobs:** This tells the engine how many processors are allowed. If the value is 1, only one processor is used, if the value is -1, there is no limit.
- **random_state:** Controls the randomness of the pattern. If the model has certain state values and takes the same hyperparameters and the same training data, the model will always produce the same results.
- **oob_score:** OOB means out of the bag. It is a random forest cross-validation method. In the third part of this model, data is used not for training but for evaluating performance. These models are called out-of-bag models

1.5 Hepatitis

Accurate identification is necessary for medical diagnosis, which is a significant and challenging undertaking. It's critical that the illness be identified in a timely manner and treated at the earliest. The liver is an essential component of the human body. Hepatitis, which results in liver inflammation, is one of the serious conditions that impair liver function. The presence of a virus in the liver is the primary cause of hepatitis [10]. Hepatitis has a high death rate and is a global illness. The body's essential processes could be impacted, and there could be cirrhosis, severe scarring, and an increased risk of liver cancer if prompt action is not done [11].

The condition can be cured by early detection achieved by appropriate diagnosis and treatment. The two key components for diagnosing any illness are: (i) The choosing the appropriate diagnostic settings and (ii) properly analyzing the findings with a skilled professional.

III. SYSTEM DESIGN

A machine learning-based model was created to correlate hepatitis with the categories. Numerous hepatitis features that are accessible online are taken into consideration as model inputs. Approximately 155 hepatitis samples were gathered online for the aforementioned repository. Following that, the hypothyroid samples were divided into two groups: DIE and LIVE. WEKA Explorer is used to process the classification in the end.

IV. DATA COLLECTION

Thus, the idea of creating hepatitis data was put out. As a result, both domestic and foreign resources were employed for the study. The internet has been used to gather data for the study from a variety of online sources. The necessary data set for this work was selected from the UCI repository while taking various clinical scenarios into account. There are 155 occurrences in this dataset with 20 properties, one of which is the same characteristics are used to determine a hepatitis patient's life expectancy. The details are as shown in following table 1 and The 19 attributes for classification are shown in table 2.

Table 1: Collection of hypothyroid Dataset

Name of Disease	Number of Features	Number of Samples
Hepatitis	20	155

Table 2: Attributes of hepatitis Dataset

Attributes	Value
Age	Numerical value
Sex	Male(1), female(2)
Steroid	no(1), yes(2)
Liver Big	no(1), yes(2)
Liver Firm	no(1), yes(2)
Spiders	no(1), yes(2)
Antivirals	no(1), yes(2)
Fatigue	no(1), yes(2)
Malaise	no(1), yes(2)
Spleen Palpable	no(1), yes(2)
Ascites	no(1), yes(2)
Varices	no(1), yes(2)
Varices	no(1), yes(2)
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alkaline Phosphate	33, 80, 120, 160, 200, 250
Aspartate transaminase	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Pro-time	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	no(1), yes(2)

V. PERFORMANCE ANALYSIS

The Data so collected need a processing. Hence as given in the system design phase, all the 155 data were processed into 2 categories. The test mode "Cross-Validation 5 FOLDS" and "Cross-Validation 10 FOLDS" used for ZeroR, Multilayer Perceptron and RandomForest. For processing WEKA APIs were used. The following tables shows the Confusion Matrix and True positive (TP) and False Positive (FP) rate of ZeroR, MultilayerPerceptron and RandomForest.

Table 3: Confusion Matrix for ZeroR for Test Mode: Cross-Validation 5 FOLDS

Classified as ⟶	DIE	LIVE
DIE	0	32
LIVE	0	123

Table 4: TP and FP Rate of ZeroR for Test Mode: Cross-Validation 5 FOLDS

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DIE	0.000	0.000	0.000	0.000	0.000	0.476
LIVE	1.000	1.000	0.794	1.000	0.885	0.476
Weighted Avg.	0.794	0.794	0.000	0.794	0.000	0.476

Table 5: Confusion Matrix for Multilayer Perceptron for Test Mode: Cross-Validation 5 FOLDS

Classified as ⟶	DIE	LIVE
DIE	13	19
LIVE	15	108

Table 6: TP and FP Rate of Multilayer Perceptron for Test Mode: Cross-Validation 5 FOLDS

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DIE	0.406	0.122	0.464	0.406	0.433	0.786
LIVE	0.878	0.594	0.850	0.878	0.864	0.786
Weighted Avg.	0.781	0.496	0.771	0.781	0.775	0.786

Table 7: Confusion Matrix for Random Forest for Test Mode: Cross-Validation 5 FOLDS

Classified as ⟶	DIE	LIVE
DIE	14	18
LIVE	6	117

Table 8: TP and FP Rate of Random Forest for Test Mode: Cross-Validation 5 FOLDS

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DIE	0.438	0.049	0.700	0.438	0.538	0.857
LIVE	0.951	0.563	0.867	0.951	0.907	0.857
Weighted Avg.	0.845	0.456	0.832	0.845	0.831	0.857

Table 9: Confusion Matrix for ZeroR for Test Mode: Cross-Validation 10 FOLDS

Classified as ⟶	DIE	LIVE
DIE	0	32
LIVE	0	123

Table 10: TP and FP Rate of ZeroR for Test Mode: Cross-Validation 10 FOLDS

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DIE	0.000	0.000	0.000	0.000	0.000	0.467
LIVE	1.000	1.000	0.794	1.000	0.885	0.467
Weighted Avg.→	0.794	0.794	0.000	0.794	0.000	0.467

Table 11: Confusion Matrix for Multilayer Perceptron for Test Mode: Cross-Validation 10 FOLDS

Classified as ⟶	DIE	LIVE
DIE	18	14
LIVE	17	106

Table 12: TP and FP Rate of Multilayer Perceptron for Test Mode: Cross-Validation 10 FOLDS

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DIE	0.563	0.138	0.514	0.563	0.537	0.823
LIVE	0.862	0.438	0.883	0.862	0.872	0.823
Weighted Avg.→	0.800	0.376	0.807	0.800	0.803	0.823

Table 13: Confusion Matrix for Random Forest for Test Mode: Cross-Validation 10 FOLDS

Classified as ⟶	DIE	LIVE
DIE	15	17
LIVE	6	117

Table 14: TP and FP Rate of Random Forest for Test Mode: Cross-Validation 10 FOLDS

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DIE	0.469	0.049	0.714	0.469	0.566	0.868
LIVE	0.951	0.531	0.873	0.951	0.911	0.868
Weighted Avg.→	0.852	0.432	0.840	0.852	0.839	0.868

VI. CONCLUSION

The following table 15 shows the summary of Classification.

Table 15: Summary of Classification

Classifier	ZeroR		Multilayer Perceptron		RandomForest	
Test Mode	Cross-Validation 5 FOLDS	Cross-Validation 10 FOLDS	Cross-Validation 5 FOLDS	Cross-Validation 10 FOLDS	Cross-Validation 5 FOLDS	Cross-Validation 10 FOLDS
Correctly Classified Instances	123 (79.35%)	123 (79.35%)	121 (78.06%)	124 (80%)	131 (84.51%)	132 (85.16%)
Incorrectly Classified Instances	32 (20.64%)	32 (20.64%)	34 (21.93%)	31 (20%)	24 (15.48%)	23 (14.83%)

In this paper as per the previous performance analysis, Table 15 Summary of Classification shows that the Classifier Random Forest has the accuracy for test mode evaluate on training data for 5 Folds cross-validation is 84.51% and for 10 FOLDS cross-validation is 85.16%, the Classifier Multilayer Perceptron has accuracy for both cross-validations evaluate is below 80% and the Classifier ZeroR has accuracy for test mode evaluate on training data is 79.35% for both cross-validations. The accuracy Classifier ZeroR for test mode evaluate on training data is worst as there is no predictability power in ZeroR. So it is concluded that Classifier Random Forest is the best classifier for classifying hepatitis data.

From all the above result in the Table 3 to Table 14, it is observed that performance of Classifier Random Forest is Excellent as compared to Classifier ZeroR and Multilayer Perceptron.

REFERENCES

[1] AHMED M. KHEDR, ZAHER AL AGHBARI, AMAL AL ALI, AND MARIAM ELJAMIL,2021, An Efficient Association Rule Mining From Distributed Medical Databases for Predicting Heart Diseases, IEEE Access, VOLUME 9, pp. 15320- 15333.

[2] Khedr, Z. A. L. Aghbari, and I. Kamel,2018, Privacy preserving decomposable mining association rules on distributed data, Int. J. Eng. Technol., vol. 7, nos. 313, pp. 157162.

[3] M. A. Cifci and S. Hussain, 2018, Data mining usage and applications in health services, Int. J. Informat. Vis., vol. 2, no. 4, p. 225, doi: 10.30630/joiv.2.4.148.

[4] R. Ray,2018, Advances in data mining: Healthcare applications, Int. Res. J. Eng. Technol., vol. 5, no. 3, pp. 23562395.

[5] P. Nayak,2017, A survey on medical data by using data mining techniques, Int. J. Advance Res., Ideas Innov. Technol., vol. 3, no. 6, pp. 13301335.

[6] Vijayarani, S., and S. Dhayanand, 2015, Liver disease prediction using SVM and Naïve Bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) 4, no. 4 (2015): 816- 820.

[7] Sartakhti, Javad Salimi, Mohammad Hossein Zangooei, and Kourosh Mozafari, 2012, Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVMSA), Computer methods and programs in biomedicine 108, no. 2 (2012): 570-579.

[8] Uttreshwar, Ghumbre Shashikant, and A. A. Ghatol,2009, Hepatitis B diagnosis using logical inference and generalized regression neural networks, In 2009 IEEE International Advance Computing Conference, pp. 1587-1595. IEEE.

[9] Ba-Alwi, Fadl Mutaher, and Houzifa M. Hintaya, 2013, Comparative study for analysis the prognostic in hepatitis data: data mining approach, spinal cord 11: 12.

[10] Ghumbre S. U.; Ghalot A.A,2008, Hepatitis B Diagnosis using Logical Inference And Self Organizing Map, Journal of Computer Science ISSN 1549-3636.

[11] M. A. Chinnaratha, G. P. Jeffrey, G. Macquillan, E. Rossi, B. W. D. Boer, D. J. Speers, and L. A. Adams,2014, Prediction of morbidity and mortality in patients with chronic hepatitis c by non-invasive liver fibrosis models, Liver International, vol. 34, no. 5, pp. 720–727.