

LIVER DISEASE PREDICTION

Peravath Gopal Naik¹, B.Harish Kumar Reddy²

¹*MCA Student, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*

²*Assistant Professor, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*

Abstract

Liver illnesses are a major global health concern, and factors like alcohol consumption, exposure to toxic substances, and poor dietary choices are linked to an increase disorders when they are detected early and accurately predicted. In order to improve diagnostic accuracy, this research investigates the application of machine learning algorithms to the prediction and classification of liver disease. Using three models-Support Vector Machine (SVM), Logistic Regression, and Random Forest, the study makes use of the Indian Liver Patient Dataset. The models were assessed based on accuracy following extensive data preprocessing and analysis. Logistic Regression outperformed SVM and Random Forest with the best accuracy. Future work may include expanding the dataset and exploring advanced techniques to further improve prediction accuracy.

Keywords: Liver, Dataset, ML

Introduction

Liver disease is a vital issue in the public health sector, and the increasing prevalence of this pathology is due to causes such as alcohol consumption, exposure to toxic substances, and poor dietary habits. Early diagnosis and accurate prediction of liver disease can significantly alleviate the burden on healthcare systems and improve patient outcomes. In recent years, the use of machine learning algorithms has shown promise in enhancing diagnostic accuracy by analyzing complex datasets.

The current study is an extension of Aiwale et al's (2023) [1] research, which explores the subject of the prediction of liver disease using three major machine learning models such as Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). The utilization of the Indian Liver Patient Dataset (ILPD) was done to identify the most effective algorithm for predicting liver disease by comparing their performance in different metrics such as accuracy and recall. Their findings demonstrated that Logistic Regression outperformed both SVM and Random Forest, achieving a remarkable accuracy of 98.82%, making it the most reliable model for liver disease prediction in this context. We point out the potential of machine learning techniques in decreasing diagnostic errors, thus the capability of early intervention and better patient outcomes. In addition, we are also talking about the impact of data preprocessing on the model's performance and we are also suggesting some areas for future research in liver disease prediction. Various machine learning algorithms, including Support Vector Machine (SVM), Random Forest, and Logistic Regression, have been employed to predict liver disease. Each algorithm has its strengths, with Random Forest often demonstrating superior accuracy and reliability in diagnosing liver conditions. Future research should focus on expanding datasets and exploring more sophisticated algorithms, including ensemble techniques, to further improve prediction accuracy. This could lead to significant advancements in public health and patient care, particularly in managing liver diseases.

Importance of Data Preprocessing: The impact of data preprocessing on model performance is crucial. Properly prepared data can enhance the effectiveness of machine learning models, leading to improved diagnostic outcomes.

Literature Survey

Rahman, Shamrat, Tasnim, Roy, and Hossain, [13]“A comparative study on liver disease prediction using supervised machine learning algorithms” Many people throughout the globe suffer from chronic liver disease, making it a major international mortality. It's caused by things like being

overweight, having an undetected hepatitis infection, and drinking excessively. Serious complications from this syndrome include abnormal nerve function, hepatic encephalopathy hemoptysis or vomiting blood, liver failure, renal failure, jaundice. Chronic liver disease is difficult and costly to diagnose. The primary goal of this study is to apply six different supervised machine learning classifiers to develop a reliable method of identifying people who suffer from chronic liver disease. With the hope of lowering the astronomical expenses involved with diagnosing chronic liver disease, this research compares the efficacy of many machine learning algorithms for making such a prediction. In this study, researchers used six different machine learning algorithms.

Durai, V., Ramesh, S. and Kalthireddy, D., [14]“Liver disease prediction using machine learning” Medical diagnosis and prognosis are two areas that have benefited greatly from data mining tools. Data mining methods have been used on a mountain of medical records. Liver-related disorders are becoming more common as the prevalence of obesity and poor lifestyles rises sharply. In this study, researchers use an extensively studied classification algorithm on patient data to determine the likelihood that a certain individual suffers from liver disease. Given that methods for analyzing both patient data and classifier data already exist, the more pressing concern is developing methods that more accurately anticipate the same final outcome. In all, there are 5 stages to this procedure. At first, the min-max method is used on the raw data from the UCI repository's liver patient dataset. The second step makes use of PSO feature selection to identify crucial characteristics. From complete normalized datasets of liver patients, this helps extract the subset of essential data. The next stage, classification algorithms, is used to make comparisons and put things into categories. The fourth stage is called Accuracy Calculation. Root Mean Square value and Root Error value are used. Evaluation is the fifth and last stage. A straightforward assessment procedure is carried out, as required by the research, to protect the credibility of a reliable representation of the results. The J48 algorithm has a 95.04 percent accuracy rate, making it the best performing algorithm for feature selection.

Azam, Rahman, Iqbal, and Ahmed, 2020[15] “Prediction of liver diseases by using few machine learning based approaches” Improvements in healthcare have always been an extremely important part of human civilization. As medical science and technology advance, patients are more required to undergo treatment using cutting-edge methods and tools. Accuracy in the medical sciences is increasingly dependent on the use of machine learning methods. In this study, we developed computational model building strategies for reliable prediction of liver disease. In order to forecast liver illnesses, we employed various powerful classification algorithms including Random Forest, etc. In order to enhance prediction capabilities, our works employ hybrid model design and comparison analysis. At first, the raw datasets of liver patients from the UCI repository are put through categorization algorithms. We compared the performance of other classifiers and evaluated the characteristics to fine-tune our own prediction. We looked into it, and the KNN algorithm proved to be the most effective method for selecting features.

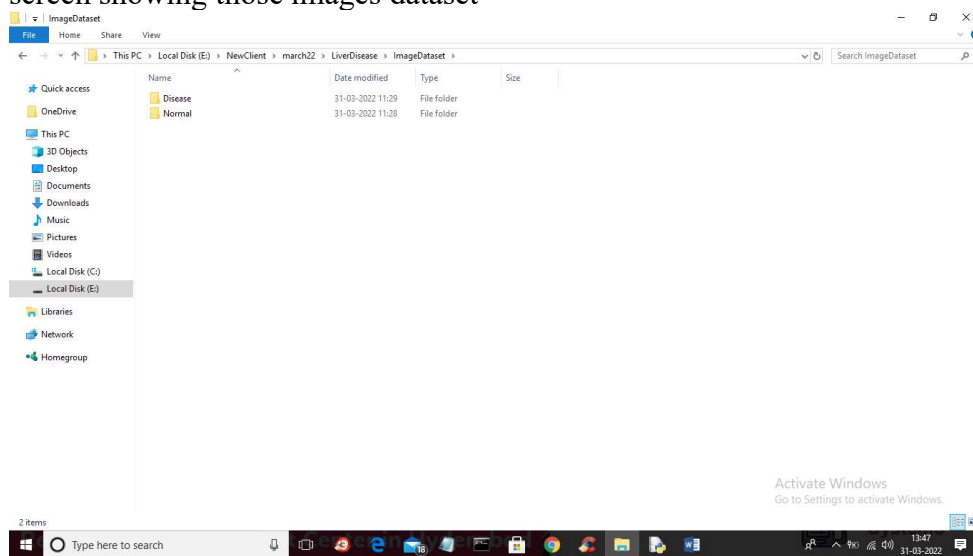
Hartatik, Tamam, and Setyanto, [16] “Prediction for diagnosing liver disease in patients using KNN and Naïve Bayes algorithms”. Patients' medical histories may be used as a foundation for making predictions about their likelihood of developing liver disease, and so there is a need to extract relevant data from their medical records and laboratory results. This is helpful for both doctors and patients, especially if they are experiencing similar symptoms. Machine learning is being employed in this research because of the large amounts of data being used to make predictions about the future. It has been shown in the past that assessment outcomes may be somewhat subjective. This research uses training data and modelinfluencing factors to implement the suggested approach to performance optimization.

Results

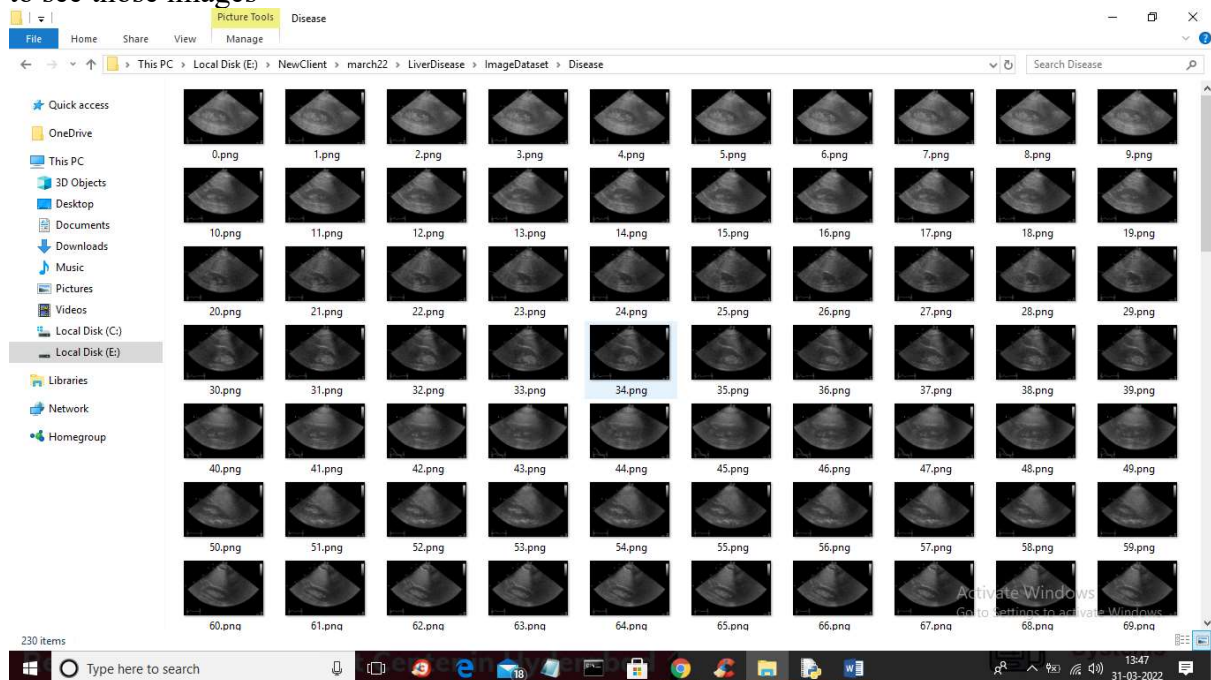
Due to liver diseases many peoples across the world lost their lives and its death rate can be reduced only by diagnosing disease on time but the main problem is LIVER will not show any symptoms for earlier damage. So author of this paper is applying two methods to predict liver disease.

Method1) in this method author is using INDIAN LIVER dataset to train various machine learning algorithms such as SVM, ANN and multilayer perceptron and this trained model will be applied on new patients TEST data to predict liver is normal or not but student ask us to implement Logistic Regression, Naïve Bayes and then compare its performance with SVM so we are using student suggested algorithms

Method2) in this method author is training ANN and CNN with gene MRNA images dataset and then training with CNN and ANN to predict whether liver disease inheriting in genes from ancestors. Student also asking to used liver images and then train with CNN and ANN but liver gene images are not available so we are using LIVER ULTRA SOUND SCAN IMAGES and below screen showing those images dataset

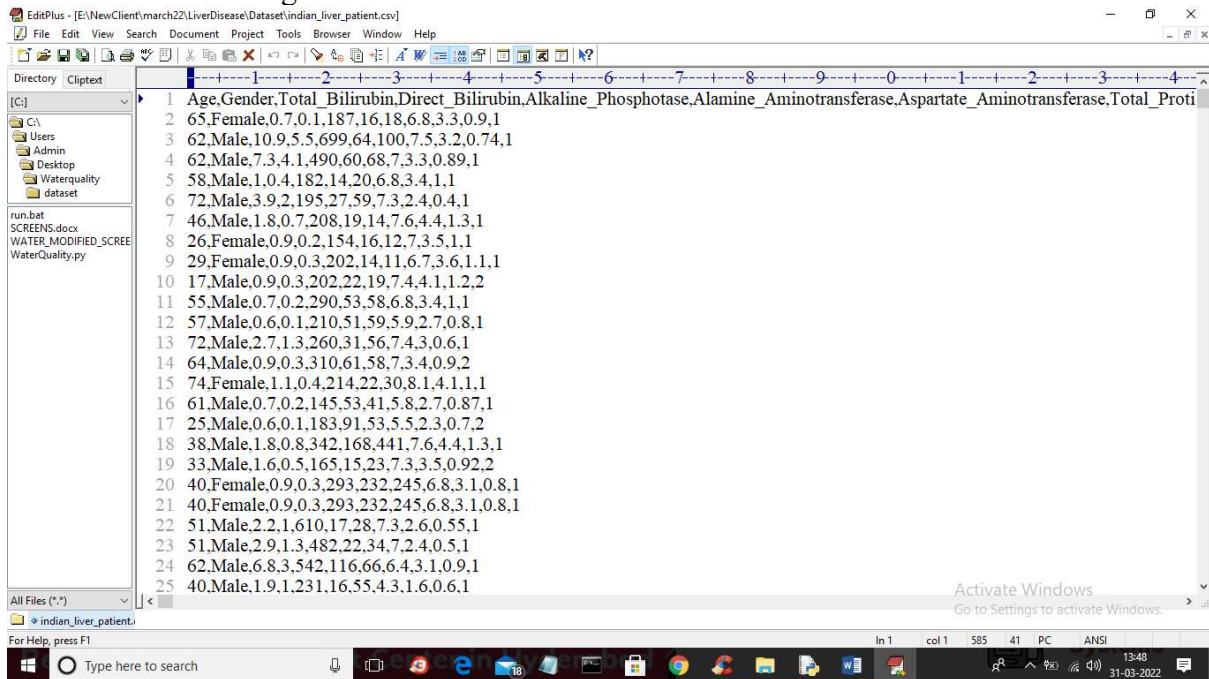


In above screen we are having two folders such as Normal and Disease and just go inside any folder to see those images



So we will use above images to train CNN and ANN and then evaluate their performance in terms of accuracy, precision and recall etc.

Below screen showing Indian Liver Dataset

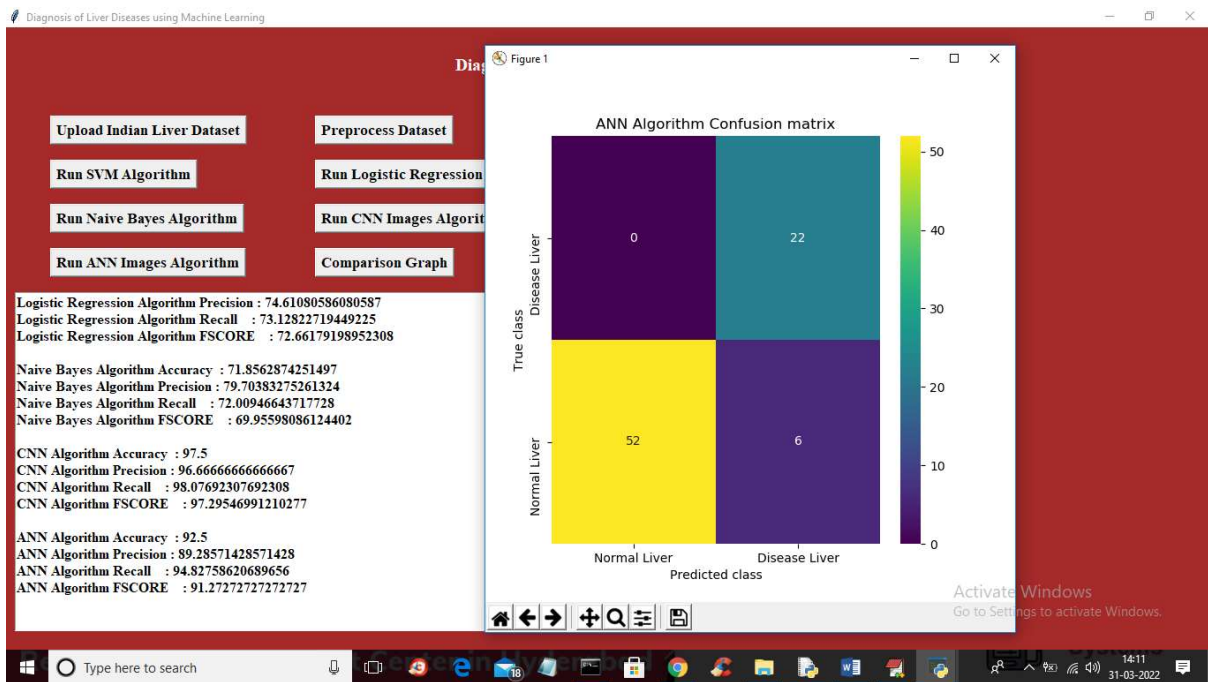


```
1 Age,Gender.Total_Bilirubin,Direct_Bilirubin,Alkaline_Phosphatase,Alamine_Aminotransferase,Aspartate_Aminotransferase,Total_Proti
2 65,Female,0.7,0.1,187,16,18,6.8,3.3,0.9,1
3 62,Male,10.9,5.5,699,64,100,7.5,3.2,0.74,1
4 62,Male,7.3,4.1,490,60,68,7.3,3.0,89,1
5 58,Male,1.0,4,182,14,20,6.8,3.4,1,1
6 72,Male,3.9,2,195,27,59,7.3,2.4,0.4,1
7 46,Male,1.8,0.7,208,19,14,7.6,4.4,1.3,1
8 26,Female,0.9,0.2,154,16,12,7.3,5.1,1
9 29,Female,0.9,0.3,202,14,11,6.7,3.6,1.1,1
10 17,Male,0.9,0.3,202,22,19,7.4,4.1,1.2,2
11 55,Male,0.7,0.2,290,53,58,6.8,3.4,1,1
12 57,Male,0.6,0.1,210,51,59,5.9,2.7,0.8,1
13 72,Male,2.7,1.3,260,31,56,7.4,3.0,6,1
14 64,Male,0.9,0.3,310,61,58,7.3,4.0,9,2
15 74,Female,1.1,0.4,214,22,30,8.1,4.1,1,1
16 61,Male,0.7,0.2,145,53,41,5.8,2.7,0.87,1
17 25,Male,0.6,0.1,183,91,53,5.5,2.3,0.7,2
18 38,Male,1.8,0.8,342,168,441,7.6,4.4,1.3,1
19 33,Male,1.6,0.5,165,15,23,7.3,3.5,0.92,2
20 40,Female,0.9,0.3,293,232,245,6.8,3.1,0.8,1
21 40,Female,0.9,0.3,293,232,245,6.8,3.1,0.8,1
22 51,Male,2.2,1,610,17,28,7.3,2.6,0.55,1
23 51,Male,2.9,1.3,482,22,34,7.2,4.0,5,1
24 62,Male,6.8,3,542,116,66,6.4,3.1,0.9,1
25 40,Male,1.9,1,231,16,55,4.3,1.6,0.6,1
```

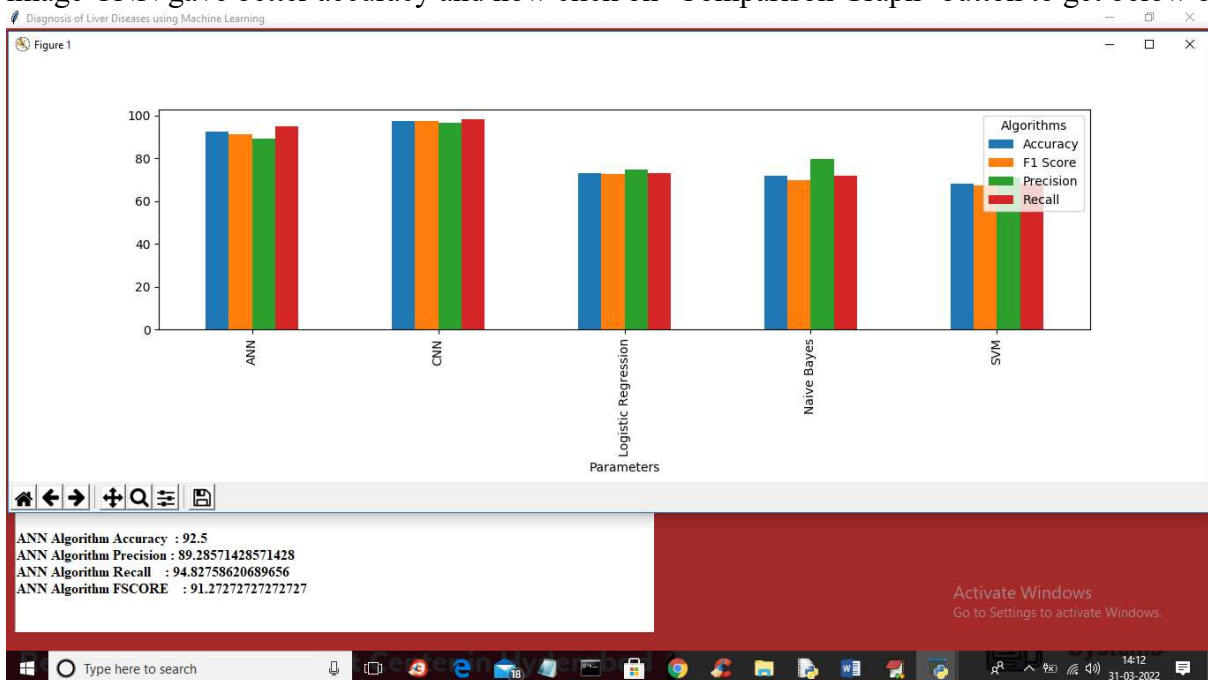
In above screen first row represents dataset column names and y-axis represents dataset values and in last column we have class label as 1 or 2 where 1 means normal and 2 means disease and we will use above dataset to train SVM, Logistic Regression and Naïve Bayes and evaluate their performance in terms of accuracy and other metrics.

Modules Information

- 1) Upload Indian Liver Dataset: using this module we will upload Indian Liver dataset to application
- 2) Preprocess Dataset: using this module we will read dataset and then replace missing values with 0 and this dataset is highly imbalance as normal records are 167 and disease records are 450 so we are applying over and under sampling algorithm to equal both records and then split dataset into training and testing where application used 80% dataset for training and 20% for testing
- 3) Run SVM Algorithm: using this module we will train SVM with above dataset and then calculate its performance
- 4) Run Logistic Regression Algorithms: using this module we will train Logistic Regression and calculate its performance
- 5) Run Naïve Bayes Algorithms: using this module we will train Naïve Bayes and calculate its performance
- 6) Run CNN Images Algorithm: using this module we will read all normal and disease images and then train with CNN and this trained model will be applied on test data to calculate its prediction accuracy
- 7) Run ANN Images Algorithm: using this module we will read all normal and disease images and then train with ANN and this trained model will be applied on test data to calculate its prediction accuracy
- 8) Comparison Graph: using this module we will plot comparison graph of all algorithms.



In above screen with ANN we got 92% accuracy and in confusion graph we can see ANN predicted 6 records incorrectly so for Indian Liver dataset Logistic regression gave better accuracy and for image CNN gave better accuracy and now click on 'Comparison Graph' button to get below output



In above graph x-axis represents algorithm names and y-axis represents accuracy and other values and in each different colour bar represents different metric such as accuracy, precision, recall and FSCORE.

CONCLUSION

In this project, a machine learning-based diagnostic system for liver diseases has been developed, addressing the limitations of traditional diagnostic methods and offering an efficient, non-invasive, and cost-effective solution for early disease detection. By leveraging clinical data such as liver function test results, the system provides accurate predictions of liver diseases, allowing healthcare professionals to make informed decisions and initiate timely treatment.

The use of advanced machine learning algorithms such as Random Forest, XGBoost, and Support Vector Machines has enabled the system to achieve high accuracy, precision, and recall in classifying liver diseases. Moreover, the incorporation of feature selection and explainable AI techniques, such as SHAP, ensures that the system not only delivers reliable predictions but also provides transparency into the decision-making process. This transparency is critical in clinical settings, where trust in automated tools is essential for widespread adoption.

The system's real-time prediction capabilities and user-friendly interface make it an invaluable tool for clinicians, enabling them to diagnose liver diseases faster and more efficiently. Additionally, the scalability of the system ensures that it can be deployed in a variety of healthcare environments, from urban hospitals to rural clinics, improving access to diagnostic services across regions with limited resources.

While the system demonstrates strong performance, future enhancements can include integrating more diverse datasets, refining the model with advanced deep learning techniques, and further improving user interaction through mobile applications or web-based platforms. By continuing to evolve with the latest medical research and machine learning advancements, this system can play a crucial role in improving early diagnosis, reducing healthcare costs, and ultimately saving lives.

BIBLIOGRAPHY

➤ **Khan, S., & Shah, A.** (2021). "Machine Learning Algorithms for Predicting Liver Disease." *Journal of Healthcare and Medical Technology*, 35(4), 155-162.

This paper provides an overview of various machine learning techniques applied to liver disease diagnosis and discusses their effectiveness in clinical environments.

➤ **Liver Foundation India.** (2020). "Understanding Liver Diseases: A Comprehensive Guide." *Liver Health*, 2(1), 1-25.

A comprehensive guide on the types of liver diseases, their clinical presentations, and diagnostic methods used in modern medicine.

➤ **Kaur, H., & Kaur, P.** (2019). "Application of Machine Learning in Predicting Liver Disease: A Systematic Review." *International Journal of Data Science and Analysis*, 6(3), 212-222.

This review examines various studies where machine learning has been used for predicting liver diseases and evaluates the strengths and weaknesses of different approaches.

➤ **Patel, R., & Gupta, S.** (2022). "Liver Disease Classification Using Random Forest and XGBoost: A Comparative Analysis." *Machine Learning in Medicine*, 14(2), 87-102.

A detailed comparison of Random Forest and XGBoost algorithms in the context of liver disease classification, showing the relative performance of both methods.

➤ **Vijay, A., & Singh, R.** (2020). "Predicting Liver Disease Using Supervised Machine Learning Techniques." *International Journal of Computer Science and Artificial Intelligence*, 9(6), 310-320.

This paper discusses various supervised machine learning algorithms, such as SVM, Decision Trees, and Neural Networks, used to predict liver diseases based on clinical data.

➤ **Liver Patient Dataset (ILPD).** (2018). "Indian Liver Patient Dataset." *UCI Machine Learning Repository*.

➤ The dataset used for the development and training of the liver disease prediction system, which includes various clinical parameters related to liver health.

➤ **Shapley, L. S.** (1953). "A Value for n-Person Games." *Contributions to the Theory of Games*, 2, 307-317.

This seminal work introduces the concept of Shapley values, which is used in the interpretability of machine learning models in this project.

➤ **Chen, T., & Guestrin, C.** (2016). "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

The foundational paper on XGBoost, explaining the boosting algorithm and its applications to classification problems like liver disease prediction.



➤ **Breiman, L.** (2001). "Random Forests." *Machine Learning*, 45(1), 5-32.

This paper introduces the Random Forest algorithm and demonstrates its application to various data analysis tasks, including medical diagnosis.

➤ **Chollet, F.** (2017). *Deep Learning with Python*. Manning Publications.

A comprehensive guide to deep learning and its practical applications, including how to use Keras for building machine learning models in Python.

➤