e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

A COMPARATIVE STUDY OF PROMINENT APPROACHES IN MACHINE LEARNING FOR PREDICTION AND DIAGNOSIS OF CHRONIC LIVER DISEASE

Asma Khatoon*1, Dr. Raisul Hasan*2

*1,2Dept. Of Computer Science And Engineering Rameshwaram Institute Of Technology And Management (AKTU) Lucknow, Uttar Pradesh, India.

ABSTRACT

Influenced by lifestyle factors such as excessive alcohol consumption, exposure to environmental toxins, consumption of contaminated or processed foods, and excessive misuse of prescription drugs, liver diseases are rapidly becoming a major health burden worldwide. Additionally, conditions such as viral hepatitis, obesity, and metabolic syndrome are major contributors to the increasing prevalence of liver disease. The liver plays a central role in metabolic processes, detoxification, and nutrient storage, making its dysfunction a major concern for overall health. Chronic liver disease (CLD) has become a widespread problem, affecting populations in both developed and developing countries, with a worrying trend being that it is being diagnosed more often in younger individuals, often in their mid-20s. This demographic shift highlights the urgent need for strong public health strategies, early detection tools, and preventive care to reduce its impact.

This paper presents a thorough review of the prominent ML applications developed for diagnosing and managing liver disease. It highlights integrating data mining techniques to enhance diagnostic accuracy and provide early intervention support for medical practitioners. Additionally, a comparative analysis of selected studies is performed to assess their methodology, results, and its practical deductions. This comparative study results are intended to contributing in the development of more efficient, reliable, and accessible diagnostic tools, ultimately helps in improving patient outcomes and addressing the growing burden of liver diseases.

Keywords: Chronic Liver Disease, Machine Learning, Diagnosis, Algorithm.

I. INTRODUCTION

The liver is a largest organ that serves as a crucial barrier between the gut and the bloodstream, filtering out harmful substances viz. toxins, antigens etc. in the body. It executes many essential functionalities including hormones, enzymes, and proteins, which are essential for metabolism, digestion, and immune regulation. The liver also detoxifies harmful substances by converting them into excretable forms, thus maintaining the body's metabolic balance and overall health [1].

Liver related disease includes a broad range of conditions and disorders that can impair the liver's ability to function optimally. While some liver disorders originate within the liver itself (primary liver disease), others occur as a secondary effect of systemic disorders that affect other organs. Given the liver's central role in metabolic and detoxification processes, its dysfunction can lead to a wide

range of complications, highlighting the importance of early diagnosis and effective management [1].

The diagnosis and prognosis of chronic liver disease (CLD) have received significant attention due to the critical role of the liver in maintaining metabolic balance, detoxification, and homeostasis. CLD represents a range of conditions, including hepatitis, cirrhosis, and hepatocellular carcinoma (HCC), which collectively impose a substantial global health burden. Alcohol intake, viral hepatitis infections (HBV & HCV), liver diseases related to obesity, and environmental toxins contribute to the increased prevalence of liver diseases. The irreversible nature of cirrhosis and its progression to HCC emphasize the need for early detection and effective clinical intervention to reduce morbidity and mortality.

Hepatic fibrosis, a progressive accumulation of extracellular matrix proteins in the liver, and its end stage, cirrhosis, have become significant global health concerns. Grief struck Cirrhosis represents the irreversible consequences of long-term fibrosis, where normal liver tissue is replaced by dense clumps of fibrotic tissue. These bands are caused by structural and functional defects of the liver. The main causes of cirrhosis are chronic hepatitis B and C infection as well as excessive alcohol use [2]. These etiological factors trigger ongoing liver damage and inflammation, gradually leading to fibrosis. Cirrhosis is a major cause for liver-related morbidity and mortality in worldwide, early detection is important to address these growing health challenges. Emphasizes the urgent need for implementation, precautionary measures and, new treatment strategies.

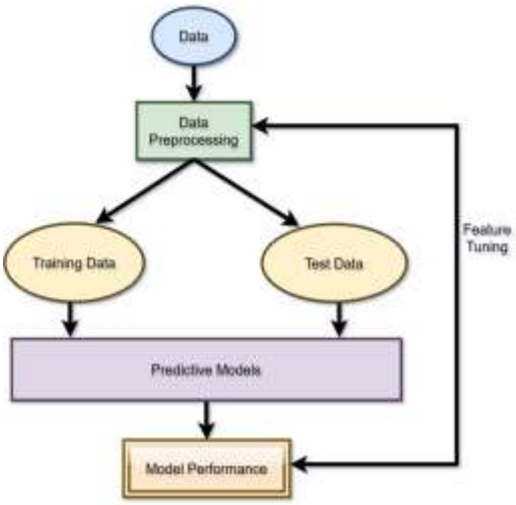
Infections due to chronic hepatitis C virus (HCV) ,it progresses disease slowly that may remain asymptomatic for years, posing challenges for timely detection and intervention. Although some individuals with chronic HCV infection have minimal or no liver damage, others may rapidly progress to liver conditions such as cirrhosis and hepatocellular carcinoma (HCC) [3]. These patients, mainly those with chronic liver diseases, are at high risk of developing HCC and require continuous monitoring for early diagnosis and effective management.

HCV is recognized globally as the leading cause of cirrhosis and HCC, with alpha-fetoprotein levels often being increased in infected patients. Rising incidence of HCC is expected to continue to increase [4]. some regions such as Egypt, where liver cancer is a major cause of death. Studies consider genetic abnormalities caused by hepatitis B virus (HBV) and external etiological factors such as HCV infection as the main cause of HCC development [5]. These factors can cause DNA damage, including mutations in genes such as p53 and disruptions in HBV genome integration , thereby emphasizing the critical need to prevent and treat hepatitis virus infection as the primary strategy against liver cancer.

With the advent of high-throughput sequencing, advanced imaging methods, and wearable health devices, the biomedical field has entered the era of big data. These technologies generate vast datasets, including biochemical markers, genetic sequences, and imaging outputs, which have enormous potential for understanding and managing the disease. However, the challenge lies in effectively analyzing and interpreting these datasets to derive meaningful insights. Thus, machine learning (ML) play an important role in addressing the challenges related with the diagnosis, management, and clinical management of chronic liver disease (CLD). These methods have significant relevance due to its ability to enhance accuracy and sensitivity of disease diagnosis and prognosis, which are important areas of focus in the contemporary medical research landscape.

Figure 1 outlines the sequential steps required to make predictions using an algorithm that must first undergo a training phase. This process highlights the significance of data preparation, feature selection, and optimization of model to ensure reliable and accurate results. By using these techniques, the likelihood of human error during the decision-making phase is significantly reduced, thereby improving clinical outcomes and promoting trust in automated diagnostic systems [6].

In light of these developments, this study attempts a comprehensive literature review of prominent ML methods applied to chronic liver disease research. These efforts aim to give insights into how these methods can be used to address clinical challenges, aid decision making, and pave the way for innovative diagnostic and therapeutic solutions.



The use of automated classification tools can reduce the burden of diagnosing and managing liver diseases. Data classification in clinical diagnostics typically involves two phases: a training phase, where the algorithm generates classifiers using the training dataset, and a classification phase, where model performance is evaluated using the test dataset [7].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

Feature selection processes, which are integral to improving the performance and accuracy of machine learning (ML) models, are broadly classified into filter and wrapper methods. Filter methods use data properties and are independent of classification algorithms [8], while wrapper methods leverage predictive accuracy of specific algorithms to determine feature relevance[9]. Among wrapper methods, advanced machine learning methods such as support vector machines (SVM) are particularly effective, because they consider the combination effects of features and consistently improve traditional filter methods in classification accuracy [10].

Machine learning models, especially those trained on datasets like Indian Liver Patient Dataset (ILPD) Techniques such as logistic regression (LR), K-nearest neighbor (KNN) and random forests (RF) consistently achieve accuracy above 70%, making them reliable alternatives to traditional statistical methods[11,12].

Main advantage of ML algorithms is their ability to analyze nonlinear relationships within data, making them highly effective for heterogeneous and multifactorial diseases such as CLD. For example, feature selection techniques – such as filter methods and wrapper methods – ensure that predictive models focus on the most relevant biomarkers, improving both sensitivity and specificity. However, the inclusion of diverse datasets and demographic information is critical to reduce biases and ensure equitable healthcare delivery, as disparities in diagnosis and treatment among different populations are well documented [[1].

In addition to improving diagnostic accuracy, these technologies offer significant benefits for disease diagnosis and management. Predictive models trained on biomarkers like aspartate aminotransferase (AST), alanine aminotransferase (ALT) and alkaline phosphatase (ALP) etc. can identify high-risk patients, monitor disease progression, and provide treatment. Integrating these methods into the clinical workflow can results into improvement of patient outcomes as well as reduced burden on healthcare professionals.

This research explores the role of machine learning techniques in addressing the challenges associated with chronic liver disease. By examining state-of-the-art models and their application to real-world datasets, we aim to provide a comprehensive understanding of their potential to transform the diagnosis and management of liver disease. Furthermore, we address these limitations and suggest future directions to enhance the applicability of these tools across various healthcare settings.

II. LITERATURE REVIEW

Diseases in humans are spreading rapidly now a days as compared to previous decades. Among them people affected by liver diseases is constantly increasing [13]. However, most liver diseases do not show significant symptoms in their starting stages. The liver is the primary target for insulin and its regulatory hormones such as glucagon. Hepatocellular carcinoma (HCC) patients involved in heavy alcohol consumption have a higher risk of developing cirrhosis than those who do not consume alcohol. Non-alcoholic fatty liver disease is the most common cause of liver failure. Cirrhosis, characterized by severe scarring of the liver, is the end stage of many liver conditions and is associated with a progressive decline in liver function, eventually leading to liver failure. Hepatocellular carcinoma is the most common primary liver cancer. Several factors influence disease progression, including age, sex, chronic alcohol use, and extent of viral exposure. The disease is more aggressive in patients infected with the hepatitis C virus (HCV) after the age of 40 and may progress more rapidly in men than women [3].

In the modern era of huge databases, it has become easier to extract data and gain insights to help treat various diseases [14]. Researchers use several strategies to extract insights from datasets, some of which involve machine learning (ML) classifiers for feature selection or extraction, while others do not.

Currently, data is being generated and stored at an unprecedented rate, providing researchers with an efficient way to solve problems of areas like medical imaging, finance, genomics, transactions, and security breaches etc. However, the presence of large amounts of both related and unrelated data can affect the performance of ML algorithms. Therefore, various methods are used for selecting and extracting of the most correlated features, which can predict diseases and other outcomes with greater accuracy. The development of machine learning (ML) has paved the way for its application in various fields, including disease diagnosis and prediction [[15], [16]]. Many studies have used ML techniques for predicting or detecting liver diseases (LDs). For example, Pasha et al. [17] proposed an LD prediction model and compared its accuracy with other ML algorithms such as random forest (RF), logistic regression (LR) and KNN etc .

III. DATA PREPROCESSING

The data preparation is considered as most important step in any analysis. A significant portion of healthcare data often contains missing values and other anomalies, which can reduce its overall utility. Data pre-processing plays an important role in improving data quality and reliability, ensuring that the results acquired from data mining process are accurate and valuable.

This approach is essential for delivering reliable results and making accurate predictions, which enables us to maximize the potential of the dataset by effectively applying machine learning techniques. As a result, several steps are taken to transform the data into a smooth and error-free dataset. This process serves as a prerequisite before starting the sequential steps of iterative analysis. The term 'data preprocessing' comprises a series of operations designed to refine and prepare data for analysis [18].

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction

Given the availability of raw data from real-world sources, data preprocessing becomes a critical requirement. Most real-world data contain errors, missing values, and anomalies, which may arise due to various reasons, such as non-collection of data, errors during data entry, technical issues in biometric devices, malfunctions, and more. Noisy data, which is characterized by errors and outliers, may arise from equipment failures, human errors in data entry, or other unpredictable factors. Addressing these issues are important to ensure the quality and reliability of the dataset for further analysis.

IV. TRAINING AND TESTING DATA

Generally, data is divided into two different categories: 1st is training data and second is testing data. The training data set has a known output and it is used for model training, enabling it to generalize its result to new data in the future. A testing data set, also called a testing subset which is used to predicting accuracy of model. In general, size of the training set is inversely proportional to the size of the testing set.

Both training and testing are performed to ensure accuracy and reliability of the model. The clinical measurements provided in the dataset aim to create a predictive model that determines whether a patient has liver disease or not. This prediction is generated based on clinical data available within the dataset, allowing informed and accurate predictions to be made.

V. DATASET AND ITS FEATURES

The ILPD dataset known as Indian Liver Patient Dataset, which can be found in Repository of Machine Learning Dataset at the University of California, Irvine (UCI), serves as the basis for the dataset selected for classification & experimental simulations. The Indian Liver Patient Dataset can be accessed at the following URL: https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset. This dataset provides valuable information in the diagnosis and classification of chronic liver disease.

The ILPD data collection had done from the northeastern region of Andhra Pradesh, India and focused on patients from this region. It contains a total of 583 records in which 441 patients are male, and 142 patients are female. It has a class label, named selector, which divides the dataset into two groups in 1st patients suffering from liver disease (416 cases) and in 2nd group people who do not have liver disease (167 occurrences). Each patient record contains 10 features, including age, sex, biochemical markers, and liver function test results. These attributes are given in table 1.

Table 1:

S.NO.	Features
1	Age
2	Gender
3	Total bilirubin

4	Direct bilirubin
5	Total protein
6	Albumin
7	A/G ratio
8	SGPT(alanine aminotransferase)
9	SGOT(aspartate aminotransferase)
10	Alkaline phosphatase (AlkaPhos)

A significant amount of research has been done using this dataset, with a focus on identifying liver disease and exploring clinical differences based on gender. For example, the dataset has been used to investigate differences in liver disease prediction between male and female patients, as previous studies have shown that certain biochemical markers differ based on gender.

This dataset has also supported research on algorithm fairness and bias. A notable study by Straw, and Honghan Wu, 2022 [19], analyzed supervised machine learning models to test for gender-based bias in liver disease prediction. Additionally, the dataset has been used to compare liver disease characteristics between US and Indian patients, providing insight into regional and demographic differences in disease patterns.

The Indian Liver Patient Dataset is used as a vital research resource for clinicians and researchers, it aims to develop a robust ML models for diagnosis of liver disease. Its diverse and well-documented features enable the creation of predictive systems that address gaps in healthcare assessment while striving for accuracy and objectivity.

VI. MACHINE LEARNING ALGORITHMS

ML is a statistical learning technique where each object in a dataset is characterized by a set of features 'or' attributes. This approach, conceived by computer scientists, has gained widespread popularity in recent years. Jerome Bruner is credited with introducing this concept, which has been further developed, particularly through the efforts of the IBM Watson Research Center.

A. Logistic Regression Algorithm

Logistic regression is a statistical method which is used for analysis of relationship between a dataset and one or more independent variables that act as predictors for an outcome. This type of dataset can have a variety of determinants. An outcome is evaluated using a binary variable, which means there are only two possible outcomes, such as binary categories like 1/0, yes/no, or true/false. Logistic regression is used to predict binary outcomes based on a set of independent variables. Dummy variables are used whenever the outcomes need to be presented in binary or categorical form.

LR is a special form of linear regression where the dependent variable is the log of the probability, making it suitable for categorical outcome variables. This statistical technique estimates the probability of a particular event by applying the collected data to a reliable statistical model. Basically, it predicts the probability of a particular outcome occurring based on given predictors.

The concept of LR was first introduced by one of the statistician David Cox in 1958. Using the binary logistic model, it is possible to calculate the probability of a binary response given one or more independent variables or predictors. This approach allows fr the interpretation of how risk factor presence can increases the probability of a particular output by a quantifiable percentage.

Sigmoid function $P = 1/1+e - (a+bx)$

Here P represents probability, a and b are parameter of Model.

B. Random Forest Algorithm

Random forest is a versatile method which is used for classification and regression tasks and is a part of the ensemble learning family. It is particularly effective when handling large datasets. The algorithm, developed by Leo Breiman, is recognized for its ability to increase the efficiency of decision trees by reducing the overall variance. During the training process, it creates multiple decision trees. The algorithm determines the class based on the majority vote (ranking) or computes the average prediction (regression) of individual trees in the end of process [20].

The process begins by exploring and selecting different potential features. Each randomly generated decision tree contributes its prediction, which is then aggregated across all trees in the forest. The second step involves counting the votes for each predicted outcome. In the last step, the prediction with the most votes becomes the final output of the random forest algorithm.

Random forest parameters provide flexibility, allowing users to make accurate predictions in a variety of applications. This adaptability makes random forests a powerful tool for solving complex classification and regression problems.

C. K Nearest Neighbour Algorithm

KNN is one of the most important chance-based category instance-based classification algorithms in machine learning. In any case, KNN takes a shot KNN works on the assumption that instances are close to similar instances in the category of similar instances [21]. KNN assigns an instance to the type.

The magnitude is fixed to the maximum of K neighbors. K is one Classification Algorithm Adjustment Problem [22].

VII. EVALUATION METRICS

We will make evaluation of our designed algorithm using the generated dataset for checking its effectiveness [23]. Additionally, the generated dataset will be used to estimate our model. We consider accuracy as a performance metric to measure effectiveness of our newly developed classification system and can be compared with existing methods.

True Positives (TP): These are instances in wehich model correctly identifies positive instances as positive.

True Negatives (TN): These are instances in which model correctly identifies negative instances as negative.

False Positives (FP): When negative instances are incorrectly classified as “positive” by the model then false positive condition occurs..

False Negatives (FN): When model incorrectly labels positive as negative instances in that case FN occurs.

Table 2:

Conditions→ ↓	**Actually Positive**	**Actually Negative**
Predicted Positive	True Positives (TPs)	False Positives (FPs)
Predicted Negative	False Negatives (FNs)	True Negatives (TNs)

Table 3: Performance comparison of the recent studies with the proposed method on the ILPD dataset

Author & Reference	**Year**	**Classifier**	**Accuracy**
K. Gupta et al.[15]	2022	LR	57.00
		RF	63.00
		KNN	57.00
J. Singh et al. [24]	2020	LR	74.36
		RF	71.87
Kuzhippallil et al.[25]	2020	LR	76.00
		RF	88.00
		K-NN	79.00
Amare et al.	2019	KNN	68.61

M. Babu et al.[26]	2016	KNN	64.00

VIII. CONCLUSION

Detecting chronic liver disease at an early stage remains a significant challenge, as the liver often continues to function normally despite partial damage. This study provides a comprehensive review of clinical methods and machine learning (ML) methods for automatic detection of chronic liver disease (CLD). The research systematically analyzed four major aspects: i) available databases, ii) ML-based classification and prediction methods, iii) AI-powered smart assistants for CLD patients, iv) and performance used in existing studies.

The results show that algorithms such as random forests (RFs),Logistic regression and K nearest neighbour(KNN) in particular, demonstrate a remarkable ability to learn deeply and efficiently to automatically retrieve and classify CLD data.

However, there remains an urgent need for more reliable and robust methods that can accurately predict the probability of a patient having CLD while recommending preventive measures. In addition, future research should focus on developing models that are able to predict comorbidities often associated with CLD. Achieving high accuracy rates is an ongoing challenge, highlighting the need for continued development in this field.

ACKNOWLEDGEMENTS

I would like to express my heartiest gratitude to Sayed Azhar Sabri sir (Full stack developer at Doc Cirrus, Germany) and Samar Fatma ma'am for their invaluable help and guidance during this research. Special thanks go to Dr Raisul Hasan sir (director of Rameshwaram Institute of Technology and Management) for their insightful feedback and encouragement. Additionally, we appreciate all the CS department faculties of Rameshwaram Institute of Technology & Management (RITM) for providing the necessary resources for this work.

IX. REFERENCES

[1] Hashem, Esraa M., and Mai S. Mabrouk. "A study of support vector machine algorithm for liver disease diagnosis." American Journal of Intelligent Systems 4.1 (2014): 9-14.

[2] Golla, Keerthi, Joel B. Epstein, and Robert J. Cabay. "Liver disease: current perspectives on medical and dental management." Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology 98.5 (2004): 516-521.

[3] Liang, T. Jake, et al. "Pathogenesis, natural history, treatment, and prevention of hepatitis C." Annals of internal medicine132.4 (2000): 296-305.

[4] Johnson, P. J. "Hepatocellular carcinoma: is current therapy really altering outcome?." Gut 51.4 (2002): 459-462.

[5] Mabrouk, Mai S., Esraa M. Hashem, and Amr Sharawy. "Statistical approaches for hepatocellular carcinoma (HCC) biomarker discovery." American Journal of Bioinformatics Research 2.6 (2012): 102-109.

[6] Bellazzi, Riccardo, and Blaz Zupan. "Predictive data mining in clinical medicine: current issues and guidelines." International journal of medical informatics 77.2 (2008): 81-97.

[7] Mitchell, Tom M., and Tom M. Mitchell. Machine learning. Vol. 1. No. 9. New York: McGraw-hill, 1997.

[8] Ding, Chris, and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data." Journal of bioinformatics and computational biology 3.02 (2005): 185-205.

[9] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46 (2002): 389-422.

[10] Chai, Hong, and Carlotta Domeniconi. "An evaluation of gene selection methods for multi-class microarray data classification." Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics. 2004.

[11] Jin, Hoon, Seoungcheon Kim, and Jinhong Kim. "Decision factors on effective liver patient data prediction." International Journal of Bio-Science and Bio-Technology 6.4 (2014): 167-178.

[12] Adil, Syed Hasan, et al. "Liver patient classification using logistic regression." 2018 4th International Conference on Computer and Information Sciences (ICCOINS). IEEE, 2018.

[13] Pasyar, Pezhman, et al. "Hybrid classification of diffuse liver diseases in ultrasound images using deep convolutional neural networks." Informatics in Medicine Unlocked 22 (2021): 100496.

[14] Joloudari, Javad Hassannataj, et al. "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection." Informatics in medicine unlocked 17 (2019): 100255.

[15] Gupta K, Jiwani N, Afreen N, Divyarani D. Liver disease prediction using machine learning classification techniques. Proc. - 2022 IEEE 11th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2022 2022; (August): 221–6.

[16] Shaikh, F. J., and D. S. Rao. "Prediction of cancer disease using machine learning approach." Materials Today: Proceedings 50 (2022): 40-47.

[17] Pasha, Syed Nawaz, et al. "Liver disease prediction using ML techniques." AIP Conference Proceedings. Vol. 2418. No. 1. AIP Publishing, 2022.

[18] Bosnić, Zoran, et al. "Mining data from hemodynamic simulations for generating prediction and explanation models." IEEE Transactions on Information Technology in Biomedicine 16.2 (2011): 248-254.

[19] Straw, Isabel, and Honghan Wu. "Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction." BMJ health & care informatics 29.1 (2022).

[20] Nelwamondo, Fulufhelo V., Shakir Mohamed, and Tshilidzi Marwala. "Missing data: A comparison of neural network and expectation maximization techniques." Current Science (2007): 1514-1521.

[21] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." Pattern recognition 40.7 (2007): 2038-2048.

[22] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN Model-Based Approach in Classification (pp. 986–996). Springer,Berlin, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39964-3_62.

[23] Islam, Md Tanvir, et al. "Typical and non-typical diabetes disease prediction using random forest algorithm." 2020 11th International conference on computing, communication and networking technologies (ICCCNT). IEEE, 2020.

[24] Singh J, Bagga S, Kaur R. Software-based prediction of liver disease with feature selection and classification techniques. Procedia Comput Sci 2020;167(2019): 1970–80.

[25] Kuzhippallil MA, Joseph C, Kannan A. Comparative analysis of machine learning techniques for Indian liver disease patients. In: 2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020; 2020. p. 778–82. https://doi.org/10.1109/ ICACCS48705.2020.9074368

[26] Babu MSP, Ramjee M, Katta S, Swapna K. Implementation of partitional clustering on ILPD dataset to predict liver disorders. Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS 2016:1094–7.

https://doi.org/10.1109/ICSESS.2016.7883256. 0