International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8339
Article Received: 12 July 2023 Revised: 05 September 2023 Accepted: 18 September 2023

Forecasting Liver Disorders with Machine Learning Models

Pragati Singh[1], Ashok Kumar Yadav[2], Sanjeev Gangwar[3]
[1]Department of Computer Science & Engineering
U.N.S.I.E.T VBS Purvanchal University Jaunpur, Uttar Pradesh, India.
E-mail: pragatisingh552@gmail.com
[2]Department of Information Technology
U.N.S.I.E.T VBS Purvanchal University Jaunpur, Uttar Pradesh, India.
E-mail: ashok231988@gmail.com
[3]Department of Computer Applications
U.N.S.I.E.T VBS Purvanchal University Jaunpur, Uttar Pradesh, India.
E-mail: gangwar.sanjeev@gmail.com

Abstract— Liver disorders encompass a spectrum of ailments that impact the liver, a crucial organ responsible for a variety of vital bodily functions. These functions encompass metabolic processes, detoxification, protein synthesis, and the production of bile. Liver maladies can arise from various sources, such as viral infections (e.g., hepatitis), excessive alcohol consumption, conditions related to obesity (like non-alcoholic fatty liver disease), autoimmune conditions, genetic predisposition, or exposure to toxins. Common signs and symptoms may encompass fatigue, jaundice, abdominal discomfort, and digestive problems. In our study, we gather data and employ five distinct machine learning classification algorithms: Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbor, and XG Boost. After constructing models and evaluating their performance, we observed that XG Boost achieved an impressive accuracy rate of 99.8%.

Keywords- Liver disease, Machine learning, Random Forest, Naïve Bayes, K-Nearest Neighbour, Decision Tree, XG Boost.

I. INTRODUCTION

Liver disease encompasses a diverse group of medical conditions that affect the liver, one of the body's largest and most vital organs. The liver, located in the upper right portion of the abdomen, plays an indispensable role in maintaining overall health. It performs a multitude of functions, which are pivotal to various physiological processes within the body. Here's a more detailed exploration of liver disease.

A. Functions of the Liver

The liver is a true multitasker, performing a wide range of functions that are essential for survival and wellbeing:

i. Metabolism: The liver processes and metabolizes nutrients from the food we eat. It regulates blood sugar levels by storing excess glucose as glycogen and releasing it when needed. It also helps break down fats, proteins, and carbohydrates.

ii. Detoxification: One of the liver's most crucial roles is detoxification. It filters toxins, drugs, and metabolic waste products from the bloodstream, preventing them from circulating throughout the body and causing harm.

iii. Protein Synthesis: The liver produces various proteins, including blood-clotting factors and albumin. These proteins are vital for maintaining blood volume, regulating fluid balance, and preventing excessive bleeding.

iv. Bile Production: The liver produces bile, a digestive liquid that gets stored in the gallbladder and is discharged into the small intestine as required. Bile plays a crucial role in facilitating the digestion and absorption of fats and fat-soluble vitamins.

v. Storage: The liver acts as a storage facility for essential nutrients, such as vitamins and minerals. It can release these nutrients into the bloodstream when the body requires them.

B. Causes of Liver Disease

Liver disease can stem from various factors, including:

i. Viral Infections: Hepatitis viruses, particularly hepatitis A, B, and C, can cause inflammation and damage to the liver. These infections can be acute (short-term) or chronic (long-term).

ii. Alcohol Misuse: Persistent and excessive intake of alcohol can result in alcoholic liver disease, spanning from fatty liver to more severe disorders such as alcoholic hepatitis and cirrhosis.

iii. Non-Alcoholic Fatty Liver Disease (NAFLD): This condition is associated with the accumulation of fat in the liver and is often linked to obesity, metabolic syndrome, and insulin resistance.

iv. Autoimmune Disorders: Certain individuals may experience autoimmune liver disorders, including autoimmune hepatitis, primary biliary cirrhosis, and primary sclerosing cholangitis, where the immune system erroneously attacks and harms liver cells.

v. Genetic Factors: Certain genetic conditions like hemochromatosis and Wilson's disease can lead to liver problems.

vi. Cirrhosis: Persistent liver injury stemming from diverse origins can lead to cirrhosis, marked by the development of fibrous tissue within the liver. Cirrhosis has the potential to culminate in liver failure.

vii. Toxic Exposures: Exposure to specific chemicals, drugs, or toxins can harm the liver, leading to acute or chronic liver disease.

C. Symptoms and Diagnosis

The indications of liver ailments can fluctuate significantly based on the particular disorder and its progression. Typical signs encompass:

i. Tiredness

ii. Yellowing of the skin and eyes, known as jaundice.

iii. Enlargement and discomfort in the abdominal region

iv. Dark urine

v. Pale-colored stools

vi. Nausea and vomiting

vii. Unexplained weight loss

To diagnose liver disease, healthcare providers often use a combination of blood tests, imaging studies (such as ultrasounds or CT scans), and sometimes liver biopsies to assess the extent of liver damage and determine the underlying cause.

D. Treatment and Prevention

Treatment for liver disease depends on its cause and severity. It may involve lifestyle modifications, medications, dietary changes, or, in severe cases, liver transplantation. Early diagnosis and treatment are crucial to prevent further liver damage and complications like cirrhosis or liver failure.

E. Preventing liver disease involves

i. Limiting alcohol consumption.

ii. Maintaining a healthy weight and addressing obesity-related factors.

iii. Getting vaccinated against hepatitis viruses.

iv. Practicing safe sex to prevent hepatitis transmission.

v. Avoiding risky behaviours related to drug use.

vi. Managing underlying health conditions that can affect the liver, such as diabetes.

II. LITERATURE REVIEW

Deepika Bhupathi et al [2] conducted a study on the identification of liver diseases using machine learning approaches. In their research, they applied various machine learning classification algorithms, including Support Vector Machine, Naïve Bayes, K-Nearest Neighbor, Classification and Regression Tree, and Linear Discriminant Analysis. Their dataset consisted of 583 entries, with 416 instances related to liver disease and the remaining without liver disease. The authors developed models and assessed their accuracy using different algorithms, ultimately determining that K-Nearest Neighbor yielded the highest accuracy of 91.7%.

Ruhul Amin et al [3] introduced a method for predicting chronic liver disease patients through the integration of projection-based statistical feature extraction with machine learning algorithms. In their study, they employed the Indian Liver Patient Dataset (ILPD) from the University of California, Irvine (UCI) repository and applied various machine learning algorithms, including logistic regression, random forest, K-nearest neighbor, support vector machine, multilayer perceptron, and an ensemble voting classifier. Their system achieved noteworthy performance metrics, including an accuracy rate of 88.10%, precision of 85.33%, recall of 92.30%, F1 score of 88.68%, and an AUC score of 88.20% for liver disease prediction.

Biju, Kalyani et al [4] presented a study on the diagnosis of chronic liver disease utilizing machine learning methods. In their research, they utilized machine learning algorithms such as Logistic Regression, K-Nearest Neighbor (KNN), and Random Forest to identify liver-related disorders. Their system had the capability to distinguish between four stages of liver diseases, including a healthy liver, fatty liver, liver fibrosis, and liver cirrhosis. The process of predicting liver disease encompassed several steps, including data pre-processing, feature extraction, and classification. They sourced their datasets from the Kaggle database, specifically Indian liver patient records.

Karmakar et al [5] focus on the early identification of liver diseases using machine learning techniques. They emphasize the liver as the body's largest internal organ. Their study introduces a liver disease detection model that relies on multiple logistic regression and utilizes secondary data from the UCI repository. To assess the model's performance, they employ a confusion matrix and implement a 10-fold cross-validation approach, considering various metrics such as accuracy, precision, specificity, sensitivity, and kappa. The primary objective of their work is to enhance the accuracy and effectiveness of an expert system for early detection of liver diseases.

Kulkarni et al [6] conducted research on improving the preprocessing techniques for the detection of liver disease using ensemble machine learning algorithms. They applied data to training using various ensemble learning algorithms, including Gradient Boosting, XG Boost, Bagging, Random Forest, Extra Tree, and Stacking. Subsequently, they compared the outcomes of these six models with each other and with models from previous studies. Notably, their proposed model, which combines the Extra Tree classifier and Random Forest, demonstrated superior performance, achieving the highest testing accuracy rates of 91.82% and 86.06%, respectively. This highlights the practicality and effectiveness of their approach in real-world liver disease detection.

Latha et al [7] introduced a method for predicting liver disease using machine learning. In their research, they harnessed the potential of machine learning, especially in the context of healthcare and medical applications. Machine learning was employed to uncover hidden patterns in data, ultimately improving the accuracy of diagnosis and decision-making. Liver disease, which presents a substantial health hazard and can result from factors like toxins, medications, and excessive alcohol use, was the central focus of the study. The research incorporated a diverse set of pertinent features and leveraged K-Nearest Neighbors (KNN) technology to enhance the prediction of liver disease.

Geetika Singh et al [8] put forward a method for predicting and analyzing liver disease using Extreme Learning Machine (ELM). Their study employed the ILPD dataset for model evaluation. They meticulously evaluated the performance of their proposed model by experimenting with various activation functions and hidden neuron counts. According to their calculations and findings, the proposed model outperformed the existing models currently in use, indicating its superiority in the field of liver disease prediction and analysis.

Niha et al [9] conducted a study focused on comparing different machine learning algorithms for the prediction of liver disease. In this research, the author explored a range of machine learning techniques for assessing the accuracy of liver disease diagnosis, including Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest. The study revealed variations in the performance of these methods in terms of accuracy, precision, and sensitivity. The results of the analysis highlighted that Logistic Regression achieved the highest level of accuracy among the tested algorithms.

III. DATASET

The proposed system utilizes machine learning principles to enhance its predictive capabilities. It follows a structured process of model training, testing, and prediction. The initial step involves users providing a range of personal information, including age and gender, as well as specific blood test results like total Bilirubin, direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), total proteins, albumin, and the A/G (Albumin to Globulin) ratio.

The input values represent essential medical parameters and are typically obtained from a standard blood test report. Once the user provides this information, the system proceeds to analyze it. This analysis entails comparing the user's data to a well-established training dataset, where machine learning models have been fine-tuned for accuracy.

Based on the patterns and relationships identified during the model training phase, the system makes a prediction regarding the user's health status. Specifically, it assesses whether the user is at risk for liver disease or not. This prediction is delivered to the user, providing valuable insights into their health.

To ensure the system's reliability and effectiveness, it undergoes a rigorous evaluation process. This evaluation primarily revolves around assessing the accuracy of its predictions. The system's performance is typically quantified using metrics such as a confusion matrix. This matrix facilitates an in-depth analysis of true positives, true negatives, false positives, and false negatives, enabling a thorough assessment of the diagnostic capabilities of the system.

IV. PROPOSED METHODOLOGY

A. Support Vector Machine

Support Vector Machine (SVM), commonly referred to as SVM, is a widely adopted supervised learning technique that finds application in both classification and regression tasks. However, its primary utilization predominantly falls within the classification domain of machine learning. The central goal of the SVM algorithm revolves around the creation of an optimal line or decision boundary capable of effectively separating data points in an n-dimensional space into distinct classes. This ensures precise categorization of future data points into their respective classes.

The paramount achievement of SVM is the identification of the ideal decision boundary, often termed as a "hyperplane." To accomplish this, SVM strategically selects the most influential data points or vectors that play a significant role in defining the hyperplane. These remarkable instances are referred to as "support vectors," thereby giving the algorithm its distinctive name, Support Vector Machine.

The primary aim of the Support Vector Machine (SVM) algorithm is to pinpoint the ideal hyperplane that efficiently segregates data into two distinct classes, all the while maximizing the gap or margin between these two classes. This objective can be formally articulated as follows:

$$\min\ 1/2\, p\hat{}2 \qquad (1)$$

Subject to

$$a_i\ (px+b)-1\ \geq 0, i=1\ldots.m \qquad (2)$$

In this equation, p represents the weight vector, b stands for the bias term, x denotes the feature vector, a_i corresponds to the class label of the sample, and m represents the total number of samples.

Upon employing the Support Vector Machine algorithm on the dataset, we achieve an accuracy rate of 72.1%.

B. Decision Tree

A Decision Tree is a supervised learning technique that is applicable to both classification and regression problems, although it is more commonly used for classification tasks. It takes on the form of a tree-like structure where internal nodes represent attributes from the dataset, branches depict decision rules, and terminal nodes (leaves) represent the final output or classification. Decision Trees comprise two primary types of nodes: Decision Nodes and Leaf Nodes. Leaf Nodes signify the ultimate outcomes and do not have further branches, while Decision Nodes are responsible for making decisions and can branch into multiple choices. This structure is analogous to a tree, originating from a root node and extending into branches.

Decision Trees can handle both numerical and categorical data, typically in a binary YES/NO manner. The process initiates by computing the entropy of the parent node, followed by determining the information gain, which is obtained by subtracting the weighted sum of entropies of the child nodes from the entropy of the parent node.

Information Gain = Entropy - [(Weighted Average) * Entropy(each feature)] (3)

The attribute exhibiting the greatest information gain is selected as the root node, and this procedure continues until the classification is finalized. Each node within the decision tree corresponds to a specific symptom drawn from the set A = {A1, A2, A3, A4, ... Aj}, with A representing conditional attributes. The values or ranges for the i-th symptom are denoted as Vi, j, and the final decisions are furnished by the leaves. Additionally, there is a set B = {B1, B2, B3, B4, ... Bk}, and their binary representations are presented as Wdk = {0, 1}.

Upon employing the Decision Tree algorithm on the dataset, we achieve an accuracy rate of 99%.

C. K-Nearest Neighbour

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used to address both classification and regression tasks. In KNN, predictions for new data points are generated based on "feature similarity," meaning that the algorithm evaluates how closely the new data point resembles existing data points in the training dataset. KNN maintains a repository of all available examples and assigns categories to new cases by employing a similarity metric. The parameter 'K' in KNN represents the number of nearest neighbors considered to determine the predominant category through a majority vote. Typically, 'K' is set to the square root of 'n,' where 'n' signifies the total number of data points. In cases where 'n' is even, adjustments are made to ensure 'K' is an odd value, enhancing the robustness of the selection process. Importantly, KNN is classified as a "lazy learner," making it particularly suitable for scenarios where the dataset is well-labeled, devoid of noise, and relatively small in size.

The K-Nearest Neighbors (KNN) algorithm involves a series of steps:

1. Select the Value of 'k': Begin by specifying the number of neighbors, denoted as 'k,' that you wish to consider when making predictions.
2. Compute Euclidean Distances: Calculate the Euclidean distance between the data point you want to classify and all other data points in the dataset.
3. Identify Closest Neighbors: Identify the 'k' data points with the shortest Euclidean distances from the data point you're trying to classify.
4. Count Categories: Among these 'k' nearest neighbors, tally the number of data points belonging to each category or class.
5. Majority Vote: The category with the highest count among the 'k' nearest neighbors is assigned to the data point you're classifying. In other words, it's determined by majority vote.

Upon employing the K-Nearest Neighbors algorithm on the dataset, we achieve an accuracy rate of 96.5%.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8339
***Article Received:** 12 July 2023 **Revised:** 05 September 2023 **Accepted:** 18 September 2023*

D. Naive Bayes

The Naïve Bayes algorithm, a supervised learning technique, leverages Bayes' theorem to address classification problems effectively. Its primary application lies in text classification, typically involving datasets with high dimensions. The Naïve Bayes Classifier stands out as a straightforward yet highly efficient classification method, capable of constructing rapid machine learning models for swift predictions. This classifier operates on a probabilistic foundation, making predictions by assessing the probabilities associated with objects. The use of "Naïve" in the term Naïve Bayes stems from the assumption that the existence of one specific feature in the dataset is independent of the presence of other features.

For instance, in the context of identifying fruits based on attributes like color, shape, and taste, the Naïve Bayes algorithm treats these attributes (e.g., red color, spherical shape, and sweet taste) as independent factors when recognizing an apple. In other words, each feature autonomously contributes to the identification of an apple without considering interdependencies.

The term "Bayes" is derived from the algorithm's reliance on Bayes' Theorem, a fundamental probability principle that guides its decision-making process.

Being a probabilistic classifier, it makes predictions by considering the likelihood of a particular event occurring. Let's denote G as a set of features, represented as G = {G1, G2, G3, ..., Gn}, and H as the output containing {Yes/No}. To calculate the probability, we employ the following equation.

$$X(H|G) = X(G|H).X(H)/X(G) \qquad (4)$$

Where,

X(H|G) represents the posterior probability,

X(G|H) denotes the likelihood probability,

X(G) signifies the marginal probability, and

X(H) stands for the prior probability

Upon employing the Naïve Bayes algorithm on the dataset, we achieve an accuracy rate of 56%.

E. Random Forest

The Random Forest algorithm is a highly employed technique in machine learning, serving various purposes including classification and regression. In the Random Forest methodology, numerous decision trees are created using bootstrapped samples from the training data. At each decision node within these trees, a random subset of the available features is selected to determine the most suitable split.

The Random Forest algorithm follows these steps:

1. Initially, select 'm' random samples from 'n' different datasets.
2. Construct a decision tree for each training dataset.
3. Each decision tree produces a prediction.
4. Ultimately, the prediction with the highest number of votes among all decision trees becomes the final prediction result.

F. XG Boost

XG Boost, short for "Extreme Gradient Boosting," stands out as an exceptionally optimized distributed gradient boosting library designed to streamline the process of training machine learning models efficiently and at scale. It belongs to the ensemble learning category, where it consolidates predictions from multiple weaker models to generate a more robust prediction. XG Boost has gained widespread recognition as a leading machine learning algorithm, primarily owing to its ability to effectively handle large datasets and consistently deliver top-tier performance across various machine learning tasks, including both classification and regression.

One notable feature of XG Boost is its proficiency in handling missing data, making it particularly well-suited for real-world datasets that often have gaps, without requiring extensive data preprocessing. Additionally, XG Boost provides built-in support for parallel processing, enabling the training of models on substantial datasets within reasonable timeframes. XG Boost finds applications in a diverse range of areas, including Kaggle competitions, recommendation systems, and predicting click-through rates, among others. It also offers a high degree of configurability, allowing for fine-tuning of various model parameters to enhance performance.

XG Boost is essentially an implementation of the Gradient Boosted decision tree technique, which has gained significant prominence in various Kaggle Competitions. In this algorithm, decision trees are built sequentially, with a key role assigned to the concept of weights. These weights are assigned to each independent variable and serve as inputs for the decision tree, contributing to the prediction process. Importantly, variables that are incorrectly predicted by one tree receive increased weight and are subsequently incorporated into the construction of the next decision tree. These individual classifiers or predictors then collaboratively combine their outputs to produce a robust and more accurate model. XG Boost showcases its versatility by being suitable for various tasks, including regression, classification, ranking, and the handling of user-defined prediction problems. When employing the XG Boost algorithm on a dataset, an impressive accuracy rate of 99.8% was achieved.

V. RESULT & DISCUSSION

Liver diseases are increasingly becoming one of the most lethal health issues in various countries. The rising number of people afflicted by liver ailments is attributed to excessive alcohol consumption, exposure to harmful gases, consumption of contaminated food, pickles, and pharmaceuticals. Researchers are scrutinizing datasets containing information on liver patients to develop classification models for the prediction of liver disease. This dataset was employed to evaluate predictive algorithms with the goal of reducing the workload on healthcare professionals. In our research, we proposed the utilization of a diverse set of machine learning algorithms, encompassing Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Decision Tree, Random Forest, and XG Boost, to conduct an extensive evaluation of liver disease in patients.

Table I: Summary of the Model's Classification Performance

Proposed Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
SVM	72.1	72	100	84
KNN	96.5	97	98	98
Random Forest	99.3	99	100	100
Decision Tree	99	99	99	99
XG Boost	99.8	100	100	100
Naïve Bayes	56	96	40	56

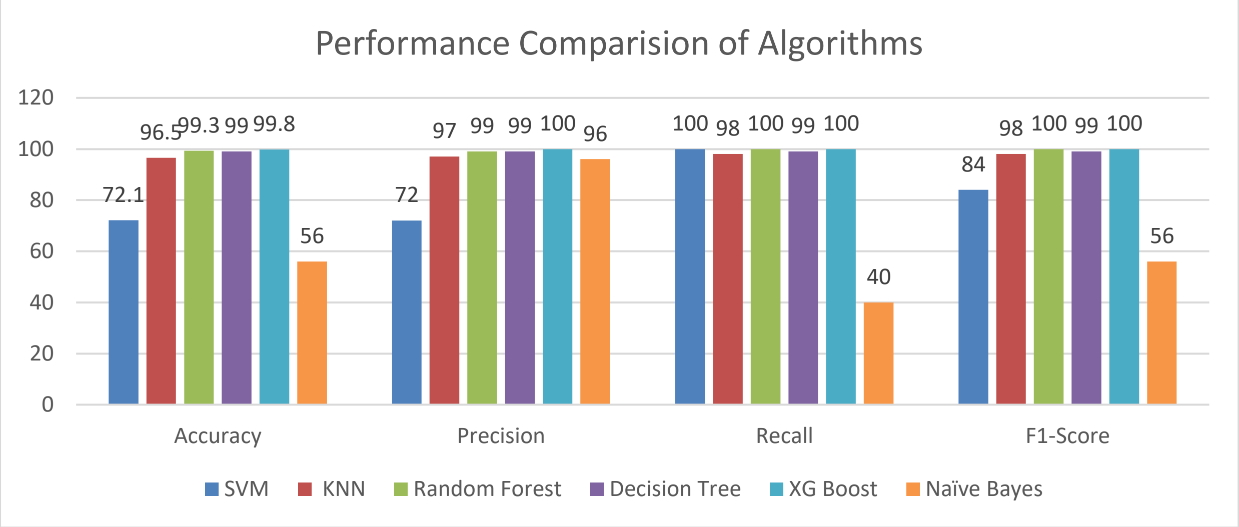

Figure 3. Visual Depiction of Classification Algorithms

The information provided in Table 1 and Figure 1 unequivocally indicates that the XG Boost classifier outperformed all other classification algorithms, achieving exceptional metrics including 99.8% precision, 100% accuracy, 100% F1-score, and 100% recall in comparison.

VI. CONCLUSION

Liver disease is a grave ailment that poses a significant risk to human health, necessitating immediate medical intervention. Medical practitioners traditionally rely on pathological approaches to generate medical reports regarding a patient's condition. However, this study's primary focus is on the early detection of liver disease through the application of machine learning techniques. Numerous algorithms, including SVM, XG Boost, K-nearest neighbor, logistic regression, and naive bayes, are at one's disposal. Among these, the XG Boost algorithm exhibits remarkable accuracy, achieving a perfect prediction rate of 99.8%. Looking ahead, the models developed through this research could potentially find integration into an application designed to predict early-stage liver disease.

REFERENCES

[1] https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset.

[2] Bhupathi, Deepika & Tan, Christine Nya-Ling & Sremath Tirumala, Sreenivas & Ray, Sayan. (2022). Liver disease detection using machine learning techniques.

[3] Yasmin, Rubia & Amin, Ruhul & Reza, Md. (2023). Design of Novel Feature Union for Prediction of Liver Disease Patients: A Machine Learning Approach. 10.1007/978-981-19-8032-9_36.

[4] Biju, Kalyani. (2023). Diagnosis of Chronic Liver Disease using Machine Learning Techniques. International Journal for Research in Applied Science and Engineering Technology. 11. 346-351. 10.22214/ijraset.2023.53305.

[5] Karmakar, Shibam & Pratihar, Subham & Roy, Shreeja & Das, Sulekha & Chaudhuri, Avijit. (2023). Early Detection of Liver Disease by using Machine Learning. international journal of

engineering technology and management sciences. 7. 271-276. 10.46647/ijetms.2023.v07i02.031.

[6] AGYEI , I. T. . (2021). Simulating HRM Technology Operations in Contemporary Retailing . International Journal of New Practices in Management and Engineering, 10(02), 10–14. https://doi.org/10.17762/ijnpme.v10i02.132

[7] Md, Abdul & Kulkarni, Sanika & Jackson, Christy & Vaichole, Tejas & Mohan, Senthilkumar & Iwendi, Celestine. (2023). Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease. Biomedicines. 11. 581. 10.3390/biomedicines11020581.

[8] M, Latha. (2022). Prediction of Liver Disease using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology. 234-241. 10.48175/IJARSCT-5673.

[9] Singh, Geetika & Agarwal, Charu. (2023). Prediction and Analysis of Liver Disease Using Extreme Learning Machine. 10.1007/978-981-19-5443-6_52.

[10] Niha, Shaik & Lakshmi, Kolla & Blessi, Pidathala & Lakshmi, Thamatam & Chowdary, Yarramasu & Rao, Mr. (2023). A Comparison of Machine Learning Algorithms for Predicting Liver Disease. International Journal for Research in Applied Science and Engineering Technology. 11. 2221-2230. 10.22214/ijraset.2023.49954.