Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 2 (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Automated Liver Disease Diagnosis using Machine Learning Techniques

Sanjay Kumar Sen [1], Sanjay Kumar*[2], Nirmal Keshari Swain[3], Shankar Prasad Mitra[4]

[1,2]Department of Computer Science & Engineering, Brainware University Kolkatta
Email ID: sanjaysen2k@gmail.com
[3]Department of Information Technology, Vardhaman College of Engineering Hyderabad
Email ID: swain.nirmal6@gmail.com
[4]Department of Computer Science & Engineering, Brainware University Kolkatta
Email ID: spmitra2016@gmail.com
***Corresponding Author**:
Sanjay Kumar
Email ID: sanjaysatyam786@gmail.com

Cite this paper as: Sanjay Kumar Sen , Sanjay Kumar, Nirmal Keshari Swain, Shankar Prasad Mitra, (2025) Automated Liver Disease Diagnosis using Machine Learning Techniques. *Journal of Neonatal Surgery*, 14 (2), 266-273.

ABSTRACT
Liver diseases are among the most deadly and panic conditions of health sector in entire world, because it is predicted that they will worsen due to a number of factors, including an increase in alcohol consumption, deteriorating global pollution from heavy industrialization and global warming, toxic gas exhaustion, contaminated water, food, and drugs, and most importantly, poor lifestyle choices. These factors all contribute to an ongoing rise in the diagnosis of liver anomalies in patients. Sometimes it is very difficult to identify the cause and symptoms by doctors. We are applying ML algorithms for the prediction of liver diseases. In order to develop forecasting system for the early identification of liver disorder, the patients liver datasets are examined to develop different forecasting system for the pre-symptomatic of liver cancer. Machine learning technique have the potential to significantly improve the prediction of accuracy. It has been observed that machine learning is improving our basic understanding of illness progression. Application of machine learning algorithm plays a crucial role in analysing the liver datasets. The major goal of the investigation is to apply various ML algorithm for the prediction of liver diseases by comparative analysis of various machine learning models indicates the best method. The prime purpose of this work is for analysing the result of LR, NB and Random forest computation in Liver disorder dataset. And Datasets are collected from the UCI database of Indian liver patient records. The suggested technique is executed using Python and outcomes are analysed ont the basis of accuracy, recall and precision. It is found that 0.75 of accuracy in Linear Regression, 0.74 of accuracy in Random Forest, 0.69 in Decision Tree, 0.64 of accuracy in Supporting voting machine, 0.62 of accuracy in knn and 0.53 of accuracy in Naïve Bayer. From above analysis Linear Regression having highest accuracy in compare to others. Our main focus on prediction of liver disease using clinical data.

Keywords: Liver, Diagnosis, ML algorithm, Logistic Regression, NB, Random forest.

1. INTRODUCTION:

Every year millions of people die from liver disease in this world. Mortality rate is more in different Asian countries [1]. As per WHO report for the year 2018, about 2480 people die of liver disease in Malaysia which is 1.7% of total death of world in the same year. This organ is one of the largest organs of the body. It plays a vital role for glucose metabolization and cholesterol metabolization. It breaks insulin and bilirubin with glucuronidation [2]. It also helps in nutrient storage, releasing of differenttoxins which are harmful for the body. The removal of toxins from the body is also another vital functions of the body that makes possible for survival. If there is a infection of virus by any means it causes damaged to the liver. When liver Is affected by any means, its function seizes and sickness arises [3]. Due to the assemblage of lipids in the liver causes non-alcoholic fatty liver disease. There is another disease called “non-alcoholic steatohepatitis” characterised by swelling and cutting of the liver cells [4]. Cirrhosis is also another fatal liver disease symptomized by the replacement of healthy tissue by scar tissue. This disease iscaused mainly due to alcoholism, chronic hepatitis B and hepatitis C[5]. As per recent research there is hardly any investigation due to any irregularities in the functioning test of liver which is recommended by national standard specification [6]. Hepatitis are in two forms acute hepatitis [7] due to the inflammation of liver rapidly and

chronic hepatis [8] due to inflammation of liver and liver damage gradually. Mostly hepatitis caused due to the infection of group of virus, different types like A, B, C, D and E [9] named chronically in order of their discovery.This type of Hepatitis of category A disease is due to the attack of category A virus. This disease is developed not only due to contamination of food and water but also through sexually contact and blood transfusion [10]. Hepatitis B also same characteristic of Hepatis A having contact with any infected person due to transmission of bodily fluid. It exists in two types, acute which cannot be treated and another is chronic that causes liver cancer [11].

Hepatitis C is also same of above two, also due to the transfusion of blood of infected person. Like hepatitis

B, it also exists in acute and chronic form [12]. The hepatitis D affects people infected with hepatitis B[13]. However, people affected with hepatitis D virus are very rare. The hepatitis E is affected by contaminated water of an infected person.

2. RELATED WORK

Many models are developed now a days that can be helpful for the diagnosis of liver disease by physicians in the medical field to support diagnosis system.[15]. Considering 6 benchmarks proposed by Christopher N.[16], that includes liver related diseases, heart disorder, diabetes, hepatitis disease. Basing on two systems developed by the authors on WSO and C4.5 for which an 64.60% was obtained and Raman[17] made some research study on liver diseases by diagnosis on some prime algorithms like NB models, K-NN and SVM and also obtained an accuracy of 51.59% on NB models, by C4.5 algorithm an accuracy of 55.94%, by BPNN accuracy of 66.66%, 62.66% of accuracy on KNN and an accuracy of 62.66% on SVM. The poor results of the training and testing of the dataset is due to insufficient in the dataset. Therefore in order to compensate this poor performance Sug[18], a method is suggested which is basing on oversampling in minor classes. C4.5 and CART[8] are two algorithms of decision tree are considered by authors to conduct this experiment. Though above two algorithm are adequate not sufficient. So more research works are required to achieve better accuracy. The misdiagnosis of the liver disorder must be prevented to make diagnosis more effective and efficient to achieve better accuracy which will help best treatment for the patient.

3. MATERIALS AND METHODOLOGY

By machine learning algorithm different types of diseases are discovered which makes correct decision. Different patterns of data discovered by data mining process are stored by computer. One of the most crucial processes in data mining is the data processing. The steps followed for the gathering observation from huge information are cleaning, aggregation, isomerisation and reduction of data[15]. For the process of classification and regression this study will be used. required. For the predicting responses classification is used having some few known values. But one or more continuous variables can be predicted through Regression [16]. By using Naïve Bayes algorithms in the classification process models are evaluated. The experiments are carried by using Python programming tool in Jupiter notebook with into a number of folds.

3.1 Proposed Model

In this system, the dataset is pre-processed and the anomalies as well as blank cells are removed from it for the purpose of effective prediction of liver dataset. For achieving enhanced clarity for the nos of correct / incorrect predictions, a Confusion matrix is constructed. The accuracy is checked after applying several classification and ML techniques. The prime objective of the process is for better accuracy. The suggested model is shown in figure .1

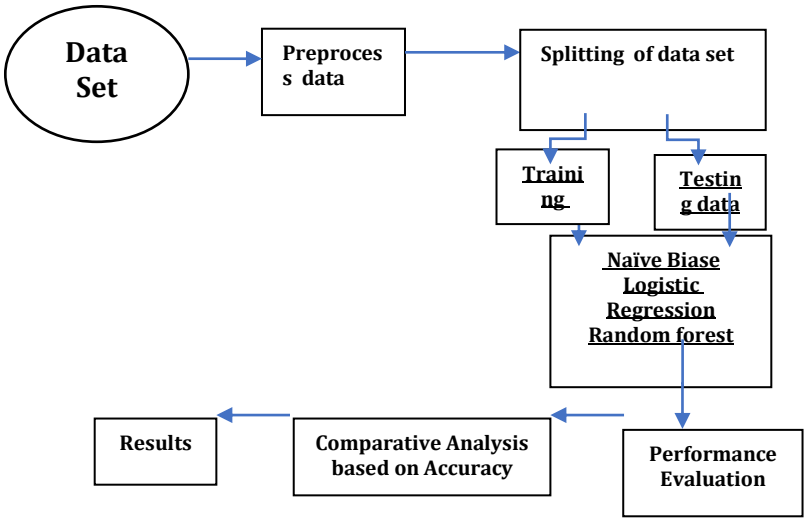

Fig.1: Proposed model

Here we collect a Liver dataset from UCI material repository. The data set contains 345 instances i.e. liver patient's data and 7 attributes. The main purpose for pre-processing is to remove duplication values in data. Then achieved data has sent for data splitting and the data is divided into two sets. One is training data of 70% and another one of 30% testing data. Different ML classifiers such as LR, GNB, and RF were applied to find the individual detection and prevention accuracy levels for level cancer.

3.2 Data preprocessing

Mean Corpuscular Volume: %1

Alkaline Phosphatase: %2 alkphos

Alanine Aminotransferase: %3 sgpt

Aspartate Aminotransferase: %4. sgot

Gamma-Glutamyl Transferase: %5. gammagt

Drinks: %6

Selector : %7 selector field used to split data into two sets

Algorithms

LR:- It is a supervised MLA which is a statistical method which is used to predict a binary outcome and categorical such as yes/no, True / False, 0 or 1 of a data set. This model is used predict a dependent data variable by applying one or more existing independent variables. This type of models are simple and it also creates predictions by applying easy-to-interpret mathematical formula. This LR model is an authorised statistical technique which is implemented comfortably for software and computation purposes.

GNB :- It is a classification technique of ML based on probabilistic approach supposing for each class succeeds a normal distribution having every parameter predicts the output variable independently. Through this algorithm various samples are classified with respect to test sample and training sample basing on functions and distance value.

RF :- This model uses multiple trees by eliminating over fitting. It is a popular ML algorithm for not only yielding accurate and precise result but also scalability and versatility with noiseless data.

diagnosed patients with liver disease: 145nos

patients not diagnosed with liver disease: 200nos

4. METRICS FOR EVALUATION

Absolute error

It is the modulus of the difference of experimental measurement and actual measurement;

Absolute error = |Experimental Measurement value –Actual Measurement value|

Relative error : It is the ratio of Absolute error and Actual error

RE=Absolute Error/Actual Error

RTMSE calculates the average difference between values predicted by a model and actual values

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

The determination of relative absolute error E_i of a particular model i is given by

$$E_i = \frac{\sum_{j-1}^{n} \left| P_{(ij)} - T_j \right|}{\sum_{j-1}^{n} \left| T_j - \overline{T} \right|}$$

Accuracy : Accuracy means the closeness of a given set of measurements to its true value.

Accuracy = (TR_PS+TR_N) / (TR_PS+TR_N+FS_P+FS_N)

Sick person– positive for disease

wellness- negative for disease

TR_PS : The nos of cases as positive classes, predicted as sick person.

TR_N : The nos of cases positive class incorrectly predicted as wellness

TR_PS: The nos of cases negative class for correctly predicted as sick person

FS_N: The nos of cases negative class for incorrectly predicted as wellness

PRECISION

It determines the accuracy of positive prediction of a model.

$$Precision = \frac{truepositives}{truepositives + falsepositives}$$

5. RESULT ANALYSIS

The main reason of experiments is to compare the performance in terms of the accuracy, of LGR, NB, RF in the Liver dataset. In terms of Accuracy, from table 1.3, Random Forest is having 0.75.

Accuracy

Algorithm	Accuracy
Logistic regression	0.69
Gaussian Naïve Bayes Algorithm	0.68
Random Forest	0.75

Name of Algorithms	MAE	MSE	RAE	RSE
Logistic regression	0.4151	0.4584	85.1648	92.8611
Bayes Algorithm	0.4597	0.5083	94.323	102.9673
Random Forest	0.3936	0.4704	80.7619	95.2954

Algorithm	Tr_pos	Tr_Neg	Fls_ pos	Fls_ neg
Logistic regression	25	54	21	14
Bayes Algorithm	19	59	27	9
Random Forest	31	55	15	13

Table 1.3 Accuracy

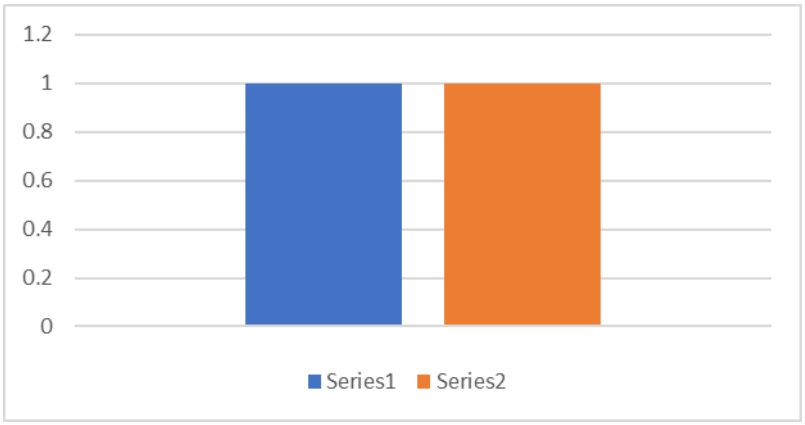

Fig.2: Series

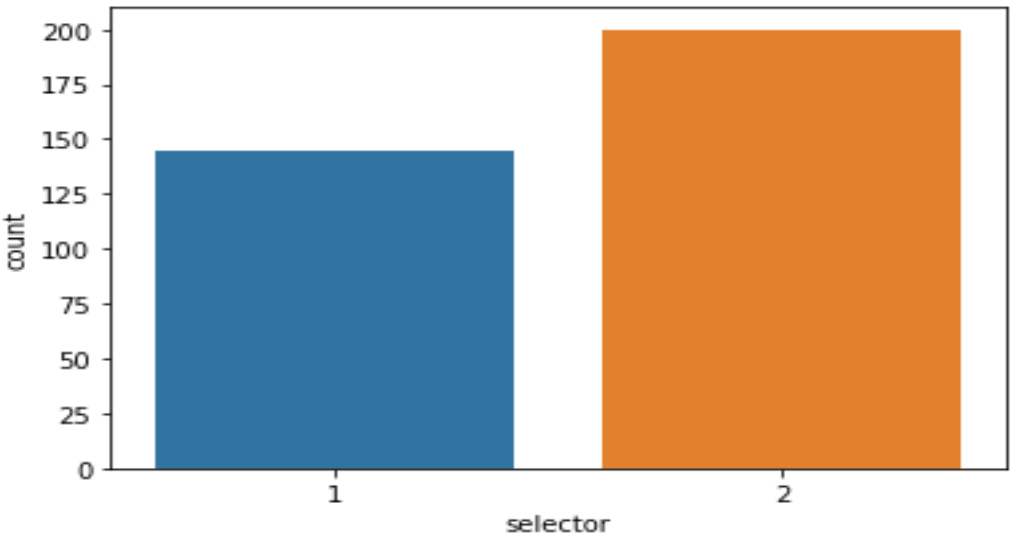

Fig.2: Evaluation

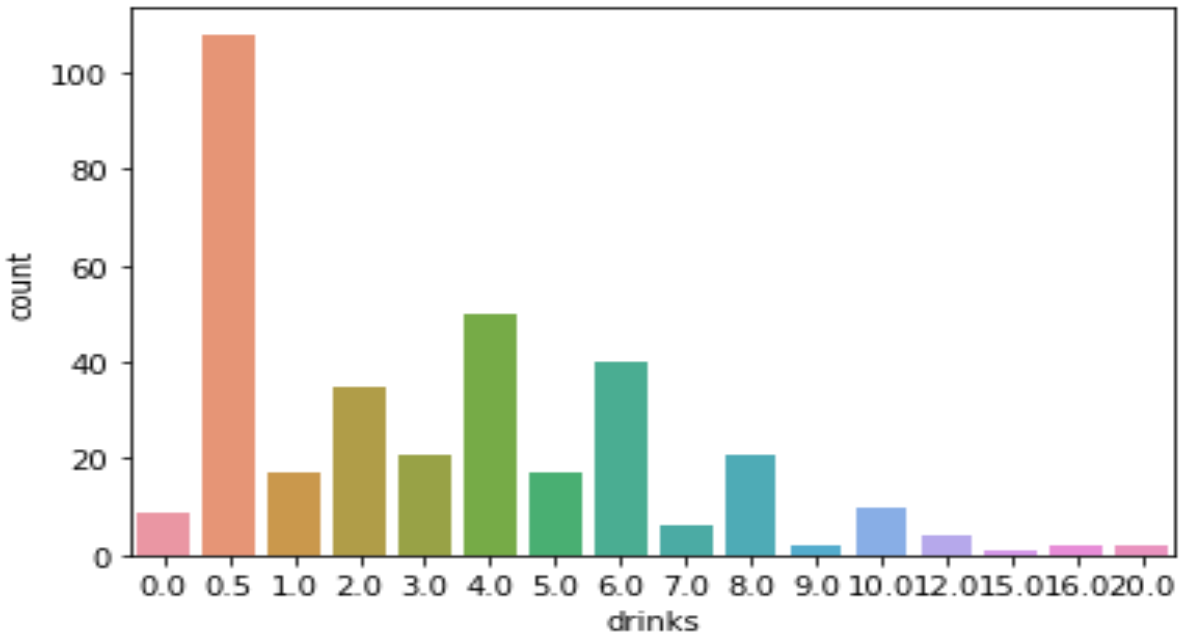

Fig.3: Model Evaluation

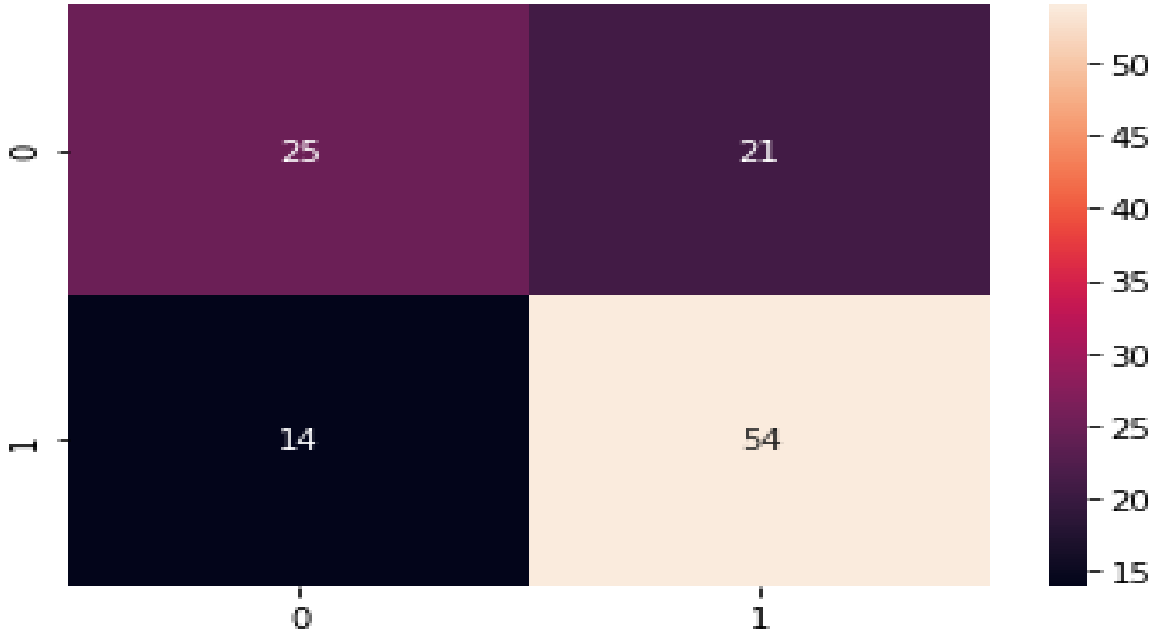

Fig.4:Gaussian Naive Bayes

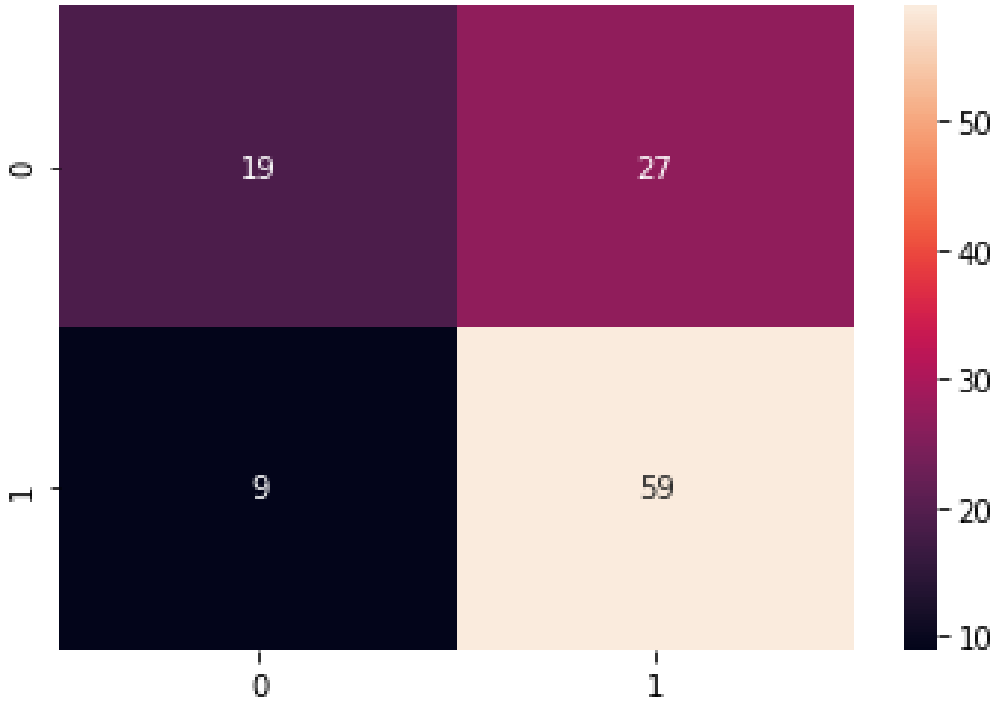

Fig.5 Random Forest

In [89]:

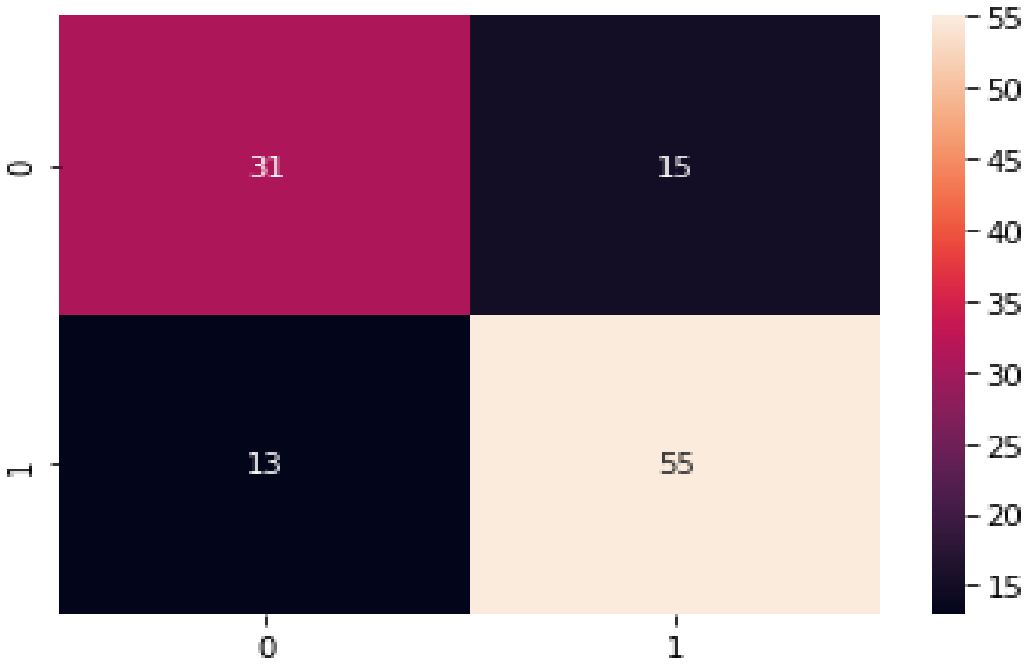

Fig.6: Logistic Regression

6. CONCLUSION

As liver related disease is a life risk, so a patient must be under constant medical supervision. Here machine learning is used for early prediction of this type of deadly disease. There are various algorithm used for analysing and prediction purposes. In this experiment RF model performs best result relating to accuracy with 0.75% and precision with 0.79% compared to other algorithms i.e. Naïve Bayes and Logistic regression. Therefore proposed model Random forest is robust which may used for prediction of liver disorder in any healthcare and diagnostic centre.

REFERENCES

[1] Asrani S, Devarbhavi H, Eaton J, Kamath P. Burden of liver diseases in the world. J Hepatol. 2019;70(1):151–71.

[2] A. Arjmand, C. T. Angelis, A. T. Tzallas, M. G. Tsipouras, E. Glavas, R. Forlano, P. Manousou, and N. Giannakeas, "Deep learning in liver biopsies using convolutional neural networks," in 2019 42nd International Conference on Telecommunications and Signal Processing (TSP). IEEE, 2019, pp. 496–499.

[3] L. A. Auxilia, "Accuracy prediction using machine learning techniques for indian patient liver disease," in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018, pp. 45–50

[4] Powell, E.E.; Wong, V.W.S.; Rinella, M. Non-alcoholic fatty liver disease. Lancet 2021, 397, 2212–2224. [CrossRef]

[5] Smith, A.; Baumgartner, K.; Bositis, C. Cirrhosis: Diagnosis and management. Am. Fam. Physician 2019, 100, 759–770. [PubMed]

[6] Macpherson I, Nobes J, Dow E, Furrie E, Miller M, Robinson E, Dillon J. Intelligent Liver Function Testing: Working Smarter to Improve Patient Outcomes in Liver Disease. The Journal of Applied Laboratory Medicine. 2020;5(5):1090–100.

[7] Rycroft, J.A.; Mullender, C.M.; Hopkins, M.; Cutino-Moguel, T. Improving the accuracy of clinical interpretation of serological testing for the diagnosis of acute hepatitis a infection. J. Clin. Virol. 2022, 155, 105239. [CrossRef]

[8] Thomas, D.L. Global elimination of chronic hepatitis. N. Engl. J. Med. 2019, 380, 2041–2050. [CrossRef]

[9] Rasche, A.; Sander, A.L.; Corman, V.M.; Drexler, J.F. Evolutionary biology of human hepatitis viruses. J. Hepatol. 2019, 70, 501–520. [CrossRef]

[10] Gust, I.D. Hepatitis A; CRC Press: Boca Raton, FL, USA, 2018.

[11] Yuen, M.F.; Chen, D.S.; Dusheiko, G.M.; Janssen, H.L.; Lau, D.T.; Locarnini, S.A.; Peters, M.G.; Lai, C.L.

Hepatitis B virus infection. Nat. Rev. Dis. Prim. 2018, 4, 1–20. [CrossRef]

[12] . Manns, M.P.; Buti, M.; Gane, E.; Pawlotsky, J.M.; Razavi, H.; Terrault, N.; Younossi, Z. Hepatitis C virus infection. Nat. Rev. Dis. Prim. 2017, 3, 1–19. [CrossRef] [PubMed]

[13] . Mentha, N.; Clément, S.; Negro, F.; Alfaiate, D. A review on hepatitis D: From virology to new therapies. J. Adv. Res. 2019, 17, 3–15. [CrossRef] [PubMed]

[14] Kamar, N.; Izopet, J.; Pavio, N.; Aggarwal, R.; Labrique, A.; Wedemeyer, H.; Dalton, H.R. Hepatitis E virus infection. Nat. Rev. Dis. Prim. 2017, 3, 1–16. [CrossRef]

[15] Han J, Kamber M and Pei J 2015 Morgan Kaufmann Series in Data Management Systems 230–24

[16] Nazim Razali1 , Aida Mustapha1 , Mohd Helmy Abd Wahab2 , Salama A Mostafa1 , Siti Khadijah Rostam1 2019 A Data Mining Approach to Prediction of Liver Diseases

[17] BUPA Liver Disorder Dataset. UCI repository machine learning databases

[18] Prof Christopher N. New Automatic Diagnosis of Liver Status Using Bayesian Classification