ISSN: 2633-4828 Vol. 5 No.2, June, 2023

International Journal of Applied Engineering & Technology

ML-BASED LIVER DISEASE DETECTION: A COMPREHENSIVE EVALUATION OF VARIOUS MACHINE LEARNING ALGORITHMS

Aman Kumar[1] and Randeep Singh[2]

[1, 2]Department of Computer Science & Technology, IEC University, Baddi, Solan, Himachal Pradesh, India

aman11304832@gmail.com[1] and [2]randeeppoonia@gmail.com[2]

ABSTRACT

Liver disease is a significant health issue worldwide, with millions of people affected each year. Early and accurate detection of liver disease is crucial for timely intervention, treatment, and improved patient outcomes. In recent years, machine learning techniques have shown promising results in aiding the detection and diagnosis of various diseases, including liver disease. This review paper focuses on the utilization of XG Boost, a popular and powerful machine learning algorithm, for liver disease detection. Assessment of different studies, methodologies, and performance metrics will be discussed to highlight the effectiveness and potential applications of XG Boost in liver disease diagnosis.

Training and testing data were used in machine learning techniques to predict categories. The study evaluated binary classifier machine learning methods (i.e., artificial neural network, random forest (RF), and XG boost & PSO) to identify patients with liver disorders, enabling improved diagnosis by medical experts using a publically available liver disease data set. The RF significantly increased the accuracy score as compared to alternative methods. As a result, research suggests that machine learning methods predict liver disease by including the risk variables, which might improve the inference-based patient diagnosis.

Keywords: Liver disease, XG Boost, machine learning, diagnosis, performance evaluation, comparative analysis.

1 INTRODUCTION

Liver disease encompasses a range of conditions, including liver infections, liver cancer, alcoholic liver disease, and viral hepatitis. Traditional diagnostic methods can be expensive, time-consuming, and invasive, leading to a growing interest in machine learning approaches for more efficient and accurate diagnosis. XG Boost, an ensemble learning algorithm, has gained significant traction due to its ability to handle complex datasets and nonlinear relationships, making it potentially suitable for liver disease detection.

Research on predicting liver illness has been ongoing, with many studies concentrating on the creation of predictive models that use medical data to identify and predict problems connected to the liver. For predicting liver illness, conventional ML approaches including LR, SVM, & decision trees have been used extensively throughout the years. Particularly when used with structured data, such patient demographics, liver function tests, and medical history, these techniques have produced encouraging outcomes. The use of DL approaches has recently become a potent strategy for predicting liver disease. Particularly when working with medical imaging modalities like ultrasound, MRI, and CT scans, DNN, in particular CNNs and RNNs, have shown amazing skills in image-based liver disease detection. Early-stage liver disease identification is challenging since there are no obvious signs. Complications of liver diseases are not identified quickly enough as the liver functions normally even when it is partially damaged [1]. Even a skilled medical professional cannot predict the early signs, yet they can be identified. The key to significantly lengthening the patient's life is early diagnosis. Modern preventive medicine relies on ML approaches to identify diseases from health care databases. The early signs can be recognised, while being undetectable even to a skilled medical professional. It has a significant role in medical decision-making and focuses on incorporating various risk indicators into prediction tools [2, 3]. As the healthcare data is increased gradually, machine learning give access to analysing massive amounts of data to rapidly [4]. Machine learning is being used by several businesses to advance medical diagnoses.

2 LITERATURE SURVEY

Bharti P, Girdhar R, and Verma AK. provides a comprehensive literature survey on the use of XGBoost, a popular machine learning algorithm, for the detection of liver diseases [5]. The authors discuss various research studies, methodologies, datasets, and performance metrics used in the field. They also highlight the challenges and limitations of existing approaches and suggest potential future research directions. Overall, the review aims to provide insights into the advancements and potential of XGBoost for liver disease detection. Phumratanaprapin W, and Sripanidkulchai explores the use of machine learning models, including XGBoost, for the diagnosis of liver diseases [6]. The authors analyze various studies published from 2008 to 2018 and discuss the performance of different models in terms of accuracy, sensitivity, and specificity. They also identify the limitations and challenges in current research and provide recommendations for future improvements in liver disease detection using machine learning. Cheng H, Cui W, and Zhang Y et al. specifically focuses on computer-aided diagnosis in breast lesions and pulmonary nodules, it discusses the application of deep learning architectures, including XGBoost, in medical imaging analysis [7]. The paper provides insights into the potential use of XGBoost and other machine learning algorithms in liver disease detection, given their success in related medical imaging applications. Hashemi SR, Ghalehjegh GT, and Ghalehjegh SH. examines the utilization of machine learning algorithms, including XGBoost, for liver disease diagnosis and classification [8]. The authors analyze various studies published from 2014 to 2019 and present an overview of different approaches, datasets, and performance metrics used in liver disease detection. They also discuss the limitations and future prospects of machine learning algorithms for liver disease diagnosis.

The researchers look at the literature on machine learning in hepatology and liver transplant medicine in this review [9]. They give a general review of the benefits and drawbacks of ML technologies as well as their possible applicability to hepatology clinical and molecular data. Research on liver illness has used ML to analyse a variety of data sources, including clinical, demographic, molecular, radiological, and pathological data. The authors believe that the clinical practise of hepatology and transplantation will alter as a result of the application of ML technologies to produce prediction algorithms. The chance to learn about the ML tools available and their applications to hepatology-related topics will be offered to readers of this review. This study [10] largely concentrated on the three image capture modalities of computed tomography, magnetic resonance imaging, and ultrasound. In particular, preprocessing, attribute analysis, and classification strategies to complete clinical diagnostic tasks were covered in their thorough study, along with the benefits and drawbacks of each preceding stage. We investigated and compared popular denoising, deblurring, and segmentation techniques during preprocessing. Nonlinear models are used most frequently for denoising. Deep neural networks, on the other hand, were commonly used for deblurring and automatically segmenting regions of interest. The most popular strategy for attribute analysis included texture characteristics. The support vector machine was primarily used for categorization across three different methods of picture capture. Comparative investigation reveals that convolutional neural networks based on deep learning provide the greatest results. The performance of the prediction can be enhanced by taking into account biopsy samples or pathological parameters such as overall stage, margin, and differentiation. Additionally, a technological breakthrough to overcome data constraint issues and enhance prediction performance is anticipated shortly with advancements in machine learning models. This study [11] presented a thorough evaluation of the advancements made in using artificial intelligence to predict and diagnose liver illnesses. It then summarized the studies' associated shortcomings and suggested further research.

The technique [12] begins the pipeline by imputing the missing values and outliers for pre-treatment. The essential characteristics needed for classification are then extracted using integrated feature extraction on top of the pre-processed data. The recommended technique was being strengthened by a simulated study. The suggested method integrates a number of ML methods, such as the ensemble voting classifier, support vector machine (SVM), multilayer perceptron (MLP), random forest (RF), logistic regression (LR), and K-nearest neighbour (KNN). In terms of predicting liver disorders, the proposed system has an accuracy of 88.10%, a precision of 85.33%, a recall of 92.30%, an F1 score of 88.68%, and an AUC score of 88.20%. In comparison to the most

recent research, our suggested approach produced outcomes that were 0.10–18.5% better. The results imply that the suggested approach may be utilised to support a doctor's liver disease diagnosis.

In this research [13], a paradigm for creating a Clinical Decision Support System (CDSS) that tackles feature selection and class imbalance was provided. In this framework, the dataset was balanced at the data level, and feature selection was carried out using a wrapper technique. For testing, the Indian Liver Patient Dataset (ILPD), Thoracic Surgery Dataset (TSD), and Pima Indian Diabetes (PID) datasets from the University of California Irvine (UCI) machine learning repository was employed. The datasets were balanced using Orchard's technique and the Synthetic Minority Over-Sampling Technique (SMOTE). For the purpose of choosing a feature subset, a wrapper strategy utilising Chaotic Multi-Verse Optimisation (CMVO) was put out. The fitness function was the arithmetic mean of the Matthews correlation coefficient (MCC) and F-score (F1), assessed by a Random Forest (RF) classifier. Following the selection of the pertinent features, classification was performed using an RF that consists of 100 estimators and employs the Information Gain Ratio as the split criterion. For the ILPD, the classifier obtained a 0.65 MCC, 0.84 F1, and an accuracy of 82.46%; for the TSD, a 0.74 MCC, 0.87 F1, and an accuracy of 86.88%; and for the PID dataset, a 0.78 MCC, 0.89 F1, and an accuracy of 89.04%. The performance of the framework was compared with other studies in the literature, and the impacts of balance and feature selection on the classifier were examined. The outcomes demonstrated that, in terms of the three performance indicators employed, the suggested framework is competitive. The Wilcoxon test results demonstrated the statistical superiority of the suggested approach.

In this study [14], a unique deep classifier made up of pre-trained deep convolutional neural networks (CNNs) was presented to categorise the liver state. ResNeXt, ResNet18, ResNet34, ResNet50, and AlexNet were employed, along with fully connected networks (FCNs). Transfer learning can be used to extract deep features that can offer enough categorization data. Then, an FCN can depict pictures of the illness in its many stages, including normal liver, liver hepatitis, and cirrhosis. To discriminate between these liver pictures, two-class (normal/cirrhosis, normal/hepatitis, and cirrhosis/hepatitis) and three-class (normal/cirrhosis/hepatitis) classifiers were developed. A hybrid classifier is presented in order to combine the weighted probabilities of the classes produced by each separate classifier since two-class classifiers performed better than three-class classifiers. The class with the highest score is then chosen using a majority voting technique. The experimental findings demonstrated a classification accuracy of 86.4% for liver pictures divided into three groups using ResNet50 and a hybrid classifier. The results reveal that the first group's sensitivity and specificity are 90.9% and 86.4% for the differentiation between normal and cirrhosis liver, respectively, as well as 90.9% and 81.8% for normal and hepatitis liver. AUC—an additional evaluation metric of the novel feature reduction along with accuracy, for robust feature selection aimed at effective disease risk prediction—was proposed as part of an advanced hybrid ensemble gain ratio feature selection (AHEG-FS) model [15]. This model consists of four major feature selection techniques: an ensemble feature selection, a gain ratio feature selection, a backward feature elimination, and an AUC. With the first two methods, the subsets of significant and highly scored characteristics are produced. Next, nine ML algorithms are synchronised with the suggested model. Additionally, the third and fourth techniques of the proposed model evaluate the AUCs for the aforementioned ML algorithms and employ backward feature elimination to eliminate the redundant features, resulting in the acquisition of the best subsets of highly contributing features that yield the results with the highest precision. Thus, four benchmarked heart disease datasets from the University of California, Irvine ML repository—Cleveland, Hungarian, Statlog, and Switzerland—are employed. The outcomes are promising. With 46.15% less characteristics, the greatest AUC and accuracy are reached, at 99.00% and 95.47%, respectively. With convergent speed, an accuracy improvement of 6.18% over recent investigations was made.

The suggested technique [16] sought to identify liver disorders early on utilising data from liver function tests. Class imbalance is an issue with many real-world datasets, including data on the diagnosis of liver disease. The circumstance when there are more or fewer observations from one class than from another is referred to as an imbalance. Because they consider neighbours equally, traditional K-Nearest Neighbour (KNN) or fuzzy KNN

classifiers do not perform well on the unbalanced dataset. To address the difficulties with data imbalance, the weighted variation of fuzzy KNN assigns a significant weight for the neighbour belongs to the minority class data and a relatively small weight for the neighbour belongs to the majority class. An upgraded version of fuzzy-NWKNN was developed in this study called variable-neighbor weighted fuzzy K nearest neighbour approach (Variable-NWFKNN). On three real-world imbalance liver function test datasets from BUPA, ILPD from UCI, and MPRLPD, the suggested Variable-NWFKNN algorithm is put into practise. The Variable-NWFKNN's accuracy was determined to be 73.91% (for the BUPA Dataset), 77.59% (for the ILPD Dataset), and 87.01% (for the MPRLPD Dataset) when compared to the current NWKNN and Fuzzy-NWKKNN techniques. Additionally, the preprocessing method TL_RUS is utilised, which increased accuracy by 78.46% (BUPA Dataset), 78.46% (ILPD Dataset), and 95.79% (MPRLPD Dataset).

In research work [17], several data mining techniques—including KNN, DT, ANFIS—are discussed. These techniques are employed to create a decision support design that might aid the doctor in identifying liver illness from the database. Each algorithm's conductance is assessed in means of accuracy, sensitivity, precision, specificity. The effectiveness of these approaches is surveyed and provided. The suggested investigation is carried out using the ILPD from the University of California, Irvine databases [18]. The various characteristics in the liver patient database are used to predict the likelihood of developing liver diseases, include age, direct bilirubin, gender, total bilirubin, Alkphos or sgot, among others. The Liver Patient database is used to test the accuracy of several classification approaches, like LR, SMO, RF alpproach, NB, J48, & IBk. Different classifier outcomes are contrasted with & without the use of feature selection approaches. Utilising feature selection & classification estimation methodologies depend on software engineering models, ILDPS is created.

3 Proposed work

In this paper, we describe a brand-new approach for detecting liver illness early on utilizing machine learning approaches. Our study's main aim is to establish a reliable prediction design that could assist in the early identification of liver disease, enabling quick medical treatments and better patient outcomes. We will begin our analysis by pre-processing the data, obtaining the dataset for liver illness from an Excel file. We manage null or missing values by replacing them with the means of the corresponding columns, assuring data accuracy. In order to prepare the information for ML approaches, we also encode textual or category values into numerical representations, which optimizes data handling and processing. We employ PCA to reduce dimensionality because the database is highly dimensional.

3.1 XGBoost Algorithm

A comprehensive overview of the XG Boost algorithm will be presented, including its hierarchy of decision trees, gradient boosting principles, regularization techniques, and hyperparameter tuning. The unique advantages of XG Boost, such as handling missing data, feature importance analysis, and model interpretability, will be explored in the context of liver disease detection.

3.2 Principal Component Analysis (PCA)

In data mining and machine learning, dimensionality reduction is frequently accomplished using the PCA approach. It attempts to maintain as much pertinent information as possible while transforming high-dimensional data into a lower dimensional environment.

3.3 Random Forest

RF is a potent ensemble learning technique that is frequently applied to classification jobs in data mining and ML. During the training stage, various DT are built, & their estimations are merged through voting to create the final classification.

3.4 Particle Swarm Optimization (PSO)

PSO is a metaheuristic optimization technique that draws its inspiration from the social behavior of bird flocks and fish schools. In data mining and other technical domains, it is frequently utilized to resolve optimization issues. PSO uses a population of moving particles to explore the search space for the best answer. Each particle symbolises a potential solution to the issue in its place, and it moves according to its own experience as well as the experiences of the particles around it.

4 Proposed Methodology

Using a combination of data pre-processing, dimensionality reduction, feature selection, and sophisticated machine learning algorithms, we provide a methodology for the prediction of liver disease in this research study.

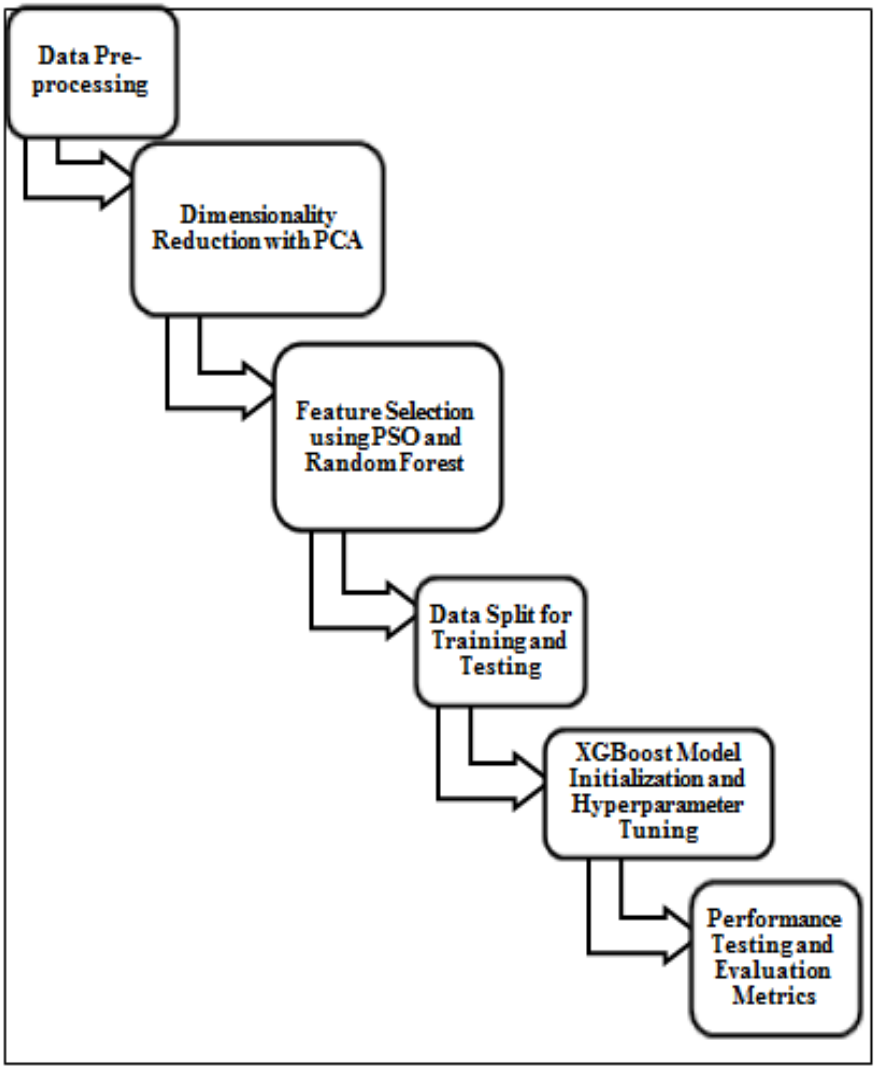

Fig. 1. Methodology for the Prediction of Liver Disease

4.1 Data collection

The first step in any machine learning study is to collect relevant data. In this case, medical data related to liver diseases would need to be collected. This could include patient demographics, medical history, lab test results, imaging data, and any other relevant information.

4.2 Data Preprocessing

Once the data is collected, it needs to be preprocessed to remove any noise or outliers and ensure it is in a suitable format for analysis. This may involve cleaning the data, handling missing values, normalizing or scaling features, and splitting the dataset into training and testing subsets.

4.3 Feature Selection/Extraction

In this step, relevant features or variables that are most likely to be predictive of liver disease are selected. This can be done using various techniques, such as statistical analysis or domain expertise. Feature extraction techniques like principal component analysis (PCA) or other dimensionality reduction methods may also be applied to reduce the number of features.

4.4 XGBoost Model Training

XGBoost is a machine learning algorithm that is commonly used for classification tasks. It is an ensemble algorithm that combines multiple weak learners (decision trees) to create a strong predictive model. The training dataset is used to train the XGBoost model, where the algorithm learns to predict the presence or absence of liver disease based on the selected features.

4.5 Model Evaluation

After training the model, its performance needs to be evaluated. This is typically done using evaluation metrics like accuracy, precision, recall, or area under the ROC curve. Cross-validation techniques can also be employed to validate the model on different subsets of the dataset.

4.6 Hyperparameter Tuning

XGBoost has various hyperparameters that can be adjusted to optimize the model's performance. Techniques like grid search or Bayesian optimization can be used to find the best combination of hyperparameters that maximize the model's accuracy.

4.7 Model Deployment and Testing:

Once the model is trained and optimized, it can be deployed for liver disease detection on new, unseen data. The model should be tested on this data to assess its performance in real-world scenarios.

5 Performance Evaluation

Different evaluation metrics used in assessing the performance of the XG Boost model for liver disease detection will be reviewed. Accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), and other relevant metrics will be analyzed to determine the algorithm's effectiveness in correctly identifying liver disease.

6 Comparative Analysis

The comparison of XG Boost with other machine learning algorithms commonly used in liver disease detection, such as logistic regression, random forest, and support vector machines. Their respective strengths and weaknesses will be evaluated to ascertain whether XG Boost outperforms these algorithms.

7 Challenges and Future Directions

Challenges in liver disease detection using ML approaches include data quality issues, imbalanced datasets, interpretability concerns, generalization challenges, and the need for clinical adoption.

Future scope lies in personalized medicine, leveraging multi-modal data fusion for a comprehensive understanding, real-time monitoring for early detection, incremental learning for continuous improvement, and advancements in explainable AI for better interpretability and clinical integration. These avenues hold promise for enhancing liver disease diagnosis, prognosis, and treatment, ultimately improving patient outcomes.

8 CONCLUSION

The conclusion summarizes the key findings of this review paper, highlighting the efficacy of XG Boost in liver disease detection and its potential for clinical implementation. Recommendations for future research areas and the promise of XG Boost for early and accurate liver disease detection are emphasized.

REFERENCES

1. Chieh Chen Wu, Wen Chun Yeh, Wen Ding Hsu, Md Mohaimenul Islam, Phung Anh (Alex) Nguyen, Tahmina Nasrin Poly, Yao Chin Wang, Hsuan Chia Yang, Yu Chuan (Jack) Li, "Prediction of fatty liver disease using machine learning algorithms," TMU Research Centre of Artificial Intelligence in Medicine, College of Medical Science and Technology, Taipei Municipal Wanfang Hospital TMU Research Centre of Cancer Translational Medicine.
2. P. Groves, B. Kayyali, D. Knott, S.V. Kuiken, "The big data revolution in healthcare: Accelerating value and innovation," 2016.
3. Charleonnan, T. Fufaung, T. Niyomwong, W Chokchueypattanakit, S. Suwannawach, N. Ninchawee, "Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques," MITiCON2016.
4. Bohr A, Memarzadeh K., "The rise of artificial intelligence in healthcare applications," Artificial Intelligence in Healthcare, pp. 25–60, doi: 10.1016/B978-0-12-818438-7.00002-2, Epub 2020 Jun 26, PMCID: PMC7325854.
5. Bharti P, Girdhar R, and Verma AK. "Machine learning based early detection of liver diseases: a review". Journal of Ambient Intelligence and Humanized Computing, 11(6), 2387-2406. doi: 10.1007/s12652-019-01534-w, 2020.
6. Phumratanaprapin W, and Sripanidkulchai K., "Machine learning models for liver disease diagnosis: a systematic review". EXCLI Journal, 17, 1004-1028. doi: 10.17179/excli2018-1421, 2018
7. Cheng H, Cui W, and Zhang Y et al., "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans". The Scientific World Journal, 2017, 9326846. doi: 10.1155/2017/9326846, 2017.
8. Hashemi SR, Ghalehjegh GT, and Ghalehjegh SH. "Utilizing machine learning algorithms for liver disease diagnosis and classification: a systematic literature review". Computer Methods and Programs in Biomedicine, 175, 153-165. doi: 10.1016/j.cmpb.2019.02.034, 2019.
9. Spann, A. Yasodhara, J. Kang, K. Watt, B. Wang, A. Goldenberg, and M. Bhat, "Applying machine learning in liver disease and transplantation: a comprehensive review," Hepatology, vol. 71, no. 3, pp. 1093–1105, 2020.
10. R. A. Khan, Y. Luo, and F.-X. Wu, "Machine learning based liver disease diagnosis: A systematic review," Neurocomputing, vol. 468, pp. 492–509, 2022
11. Z. Yao, J. Li, Z. Guan, Y. Ye, and Y. Chen, "Liver disease screening based on densely connected deep neural networks," Neural Networks, vol. 123, pp. 299–304, 2020.
12. Ruhul Amin, Rubia Yasmin, Sabba Ruhi, Md Habibur Rahman, Md Shamim Reza, "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms," Informatics in Medicine Unlocked, vol. 36, 2023, 101155, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2022.101155.
13. S. Sreejith, H. Khanna Nehemiah, A. Kannan, "Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection," Comput. Biol. Med, vol.126, 2020, Article 103991, 10.1016/j.compbiomed.2020.103991

14. S.J. Pasha, E.S. Mohamed, “Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction”, Inform Med Unlocked, 32 (June) (2022), Article 101064, 10.1016/j.imu.2022.101064.

15. P. Kumar, R.S. Thakur, “Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach”, vol.80, Issue 11, May 2021, pp. 16515-16535.

16. R. Kalaiselvi, K. Meena and V. Vanitha, "Liver Disease Prediction Using Machine Learning Algorithms," International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675756.

17. R. Kalaiselvi, K. Meena and V. Vanitha, "Liver Disease Prediction Using Machine Learning Algorithms," International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675756.

18. Jagdeep Singh, Sachin Bagga, Ranjodh Kaur, “Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques,” Procedia Computer Science, vol. 167, 2020, pp. 1970-1980, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.226.