Liver Disease Prediction Using Machine Learning Classification

Jayakumar Sadhasivam
School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore, India.
E-mail: jayakumars@vit.ac.in

J. Senthil
Professor, Nandha Engineering College, Erode, India.

R.M. Ganesh
Associate Professor, Nandha Engineering College, Erode, India.

N. Chellapan
Assistant Professor, Nandha College of Pharmacy, Erode, India.

Received February 24, 2021; Accepted June 24, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI02/WEB18293

Abstract

People have disorder of liver that require medical care at correct time. It is utmost important to find the disease before it elapse the curable stage. Significantly, much of understanding of organ development has arisen from analyses of patients with liver deficiencies. Data mining is beneficial to find the disease at early stage based on the factors that can be gathered by performing test on the patient. Nowadays, around 65 % of the population in India are eating junk foods which minimize the metabolism rate and effect liver in many ways. In recent years, liver disorders have excessively increased and are still considered to be life threatening because it has caused low survivability. Still the patients having liver diseases are increasing and the symptoms of the diseases are difficult to identify. The doctors often failed to identify the symptoms which can cause severe damages to the patient and it requires utmost attention. So, we are applying Medical Data Mining (MDM) for predicting the liver disease by using the historical data and understanding their patterns. Here we are using prediction model i.e. Support Vector Machine (SVM) to achieve the goal.

Keywords

MDM, SVM, Classification.

Introduction

The habit of eating the junk food in many countries is enlarging day by day and due to this the health of people is getting degraded and now they are facing many types of liver disorders which can cause severe damage to their life. There are several types of liver disorders such as viral hepatitis, fatty liver disease, genetic liver disease, autoimmune liver disease, alcoholic liver disease etc. having different symptoms which cannot be ignored at any cost. These disorders can cause due to DNA damage by excessive alcohol consumption, by infection due to hepatitis B or hepatitis C virus, by obesity or other aspects like viral infection. These above mentioned common liver disorders can lead to cirrhosis which can even lead to either liver cancer or liver failure.

The liver disease hepatitis is causing liver infection which can lead to damage or harm to their body and happening of liver infection again and again can cause with the problem of cirrhosis in liver. The liver cirrhosis can be distinguished in the two stages, one is reliable and secure form, that is known as compensated cirrhosis that is accepted and the second one is critical and unhealthy, called decompensated cirrhosis. So, utmost attention should be given to the diagnosis of these disorders as soon as the patient arrives to medical practitioner.

There are several liver function tests (LFT) available for diagnosis of liver diseases. These tests are used to check the number of enzymes present in the blood. These LFTs is nothing but a group of blood tests that will help the doctor to get the information about the patient’s liver condition. These LFTs includes tests like bilirubin (both direct and total), albumin, transaminase and others. Out of these several tests available some are considered as presumptive and others are considered as confirmatory tests. What doctors do is that they will first ask the patient to get presumptive tests done and then the doctor will go through the reports. If doctor found everything normal then there is no liver disease to the patient. Otherwise doctor will ask the patient to go for confirmatory tests and those tests will decide that whether the patient is having liver disease or not.

The main concern here is that persons examining these liver disorders are facing more problems in estimating the records of patients from the enormous data available. Our main objective here is to predict the liver disease of a patient on the basis of dataset available. For this we are using Support Vector Machines (SVM). The purpose of organizing these types of test is to predict the liver diseases from the techniques that we are going to apply on the given dataset.

In SVM the hyperplane or the batch of the hyperplanes are created, as we know that SVM can be used for regression and classification. For a superior separation we have to work on hyperplane because it has the important feature as the division of the lines that take place will has the nearest training data point of the class. SVM algorithm is based on the kernel and the kernel performs the function that is used to convert the data that we have used as input into a high dimensional space where the problem is being resolved. Basically, it converts the low training data samples into higher dimensional data. The major reason to use the SVM is that the result that we are going to obtain through this will be more robust and the results that are obtained through SVM classifiers are accurate that is the reason we have given the SVM priority in our work.

Literature Survey

(Rajeswari, P., 2010) In this paper the author used the dataset from UCI repository consist of 345 instances with 7 different attributes. The liver disorders are find out by taking blood sample of the patient. Here in the dataset there are two categories of blood tests which are thought to be sensitive to liver disorders or not sensitive to liver disorders. The author used WEKA tool for classifying the data and the data is then evaluated using 10-fold cross validation. Here 70% of dataset is used for training data and remaining 30% for test data.

WEKA tool is used to compare the performance accuracy of different algorithms for liver disease dataset. The author then applied different algorithms like Naïve Bayes, FT Tree and Kstar and at last the results were compared.

First of all, classification is done using the algorithms. Two learning performance evaluators are present. The first will split the dataset into training and test data and the second one will perform the cross validation to find out the best algorithm. The algorithm is selected on the basis of its performance and the prediction of classification models are compared on the test data.

The author evaluated the performance of the algorithms on the basis of accuracy and time taken for the whole execution of the algorithm. In this paper FT Tree takes less time when evaluated on the liver dataset and it gives the more accuracy as compared to other algorithms.

(Saranya, A., 2017) The author reviewed the works carried by different authors for liver diseases, the methodologies used by them, advantage and disadvantages of their work.

The author analysed all the different works done in this field done by different authors so as to find the best classifier for manipulating the liver disorders.

The author first reviewed some of the related work done by different authors with the results of their work. The author explained C4.5 and Support Vector Machines (SVM) with their advantages and disadvantages. The performance of the algorithms was compared on the basis of speed, accuracy, performance and cost]. According to the study of the author C4.5 algorithm is proved to be better from the other classification algorithms.

(Priya, M.B., 2018) In this paper the author investigates the liver patients for building the classification models for predicting the liver diseases. The author implemented a model for predicting the liver diseases in three phases.

In first phase, the author applied min max normalization algorithm on the original dataset collected from the UCI repository. Normalization is done to eliminate redundancy and increase the reliability.

In second phase, using the PSO feature selection, subset of dataset from phase one is obtained which comprises of significant attributes only.

In the third phase the classification algorithms are applied on the dataset and after that accuracy of the applied algorithms is calculated using root mean square value and root mean error value.

The author applied different classification algorithms like Random Forest, Support Vector machine (SVM), J-48, Multilayer Perceptron and Bayesian Networks. The results show that Bayesian Networks and J-48 classification algorithms but J-48 proves to be the best as it gives 95.04% accuracy.

(Ramana, B.V., 2010) In this paper Naïve Bayes classification is used for diagnosis of liver disorders. The blood tests are analysed for this purpose. The dataset used here consists of 751 patient records from Andhra Pradesh state of India. This dataset contains 12 attributes and these are classified in two classes i.e. Liver Disorder and No Liver Disorder. The author used WEKA tool for this work. The whole dataset is divided into 10 parts and out of them 9 parts are used for the training set purpose and the remaining 1 part for testing set purpose. This 10 fold will reduce the sample bias and it will consider each part is used for training and testing.

The author then applied feature selection on male and female data present in the dataset and it is done by considering their ages as well. In both the cases the author has given the important features according to the age range. The accuracy of Naïve Bayes classifier on this dataset is 88%.

(Barnaghi, P.M., 2012) In this paper the author processes the blood test dataset and apply different classification algorithms available to identify the existence of liver disorder. Here the author used four different classification methods which includes Bayesian algorithms, decision tree**,** Neural Network classification and rough sets. To evaluate all the mentioned methods WEKA tool is used. Dataset used here is consist of 340 instances with 7 attributes which are divided into two different classes. The author used 66% of dataset as training data and rest 34% as test data.

Out of these methods the author applied several algorithms available like J48, LMT (Logistic Model Tree), Bayes Net (Bayesian Network), Naïve Bayes, MLP (Multilayer perceptron), RBF (Radial Basis Function Networks) and Rough Set.

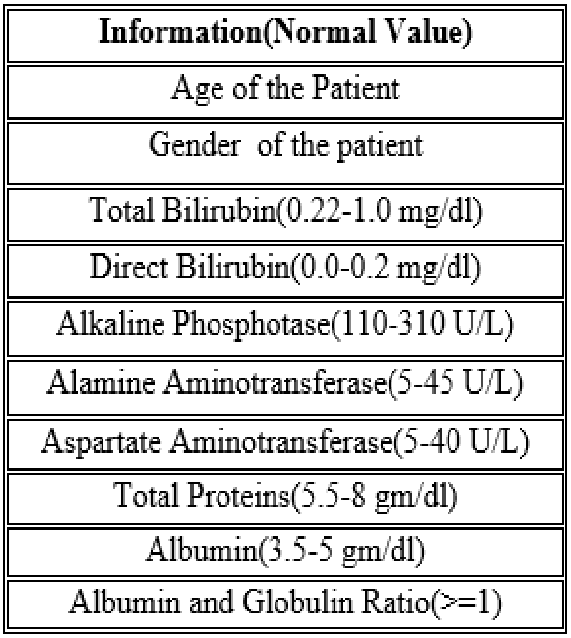
After applying several algorithms, they compared the accuracy of each algorithm with varying training size i.e. they used the combinations of training size to find out the highest accuracy of the applied algorithms and J48, MLP and RBF came out with highest accuracy i.e. 79.41%. Then they applied these algorithms by varying attributes i.e. first they take only 2 attributes then 3 then 4 like this. With different attributes MLP turned out to be the best algorithm with accuracy of 91.17% with 6 attributes.

Proposed Work

In the base paper the classification technique was used. The Classification technique which was used in the base paper named as NBTree, Decision Tree and Naïve Bayes. So, in the proposed work the support vector machine (SVM) classification technique was used and compare the result with the accuracy of the models given in the base paper. Apart from that in proposed work weka tool 3.9 was used to build the model and on the other hand the model was build using python jupyter notebook and compare to check which show the better results either python or weka tool.

So, in the proposed work the liver disease prediction model was build. There are many factors which causes the liver disease. Some of them which influence to detect the liver disease are Total Bilirubin (Total amount of bilirubin when old red blood cell breaks down inside the human body), Direct Bilirubin (It is a substance is made when the body breakdowns the old red blood cells. It is also part of bile, which your liver makes to help digest the food we eat), Alkaline Phosphotase (It is part of protein which release the enzymes to act as a catalyst which help the bile juice which produce by the liver), Alamine Aminotransferase (It is found in the plasma and various body part mostly in the liver), Aspartate Aminotransferase (It is a part of metabolism which help to digest the food we eat and keep the liver healthy), Total Proteins (Total Protein value present in the body), Albumin (Albumin is a kind of protein which is found inside human body and a major part for participation in total protein value), Albumin and Globulin Ratio (It is the ratio of both the proteins inside human body i.e. the ratio of albumin and Globulin). In, the given table 1 mention all the normal ranges of the attributes so it helps to remove the outliers from the data set using clustering technique (Vijayakumar, J., 2013) (Manchula, A., 2014).

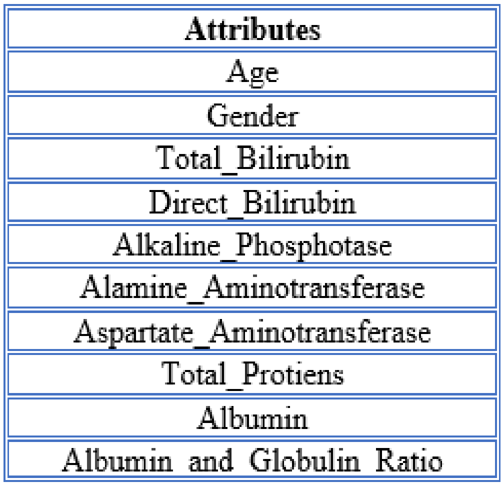
Table 1



Information(Normal Value)
Age of the Patient
Gender of the patient
Total Bilirubin(0.22-1.0 mg/dl)
Direct Bilirubin(0.0-0.2 mg/dl)
Alkaline Phosphotase(110-310 U/L)
Alamine Aminotransferase(5-45 U/L)
Aspartate Aminotransferase(5-40 U/L)
Total Proteins(5.5-8 gm/dl)
Albumin(3.5-5 gm/dl)
Albumin and Globulin Ratio(>=1)

In order to achieve the maximum accuracy of the model first thing is to be done is selection of attributes. Selection of attribute in context which is useful for the model accuracy and remove the other useless attribute. So, for the best attribute selection the author consults with the doctor which test should be consider for the liver disease and after consulting with the doctor and use the co-relation technique among the attributes. The data set having following attributes before selection of attributes.

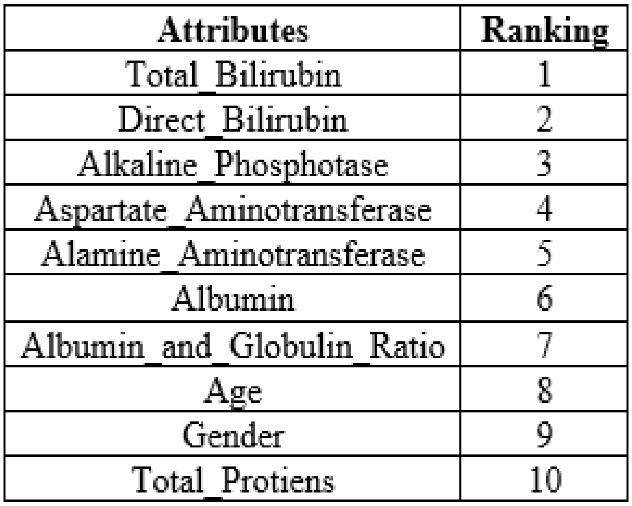
Table 2



Attributes
Age
Gender
Total_Bilirubin
Direct_Bilirubin
Alkaline_Phosphotase
Alamine_Aminotransferase
Aspartate_Aminotransferase
Total_Protiens
Albumin
Albumin and Globulin Ratio

In table 2 there are some attribute in which some of them are liver tests which should be done for knowing liver disease. After done consulted with doctor and using ranking method using chi-square is used to determine that which attribute is influence in the detection of liver disease. After the ranking process the author will get the following table 3 and pick the first six attributes to predict the model.

Table 3

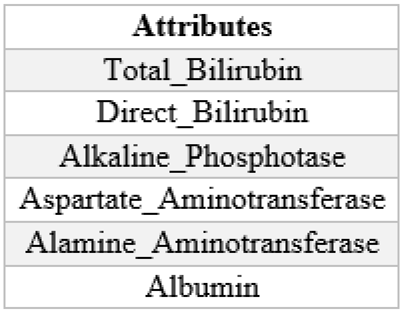


Attributes	Ranking
Total_Bilirubin	1
Direct_Bilirubin	2
Alkaline_Phosphotase	3
Aspartate_Aminotransferase	4
Alamine_Aminotransferase	5
Albumin	6
Albumin_and_Globulin_Ratio	7
Age	8
Gender	9
Total_Protiens	10

After the selection of the best attributes that influence the liver disease detection are as in the given table 4.

So, after the selection of attributes we apply the classification technique (SVM) and build the prediction model.

Table 4



Attributes
Total_Bilirubin
Direct_Bilirubin
Alkaline_Phosphotase
Aspartate_Aminotransferase
Alamine_Aminotransferase
Albumin

Experiment and Results

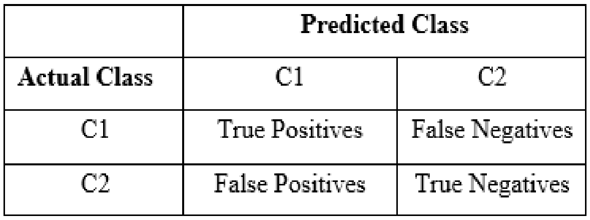
So, let's build the model but have some patience before analyzing the data set the pre-processing of data set should be done like remove the outliers using clustering, normalize the data and the advantage of using pre-processing to remove redundancy of the data, so initially the data set is as much as 200 tuples reduced to 157 tuples. For analyze the performance of the algorithm support vector machine (SVM) is used in classifying liver disease, Python using jupyter notebook and Weka tool software was used followed by 10 cross validation and compare both results and analyze which shows more accuracy of the model.

The accuracy of the model is find out by the help of the confusion matrix. The formula for measuring the accuracy of the model using confusion matrix as follows:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN)$$

Abbreviation named are TP (True Positive) is the number of positive cases that is correctly classified as positive, TN (True Negative) is the number of negative cases that is correctly classified as negative, FP (False Positive) is the number of negative cases that is incorrectly classified as positive cases, FN (False Negative) is the number of positive cases that is incorrectly classified as negative cases. Mention below table 5 present the confusion matrix.

Table 5

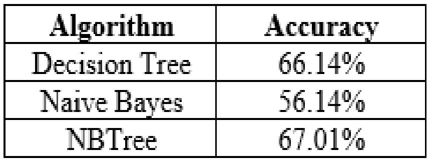


	Predicted Class	
Actual Class	C1	C2
C1	True Positives	False Negatives
C2	False Positives	True Negatives

A. Model Accuracy in the Base Paper

Below mention table 6 show the accuracy of the models in base paper.

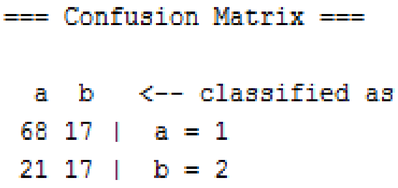
Table 6



Algorithm	Accuracy
Decision Tree	66.14%
Naive Bayes	56.14%
NBTree	67.01%

B. Model and Accuracy of Support Vector Machine Algorithm Using Weka Tool

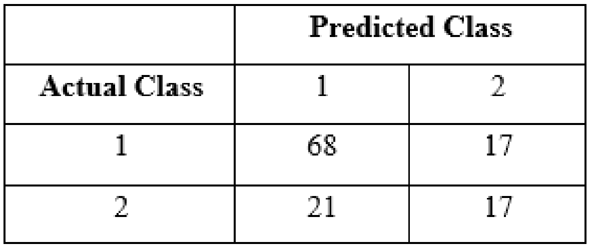
In SVM the hyperplane was created using regression technique between the classes with the widest range which separates the classes from each other. There are different kernels which help to create hyperplane as per the data complexity to separates the classes. The confusion matrix for the algorithm SVM mention using WEKA tool below in table 7.



```
=== Confusion Matrix ===

  a  b   <-- classified as
 68 17 |  a = 1
 21 17 |  b = 2
```

Table 7

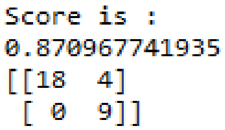


	Predicted Class	
Actual Class	1	2
1	68	17
2	21	17

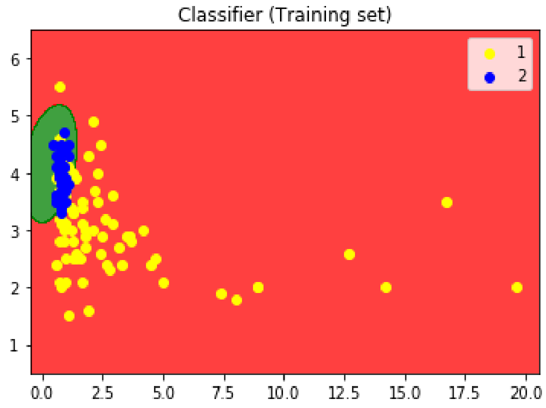
Based on the confusion matrix the accuracy value obtained by SVM using weka tool is 69.10%.

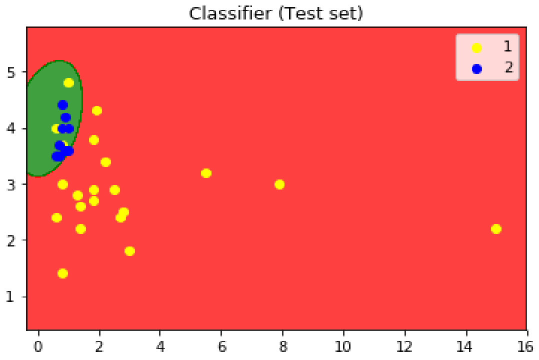
C. Model and Accuracy of Support Vector Machine (SVM) Algorithm Using Python Language

Confusion Matrix of algorithm SVM using python language as follow.



```
Score is :
0.870967741935
[[18  4]
 [ 0  9]]
```



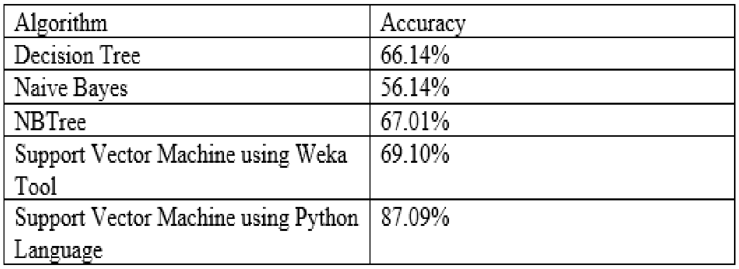


So, from the above mention confusion matrix the accuracy for algorithm SVM using python language is 87.09%.

D. Evaluation of the Algorithms

Comparison of the accuracy of the algorithm i.e, Decision tree, Naive Bayes and NBTree which all given in the base paper. The comparison of these given algorithm accuracy with the proposed algorithm support vector Machine (SVM) using Weka tool and comparison with the proposed algorithm support vector Machine (SVM) using Python language.

Table 8



Algorithm	Accuracy
Decision Tree	66.14%
Naive Bayes	56.14%
NBTree	67.01%
Support Vector Machine using Weka Tool	69.10%
Support Vector Machine using Python Language	87.09%

So, from the above mention table 8 the highest accuracy is obtained by support vector machine algorithm using python language is 87.09%

Conclusion

In this proposed paper, one data mining technique is used but with two different ways, one of them is support vector machine algorithms evaluated by using Weka tool and other support vector machine algorithms evaluated by using python language.

So, the python language is more powerful than the weka tool due it is good in pre-processing, attribute selection etc.

Due to above reason the best accuracy obtained by the support vector machine algorithm evaluated using python language was 87.09%.

Further, in this proposed work more efficient technique is used to achieve most significant attributes from the data set in identifying liver disease. By opting the most significant attributes the accuracy of the model will increased significantly in identifying the liver disease.

References

Rajeswari, P., & Reena, G.S. (2010). Analysis of liver disorder using data mining algorithm. *Global journal of computer science and technology*, *10*(14), 48-52.

Saranya, A., & Seenuvasan, G. (2017). A Comparative Study of Diagnosing Liver Disorder Disease Using Classification Algorithm. *International Journal of Computer Science and Mobile Computing*, *6*(8), 49-54.

Priya, M.B., Juliet, P.L., & Tamilselvi, P.R. (2018). Performance analysis of liver disease prediction using machine learning algorithms. *International Research Journal of Engineering and Technology (IRJET)*, *5*(1), 206-211.

Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *4*(4), 816-820.

Babu, M.P., Ramjee, M., Katta, S., & Swapna, K. (2016). Implementation of partitional clustering on ILPD dataset to predict liver disorders. *In 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 1094-1097.

Afzal, S., Masroor, I., & Beg, M. (2013). Evaluation of chronic liver disease: does ultrasound scoring criteria help?. *International journal of chronic diseases*, *2013*. http://dx.doi.org/10.1155/2013/326231

Sontakke, S., Lohokare, J., & Dani, R. (2017). Diagnosis of liver diseases using machine learning. *In International Conference on Emerging Trends & Innovation in ICT (ICEI)*, 129-133.

Ramana, B.V., Babu, M.S.P., & Venkateswarlu, N.B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, *3*(2), 101-114.

Ramana, B.V. (2010). Prof. MS Prasad Babu and BR Sarath Kumar: "New Automatic Diagnosis of Liver Status Using Bayesian Classification". *In IEEE International Conference on Intelligent Network and Computing (ICINC 2010),* 26-29.

Alfisahrin, S.N.N., & Mantoro, T. (2013). Data mining techniques for optimization of liver disease classification. *In International Conference on Advanced Computer Science Applications and Technologies*, 379-384.

Barnaghi, P.M., Sahzabi, V.A., & Bakar, A.A. (2012). A comparative study for various methods of classification. *In International Conference on Information and Computer Networks*, *27*(2), 875-81.

Vijayakumar, J., & Arumugam, S. (2013). Certain investigations on foot rot disease for betelvine plants using digital imaging technique. *In International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA)*, 1-4.

Manchula, A., & Arumugam, S. (2014). Face and fingerprint biometric fusion: Multimodal feature template matching algorithm. *International Journal of Applied Engineering Research*, *9*(22), 17295-17315.

Farzin, A., Yousefi, S., Amieheidari, S., & Noruzi, A. (2020). Effect of green marketing instruments and behavior processes of consumers on purchase and use of e-books. *Webology*, *17*(1), 202-215.