Semiconductor Optoelectronics
ISSN: 1001-5868

COMPARISON OF MACHINE LEARNING ALGORITHMS IN BIO-MEDICAL IMAGE PROCESSING

Dr. M. Hemalatha
Professor and HOD, ECE Department, Sri Rama Engineering College, Tirupati, A.P, India.

Dr. B.Lalitha
Professor, ECE department, Sri Rama Engineering College, Tirupati, A.P, India.

V. Komala Devi
Associate Professor , ECE department, Sri Rama Engineering College, Tirupati, A.P, India.

Abstract
The machine learning Algorithms are implemented on liver patient's data. The medical data set is gathered from North-East of Andhra Pradesh, India. Human mortality and human morbidity are superior due to liver disease. Now a days, liver disease is huge in number due to prevalent intake of alcohol and also due to hepatitis. The main reason of liver disease is due to more consumption of drugs, harmful food, infections and toxic substances. The Liver damage is expected to play a crucial role in inflammation, scarring, obstructions, cirrhosis, liver failure, and even liver cancer. The use of herbal medicines can be traced back several thousand years ago in ancient China. According to evidences many natural products are available as chemo protective agents against common liver diseases, such as hepatitis, cirrhosis, liver cancer, fatty liver diseases, and gallstones. This disease treatment is very costly and complicated. By considering all this facts, the work is carried out in this significant area. A novel machine leaning model has been introduced to detect liver disease. The Random forest ensemble algorithm outperformed when compared to Support vector machine (SVM), Multi Layer Perception classifier (MLP classifier) and Linear Regression algorithms. The Classification of liver disease data is carried out by using confusion matrix.
Keywords: Machine learning, SVM, Logistic Regression, Random forest, confusion matrix.

1. Introduction

The classifications of liver disease are cirrhosis, hepatitis, and liver tumor [1]. Many advances in biotechnology and more distinctively high throughput sequencing result incessantly in an easy and economic data production, thereby ushering the science of applied biology into the area of big data [2, 3]. The death rate due to liver disease is 2 million per day [4]. The author presented about naive bayes (NB) and NB tree algorithms for detecting liver disease [5]. At early stages, it is difficult to identify liver tissues that have been spoiled and it requires experts to recognize the disease [6]. The big data plays very outstanding role in

machine learning. The big data is separated into small segments, for analyzing the data by multidisciplinary machine learning algorithms. The prediction of liver disease is challenging task for the doctors [7]. It is important to know the exact diagnosis of patients by evaluation and clinical assessment Medical field produces big data about report related to patient, clinical assessment, cure, follow-ups, and tablets [8]. Development in big data needs some proper means to extract and process data effectively and efficiently [9]. Health of public is primary thing for defending and curing from health hazard diseases [10].

2. Materials and methodology

The dataset is downloaded from famous kaggle site. The datasets are processed by Jupiter notebook 3.0. The Machine learning algorithms are carried out on the liver patient datasets. The four machine learning algorithms are linear regression, SVM, MLP Classifier, and Random forest [11]. Even NB uses tress like decision tress in Random forest algorithm [12]. The Random forest algorithm gave best performance compared to linear regression, MLP classifier and SVM. The steps in machine learning model include data collection, model fitting, hyper parameter tuning, data preparation and model evaluation.

2.1 Data collection

The dataset is obtained from kaggle which consists of liver patients from North-East of Andhra Pradesh, India. The set has total patients of 583. Out of which 441 are male patients and remaining are female patients. The Figure 1 depicts count plot for liver patient datasets.

2.2 Exploratory Data Analysis

Exploratory data analysis is performed on liver patient Dataset. The male patients are affected by liver disease compared to female patients. Total number of liver disease patients in dataset is 583. The data set 1 comprises of 416 liver patient records and 167 non-liver patient records. The dataset 2 contains 441 male patients and 142 female patient details. The Figure 1 is count plot which describes the gender of liver patients.

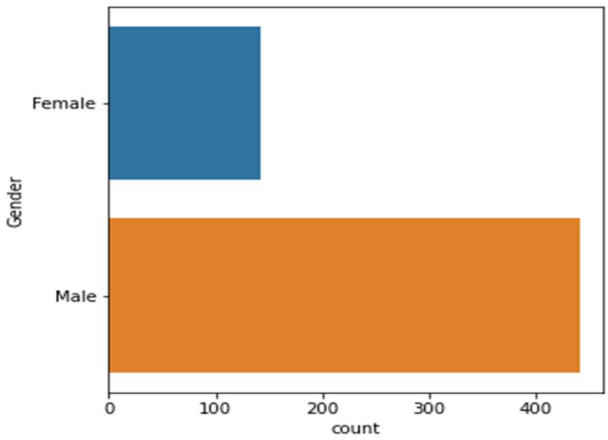

Figure 1. count plot for liver patients dataset

. 2.3 Distribution of numerical features

The dataset has eleven particular feature attributes such as ‘Age’, ‘Gender’, ‘Total Bilirubin’,'Direct_Bilirubin',‘Alkaline_Phosphotase’,‘Alamine_Aminotransferase’,‘Aspartat

e_Aminotrans','Total_Protiens', 'Albumin', 'Albumin_and_Globulin_Ratio', 'Dataset'. The Figure 2 explains clearly about various attributes of liver patient's data.

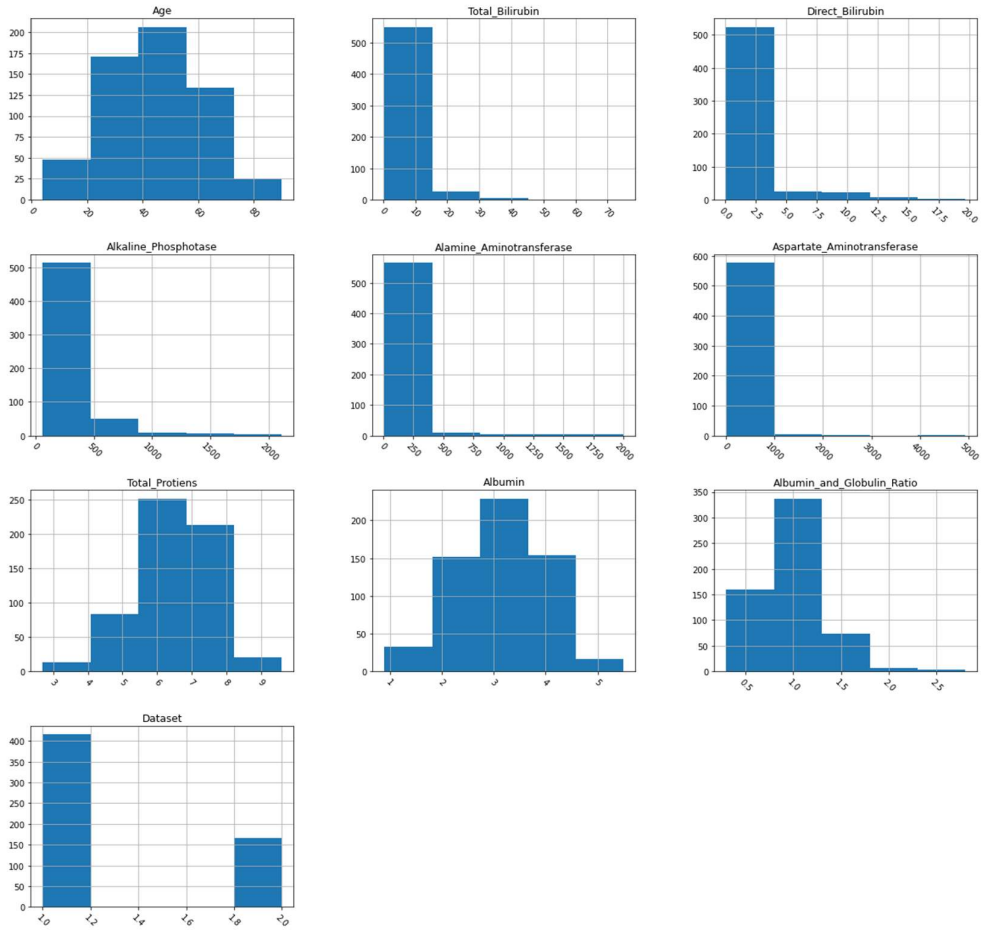

Figure 2. Plot for distribution of attributes

2.4 Distribution of categorical data

The machine learning algorithms are such as Linear regression, SVM, MLP Classifier, and Random forest carried out on Dataset1 and Dataset 2. The total male liver patients are 441 in the two datasets. The total female patients are 142 in the two datasets. In the dataset 2 more number of patients is suffering from liver disease. The male patients are more affected compared to female patients due to various reasons. The Figure 3 depicts about distribution of numerical categorical data in terms of male and female. The Figure 4 explains Direct_Bilirubin of liver patients. Here, its value exists in male patients and absent in female patients. The value decides the presence of liver disease is more in male patients.The Alkaline_phosphotase attribute is plotted in Figure 5 for both datasets.

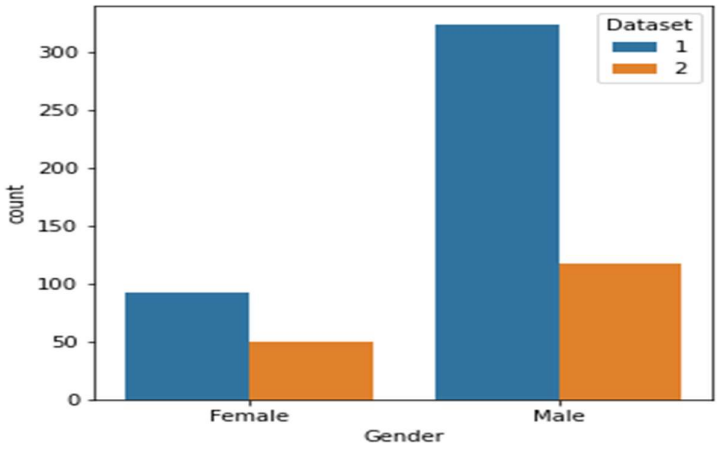

Figure 3. Categorical dataset distribution

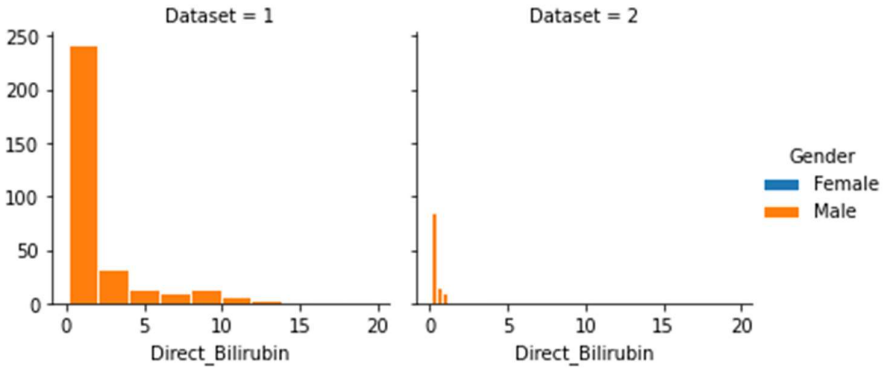

Figure 4. Plot for Direct_Bilirubin in two datasets

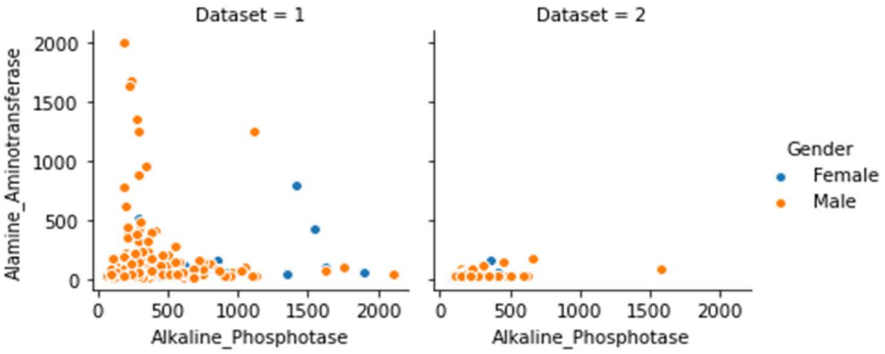

Figure 5. Plot for Alkaline_phosphotase in two datasets

2.5 Data cleaning

The duplicate data in the two datasets are filtered and the datasets are cleaned. After removal of duplicates, the data is ready for data preparation. The Data cleaning for the Alkaline_phosphotaseis depicted graphically in Figure 6. Similarly all the attributes in the data sets are cleaned and machine learning algorithms are carried out. After removing duplicates, the data analysis is implemented.

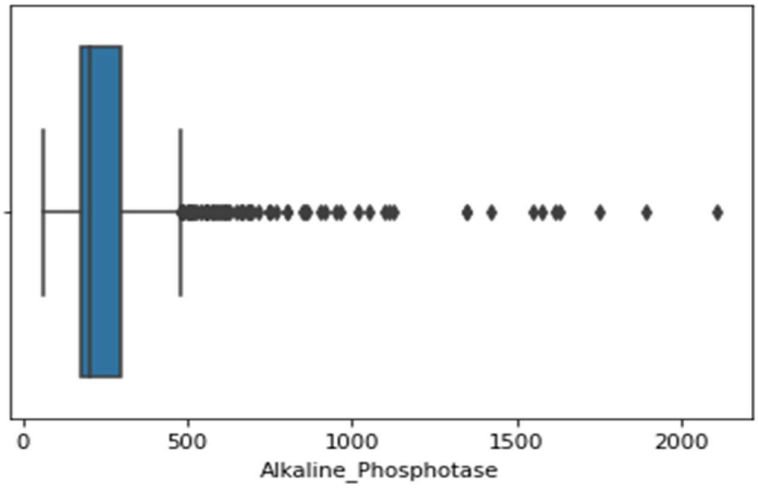

Figure 6. Plot for Alkaline_phosphotase data cleaning

3. Results and Discussions

The four machine learning algorithms i.e., Linear regression, SVM, MLP Classifier and Random forest algorithms are applied and analyzed for the two liver data sets. The proposed Random forest algorithm with machine learning preprocessing techniques outperformed and proved superior compared to SVM, MLP Classifier and Linear regression algorithms. In preprocessing it is very important to study and visualize the dataset. Sometimes the dataset has to undergo certain statistical algorithms where data can be studied and then machine learning model is carried out depending upon the requirement. Here Pearson correlation algorithm is applied to the liver datasets. If the correlation is +1 then it indicates that there is good correlation between the two classes. If the correlation is -1 then it indicates that there is negative correlation between two classes. The two classes have no correlation, if the correlation value is 0. Finally after observing and visualizing the data, the machine leaning algorithms are applied on liver patient datasets. The Random forest model is an ensemble technique which uses decision tree concept for classification of liver datasets [13]. Whenever decision trees are used for datasets then accuracy is improved much more compared to existing techniques. Figure 7 describes about confusion matrix for Logistic regression. Similarly we can calculate confusion matrix parameters for SVM, MLP classifier, and Random forest machine learning algorithms. The Figure 8 depicts comparison of Logistic regression SVM, MLP Classifier and Random forest algorithms. The Table 1 presents performance metrics for all four machine learning algorithms. The Random forest ensemble algorithm has superior performance in terms of confusion matrix metrics. Predominantly the Random forest ensemble model outperformed in terms of accuracy, kappa coefficient and F1score. The confusion matrix calculations are shown in equations 1, 2, 3, 4, and 5. The True positive Rate is also known as Recall. The Positive Predictive value is also called as Precision.

True positive Rate = TP/ (TP+FN) (1)

Positive Predictive value= FP/ (FP+FN) (2)

Accuracy= (TP+TN)/ (TP+TN+FP+FN) (3)

Kappa coefficient=

N* sum of correct pixels- sum of all row pixels* sum of all column pixels/(square of total

number of pixels-(sum of all row pixels* sum of all column pixels)) (4)

F1score= 2*(precision* Recall/ (precision+ Recall)) (5)

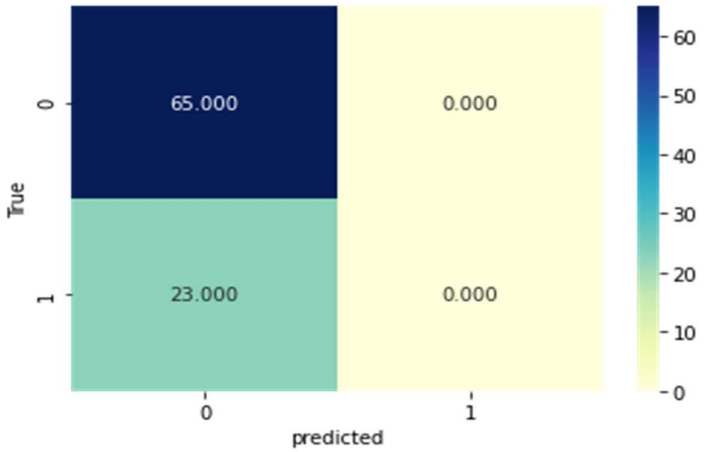

Figure 7. Confusion matrix for Logistic Regression

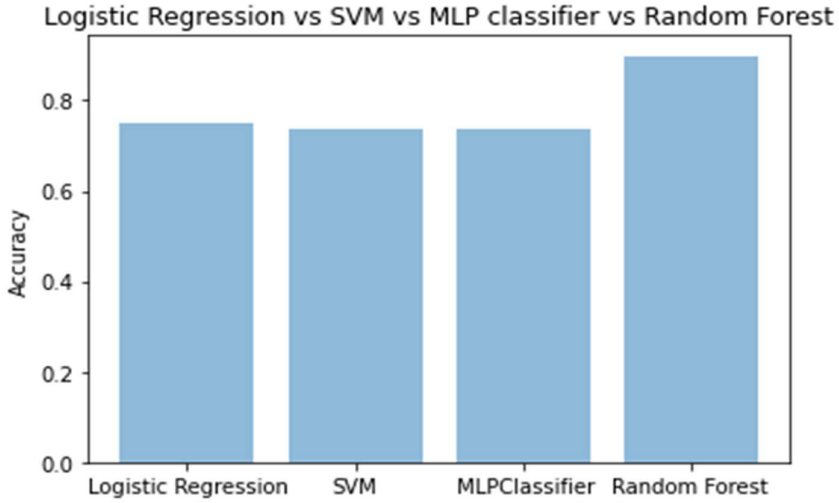

Figure 8. Comparison plot various machine learning algorithms

Table 1: Parameters for machine learning algorithms

Parameter	SVM	Logistic Linear Regression	MLP Classifier	Random Forest
Accuracy	73.2%	75.23%	73.86	89.86%
F1score	0.851	0.841	0.849	0.852

4. Conclusion

A novel machine leaning ensemble model has been proposed to detect liver disease. Total number of liver disease patients in dataset is 583. The Accuracy and F1score of dataset has

been calculated by the confusion matrix or error matrix. The proposed Random forest algorithm got good accuracy and F1score compared to logistic regression and SVM algorithms. The proposed algorithm has got 89.8% accuracy. The accuracies for logistic regression and SVM algorithms are 75% and 73.2% respectively.

Acknowledgments

We are very grateful to kaggle site for providing us liver patient data.

References

[1] World Health Organization, Global Action Plan on Physical Activity 2018-2030: More Active People for a Healthier World, World Health Organization, Geneva, Switzerland, 2019.

[2] Marx V. Biology: the big challenges of big data. Nature Jun 13 2013;498 (7453): 255–60. http://dx.doi.org/10.1038/498255a. Server/Data Center

[3] Wilson RA, Keil FC. The MIT encyclopaedia of the cognitive sciences. MIT Press; 1999.

[4] K. Sumeet, J.J. Larson, B. Yawn, T.M. Therneau, W.R. Kim. Underestimation of liver-related mortality in the United States. Gastroenterology; 145(2).pp.375–382.2013.

[5] A.A. Mokdad, A.D. Lopez, S. Shahraz, R. Lozano, A.H. Mokdad, J. Stanaway, et al. Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med 12. Article no.145. 2014.

[6] Sadiyah Noor, Novita Alfisahrin. Data Mining Algorithms for optimization of liver disease classifi-cation. Teddy Mantoro Electron-ic ISBN: 978-1-4799-2758-6 DOI: 10.1109/ACSAT.2013.81-IEEE .

[7] S. A. Gonzalez dan, E. B. Keeffe, Acute liver failure, dalam Handbook of Liver Disease Third Edition, Philadelphia: Elsevier. pp. 20-33, 2012.

[8] Roy, S., Singh, A., Shadev, S.K., Machine learning algorithm for classification of liver disorders. Far East J. Electron. Commun. 16(4), pp.789-800 ,2016.

[9] Siuly, S., Zhang, Y., Medical big data: neurological diseases diagnosis through medical data analysis. Data Sci. Eng. 1(2).pp.54–64 .2016.

[10] Luo, J., et al. Big data application in biomedical research and health care: a literature review. Biomed. Inform. Insights 8 :1-10,DOI: 10.4137/BII.S3159.2016.

[11] Support vector machine, Retrieve from: http://www.statsoft.com/textbook/support-vector-machines, Last Accessed: 5 Octobor,2019

[12] Naive Bayes, , Retrieve from: https://www.geeksforgeeks.org/naive-bayes-classifiers/, Last Accessed: 5 Octobor,2019

[13] Decision Trees, Retrieve from: https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/, Last Accessed: 5 Octobor,2019