Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 711-718

Detection and Comparative Analysis of Liver Disease Using Machine Learning Models

M. Hemalatha[1], B. Chandrababu Naik[2], K.C.Kullayappa[3]

[1]Department of ECE, Chadalawada Ramanamma Engineering College, Tirupati, A.P, India.

[2]Department of ECE, Aditya College of Engineering, Madanapalli, Tirupati, A.P, India.

[3]Department of ECE ,Chadalawada Ramanamma Engineering College, Tirupati, A.P, India.

ABSTRACT

In this paper, three different machine learning models are implemented on liver patient's data. The data set is gathered from North-East of Andhra Pradesh, India. Human mortality and human morbidity are increased due to liver disease. Now a day, liver disease is increasing because of widespread intake of alcohol and also due to hepatitis. The main reason of liver disease is due to intake of drugs, harmful food, infections and toxic substances. The Liver damage is expected to play a vital 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 Classification of liver disease data is done by using confusion matrix.

Key words- liver disease, machine learning, supervised learning, confusion matrix.

1. Introduction

The process in machine learning model includes data collection, model fitting, hyper parameter tuning, data preparation and model evaluation. The types of liver disease are cirrhosis, hepatitis, liver disease and liver tumour [1]. The death rate due to liver disease is 2 million per day [2]. The author presented about naive bayes and NB tree methods for detecting liver disease [3]. At early stages, it is difficult to identify liver tissues that have been damaged and it requires experts to identify the disease [4]. The big data plays very crucial role in machine learning. The big data is divided into small segments, for analyzing the data by multidisciplinary machine learning models. The detection of liver disease is challenging task for the doctors [5]. It is significant to understand the exact diagnosis of patients by evaluation and clinical examination. Medical field generates big data about report regarding patient, clinical assessment, cure, follow-ups, and medication [6]. Enhancement in big data needs some proper means to extract and process data effectively and efficiently [7].

2. Materials and Methodology

The dataset is downloaded from kaggle site. The software used for processing the data is Jupyter notebook 3.0. In this software three machine learning models are implemented. The three machine learning models are linear regression, SVM, and random forest. The random forest method gave best performance compared to linear regression and SVM (Support Vector machine).

2.1 Data collection

The set is gathered from North east of Andhra Pradesh, India. The data set has 11 particular features such as ‘Age’,TotalBilirubin’,'Direct_Bilirubin',‘Alkaline_Phosphotase’,‘Alamine_Aminotransferase’,‘Aspartate_Aminotrans’, 'Total_Protiens', 'Albumin', 'Albumin_and_Globulin_Ratio', Gender, 'Dataset'.

2.2 Exploratory Data Analysis

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

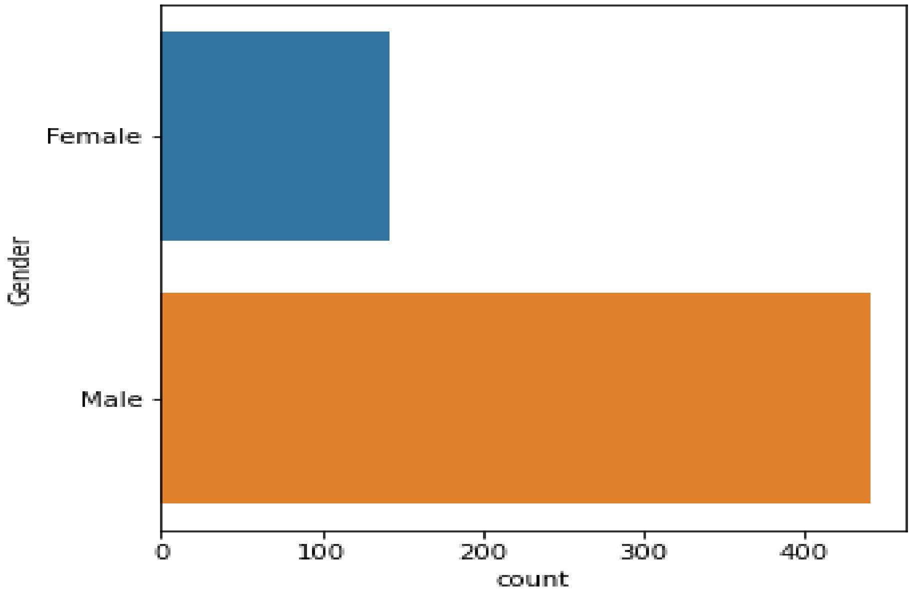

Fig.1: count plot of liver data

2.3 Distribution of numerical features

The dataset has 11 particular feature classes such as ‘Age’, ‘Gender’, ‘Total Bilirubin’, 'Direct_Bilirubin',‘Alkaline_Phosphotase’,‘Alamine_Aminotransferase’,‘Aspartate_Aminotrans’, 'Total_Protiens', 'Albumin', 'Albumin_and_Globulin_Ratio', 'Dataset'. The Figure 2 explains clearly about various numerical features of liver data.

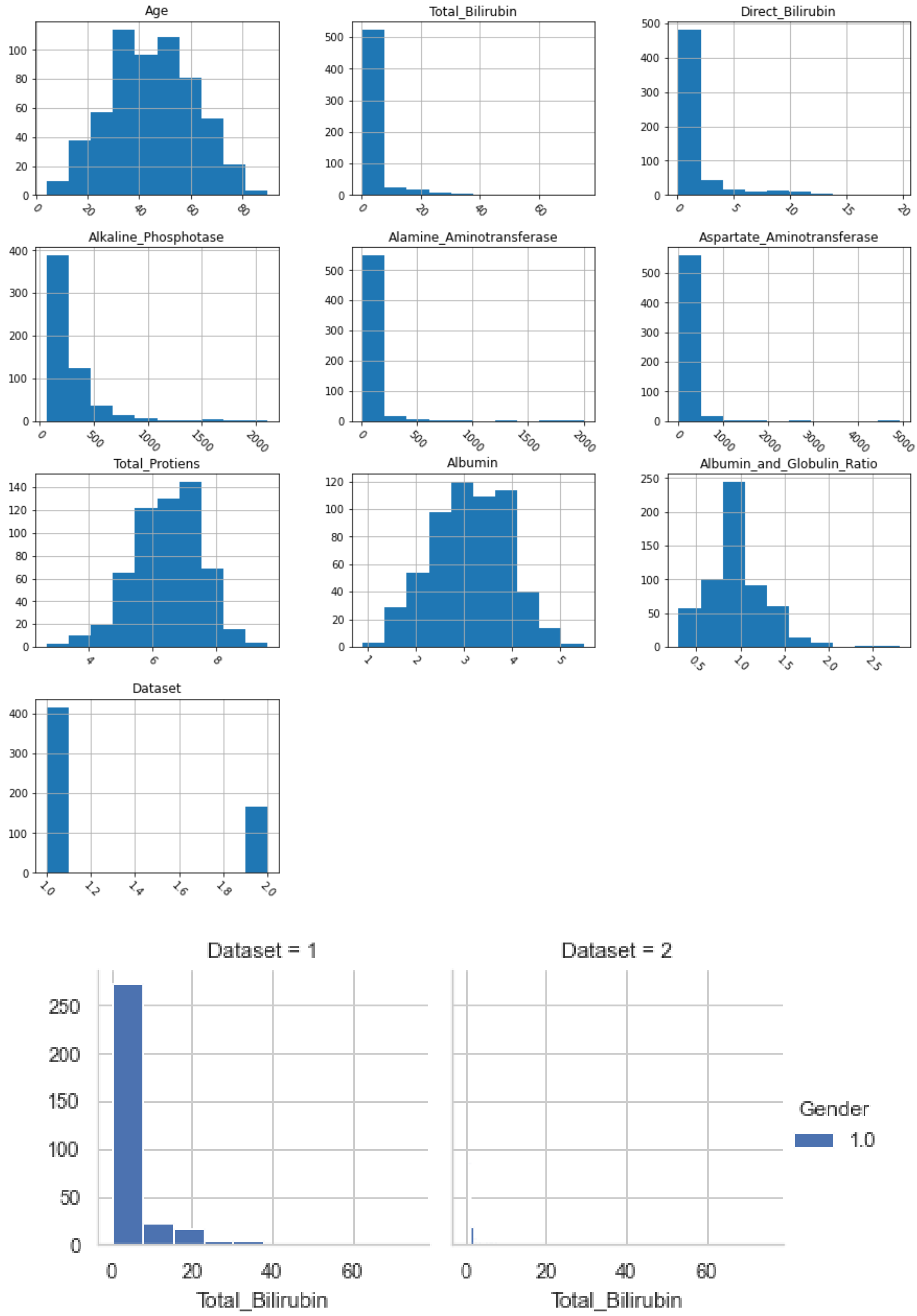

Fig. 2: Distribution of Numerical features

2.4 Distribution of categorical data

The machine learning models are implemented 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 got affected by 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.

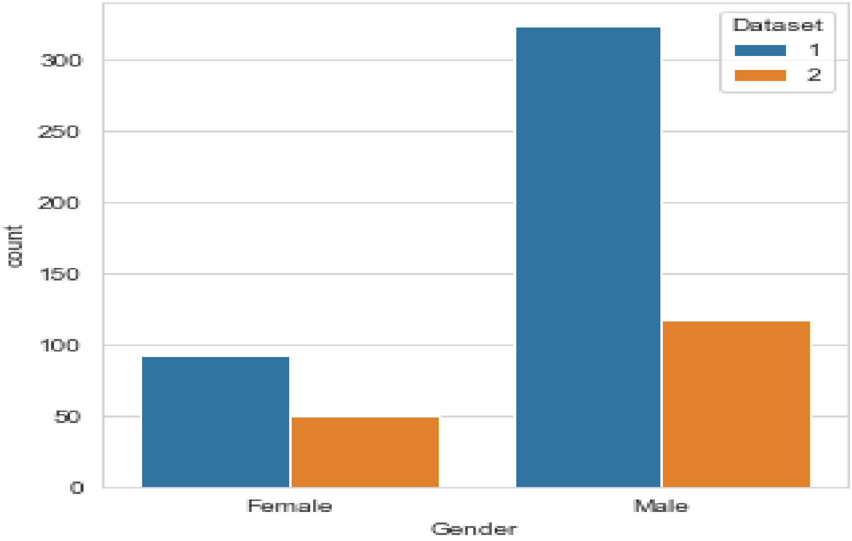

Fig. 3: Distribution of Categorical data

2.5 Data cleaning

The duplicates in the dataset are removed and data is cleaned. After removal of duplicates, the data is ready for data preparation. The Data cleaning for the Aspartate_Aminotransferase is shown graphically. Similarly all the classes in the data set are cleaned and machine learning models are implemented. After removing duplicates, the data analysis is easy and error free. The Figure 4 depicts data cleaning for Aspartate_Aminotransferase in liver data.

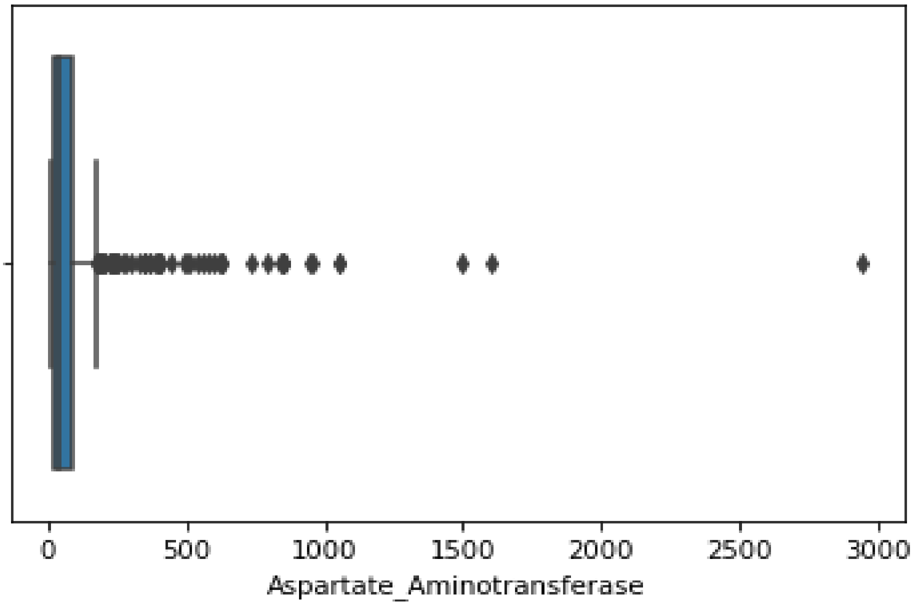

Fig. 4: Data cleaning for Aspartate_Aminotransferase

2.6 Data preparation

Linear regression, SVM, and random forest machine learning models are implemented on the two liver datasets. Then classification is carried out on the three machine learning models using confusion matrix. Finally liver disease is detected and performance metrics are calculated.

2.7 Algorithm

1. Acquisition of dataset from Kaggle.
2. Performing pre processing on the dataset. Exploratory data analysis is carried out on the dataset.
3. Extracting features from processed dataset.
4. Applying machine learning models on the dataset.
5. Performing classification for the machine learning models along with test data samples.
6. Prediction of the liver disease male and female patients in the dataset.
7. Calculating accuracy and F1score by using the confusion matrix.

3. RESULTS AND DISCUSSIONS

The comparative analysis is carried out for three machine learning models that is SVM, Logistic Regression, and Random Forest. The proposed method performed well in terms of accuracies. The Total _ Bilirubin for male and female data is shown in Figure 5. The female category has less Total _ Bilirubin in both data sets. The Figure 6 depicts about confusion matrix for linear regression. Similarly we can generate confusion matrixes for SVM and Random forest. The Figure 7 describes about accuracy’s for logistic regression, SVM, random forest machine learning models. Finally the Table 1 presents the performance metrics for three machine learning models. The proposed method is suitable for diagnosis of chronic lung cancer with 89.8% accuracy and 0.85 F1score. The proposed Random forest model outperform in terms of accuracy and F1score.

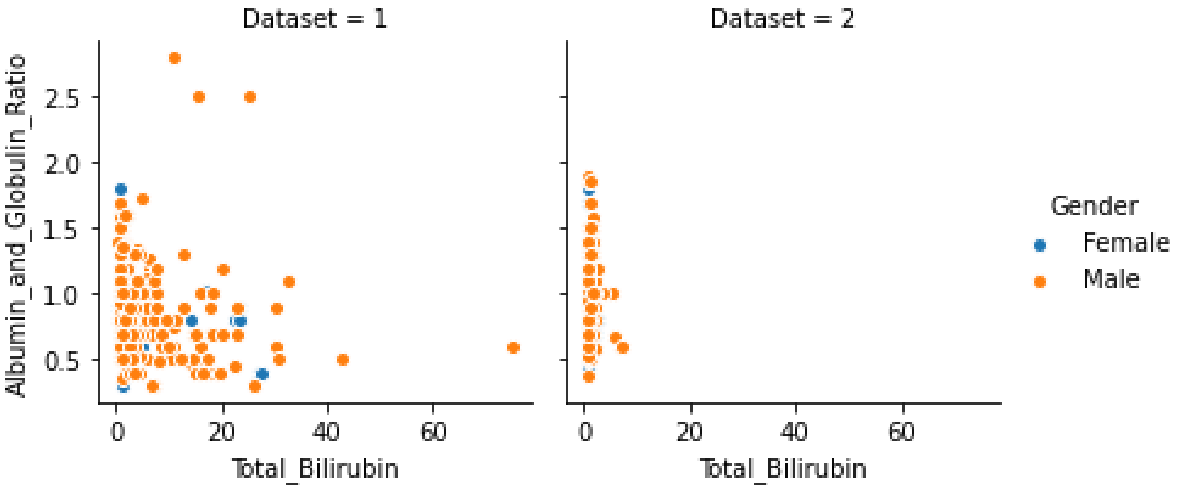

Fig.5: Total _Bilirubin for dataset 1 and dataset 2.

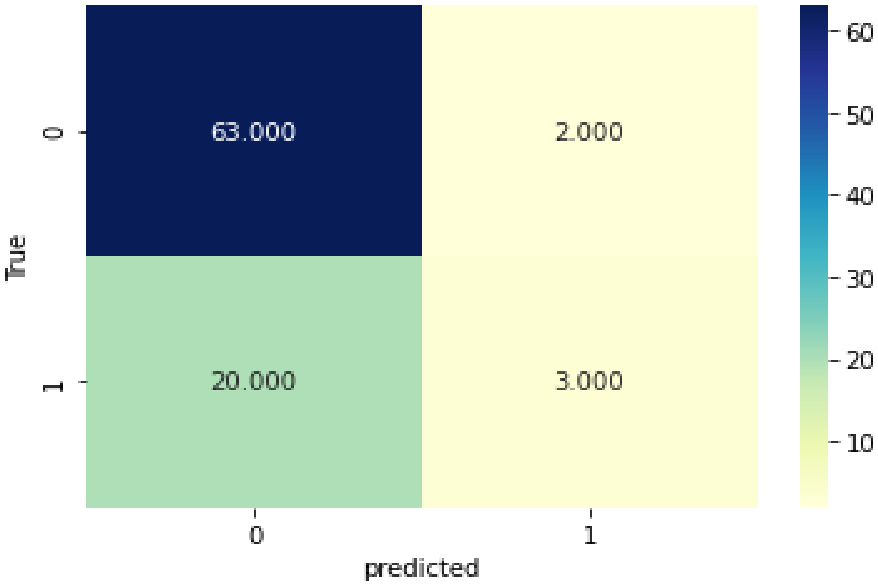

Fig.6: Confusion matrix for Logistic Regression

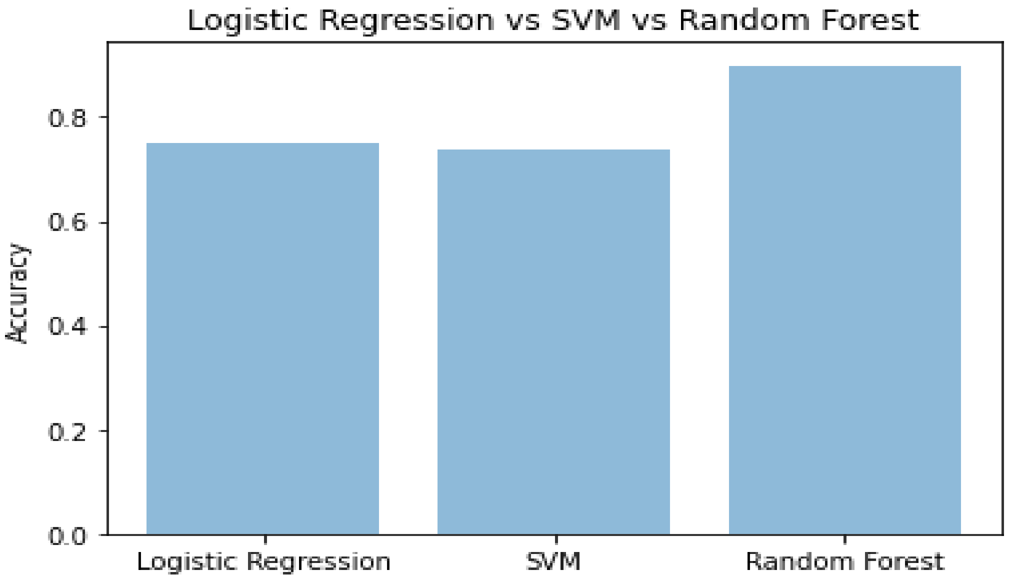

Fig.7: Comparative analysis of Logistic regression, SVM and Random forest

Table 1: Performance metrics for Logistic regression, SVM and Random forest

Parameter	SVM	Logistic Regression	Random Forest
Accuracy	73.2%	75%	89.8%
F1score	0.83	0.84	0.85

4. CONCLUSION

A novel machine leaning model has been introduced to detect liver disease. Total number of liver disease patients in dataset is 583. The Accuracy of dataset has been calculated by the confusion matrix. The proposed random forest method got good accuracy and F1score compared to logistic regression and SVM models. The proposed method has got 89.8% accuracy. The accuracies for logistic regression and SVM models are 75% and 73.2% respectively.

5.ACKNOWLEDGEMENT

I am very grateful to kaggle for providing liver disease data for my research purpose.

REFERENCES

1. 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.

2. 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.

3. Sadiyah Noor Novita Alfisahrin, Data Mining Models for optimization of liver disease classification. Teddy Mantoro Electron-ic ISBN: 978-1-4799-2758-6 DOI: 10.1109/ACSAT.2013.81-IEEE .

4. S. A. Gonzalez dan E. B. Keeffe, Acute liver failure, dalam Handbook of Liver Disease Third Edition, Philadelphia, Elsevier, pp. 20-33, 2012.
5. Roy, S., Singh, A., Shadev, S.K., Machine learning method for classification of liver disorders. Far East J. Electron. Commun. 16(4), pp.789-800 ,2016.
6. Siuly, S., Zhang, Y., Medical big data: neurological diseases diagnosis through medical data analysis. Data Sci. Eng. 1(2), pp.54–64 2016.
7. 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.