iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83039

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

A Comparative Review of Machine Learning Methods for Early Detection of Liver Disease

Sanjana Narkhede[1], Harshada Sonawane[2], Dipali Mali[3], Dr. P. S. Sanjekar[4]
[1, 2, 3]*CSE (Data Science), R C Patel Institute of Technology* Shirpur, Maharashtra
[4]*Professor, R C Patel Institute of Technology* Shirpur, Maharashtra

Abstract: ***Liver disease is like one of the major health issues affecting millions of people around the world, and it can be pretty serious in a lot of cases too. The liver performs important functions such as detoxification, digestion support, metabolism regulation, protein synthesis, and storage of nutrients. Damage to the liver can lead to severe health complications and even death if not detected early. Most liver diseases do not show symptoms during the early stages, which makes diagnosis dif- ficult. Traditional diagnostic methods depend on blood tests, imaging systems, and expert medical analysis, which may not always be available in rural and low-resource healthcare envi- ronments. Therefore, an intelligent automated system is needed for fast and accurate liver disease prediction.The proposed system applies several preprocessing techniques including missing value handling, feature scaling, feature engineering, and class balancing using SMOTETomek [8], [9]. Six machine learning algorithms are compared: Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), and Random Forest [7]. The performance of all algorithms is evaluated using Accuracy, Precision, Recall, F1-Score, and ROC- AUC Score. Experimental analysis shows that Random Forest gets the best overall performance , with higher accuracy, better recall , stronger F1 score and an improved ROC-AUC value, kinda compared to the other algorithms. [2], [3]. Random Forest seems to do better because it blends many decision trees reduces overfitting , handles messy or noisy data in an efficient way and it works really well with nonlinear patterns that show up in medical datasets. It's almost like a committee effect, though not exactly [7].***
Index Terms: Liver Disease Prediction, Machine Learning, Random Forest, SMOTETomek, Feature Engineering, Flask, Clinical Decision Support, Comparative Analysis.

I. INTRODUCTION

Liver problems are, in many ways, among the most serious health issues around, right now. The World Health Organiza- tion (WHO) notes that liver diseases cause about two million deaths each year across the world. The liver is the biggest internal organ in the human body, and it basically runs a long list of crucial tasks, doing over 500 vital functions, like filtering toxins from the blood, making bile for digestion , storing energy, and also crafting proteins that help blood clot.

Since the liver is busy with so many essential jobs, even small or moderate injury to it can badly affect a person's general wellbeing. One of the biggest problems with liver disease is that it often just does not show obvious symptoms in the early stages, so a lot of damage can quietly build up. A person can end up with serious liver damage without actually feeling sick, because the liver is a resilient organ and it can still keep functioning even when a large portion of it is already damaged. Then later, when symptoms finally show up, like yellowing of the skin (jaundice), bloating or swelling in the abdomen, vomiting blood, or extreme fatigue, the condition may already be advanced and kind of hard to treat. This is why early diagnosis and prediction of liver disease is critically important. Right now, doctors usually diagnose liver disease using a mix of blood tests, ultrasound scans, CT scans, and sometimes even a liver biopsy. But these approaches are costly, they can take a long time, and they also need specialized medical expertise. In rural or remote areas, like parts of India, access to this kind of facilities, is often very limited [12]. Patients often travel long distances, to hospitals and by the time they get a diagnosis, the illness has already progressed quite far. So there is a clear, urgent need for an intelligent but affordable automated system, that can analyze rather simple blood test results and quickly estimate whether someone might be facing liver disease risk. In the last few years , Artificial Intelligence (AI) and Machine Learning (ML) have made tremendous strides in the medical field [2], [3], [4]. Machine learning algorithms can pick up patterns from a big pile of patient data , then sort of use those patterns to make predictions about new patients. A couple of studies have already shown that ML models trained on patient blood test records can help forecast whether someone has certain diseases , like diabetes, heart disease, kidney disease, and liver disease—usually with good accuracy , even if the process looks a bit indirect at first [2], [5]. These systems can help doctors make faster, more precise choices, especially in areas where medical experts are not readily available, or kind of hard to reach.

©IJRASET: All Rights are Reserved

In this study, we bring a structured comparison of six well known machine learning algorithms aimed at early liver disease prediction. The models contrasted include Logistic Re- gression, Decision Tree, K-Nearest Neighbors, Naive Bayes, Support Vector Machine, and Random Forest [2], [3], [4]. The study covers a full end- to-end pipeline, starting from data preprocessing, moving into feature engineering, then handling the class imbalance with the SMOTETomek [8], [9], model, and finally doing model training plus performance evaluation. After a careful comparison, Random Forest shows the strongest results [7] so its selected and then deployed as a Flask web app [11] that healthcare workers can actually use in day to day clinical settings.

A. *Background*

Machine learning based disease prediction has kind of gotten a lot of attention in recent years, more and more [2], [3], [5]. The Indian Liver Patient Dataset , ILPD, is a well- known benchmark dataset for liver disease investigations [1], [12]. It was collected from patients in the Andhra Pradesh region, India and it can be found publicly via the UCI Machine Learning Repository [1]. A bunch of earlier papers used the same dataset, and they tried different kinds of machine learn- ing strategies [2], [3], [4]. Overall, many results say ensemble methods, for example Random Forest , usually do better than simpler baselines like Logistic Regression or Decision Trees, [7], especially when the dataset is treated properly first, with careful preprocessing steps to manage missing values and class imbalance [8], [9].

B. *Motivation and Contribution*

The main reason for this work is to figure out which machine learning method is most useful for early liver disease prediction, and then to package it as an easy, web based tool for clinical teams to use [11]. In many older papers, researchers would just try one or two approaches on the same dataset ,but without really taking care of class imbalance, or doing any thoughtful feature engineering, so the resulting accuracy stayed quite constrained [2], [3], [4]. In contrast, this study tries to close these gaps by evaluating six algorithms, all under essentially the same setup ,using SMOTETomek for balancing the classes [8], [9], and also crafting derived features that are clinically sensible in order to boost performance. The main contributions of this paper are:

1) A complete and reproducible comparison of six machine learning algorithms under identical experimental condi- tions.
2) Application of SMOTETomek hybrid resampling to han- dle the class imbalance problem in the dataset [8], [9].
3) Feature engineering with medically meaningful de- rived variables such as the AST/ALT ratio and log- transformed bilirubin values.
4) Deployment of the best-performing model as an accessible Flask web application [11].
5) A detailed analysis explaining why Random Forest out- performs all other algorithms for this specific prediction task [7].

II. LITERATURE REVIEW

One of the most important advancements in the health- care analytics is the usage of Machine Learning (ML) and Deep Learning (DL) techniques for Early Prediction of Liver Disorders. Liver diseases like hepatitis, cirrhosis, fatty liver disease and liver cancer are often asymptomatic, so early diagnosis is essential [2], [5]. Current testing guidelines are often slow, expensive, and less available in remote areas [12]. To combat these issues, researchers use the patient clinical data and laboratory parameters for the automated prediction [3], [4]. Many ML algorithms such as Random Forest, Support Vector Machine, Logistic Regression, and Na¨ıve Bayes, are commonly employed for classification of liver disease [2], [3], [6]. These models aid in finding hidden patterns in medical data and help to improve accuracy [10].

A. *Comparative Machine Learning Approaches for Liver Dis- ease Prediction*

Researchers have applied several machine learning algo- rithms for liver disease prediction using clinical parameters such as bilirubin levels, enzyme values, albumin, proteins, age, and gender [2], [3], [12]. The most commonly used algorithms include Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Naive Bayes, and Random Forest [2], [4]. Traditional machine learning models are capable of identifying disease patterns from patient medical records and help doctors in early diagnosis [5], [10]. Researchers observed that ensemble learning methods provide better prediction capability than individual classifiers because they combine the strengths of multiple models and reduce the effect of individual prediction errors [7]. Most studies used the Indian Liver Patient Dataset (ILPD) because it contains real patient clinical records suitable for healthcare prediction systems [1], [12]. Researchers also emphasized the importance of preprocessing techniques such as feature scaling, feature engineering, normalization, and handling missing values to improve prediction performance [2], [3]. The studies con- cluded that machine learning-based healthcare systems can assist doctors by providing faster and more reliable disease prediction results [5], [6].

©IJRASET: All Rights are Reserved

B. Feature Selection and Correlation Analysis in Liver Dis- ease Prediction

Feature selection and preprocessing play an important role in liver disease prediction systems [2], [3]. Researchers ob- served that many medical parameters in liver disease datasets are highly correlated with each other [12]. For example, Total Bilirubin and Direct Bilirubin have strong relationships be- cause both are related to liver functioning [12]. Studies showed that removing unnecessary features and selecting important attributes improve model accuracy and reduce computational complexity [2], [10]. Researchers used techniques such as cor- relation heatmaps, feature importance analysis, and statistical preprocessing to identify meaningful clinical parameters [3], [7]. Feature engineering techniques were also applied to im- prove machine learning performance [4]. Researchers created ratio-based features and transformed skewed data distributions to make patterns more understandable for machine learning models [2], [3].

C. Hybrid and Ensemble Learning Models

Hybrid and ensemble learning models have become popular in healthcare prediction systems because they improve model stability and prediction capability [6]. Researchers combined algorithms such as KNN and Random Forest to create hybrid systems that utilize the strengths of both algorithms [4]. En- semble learning methods work by combining predictions from multiple models instead of depending on a single classifier [7]. This reduces overfitting and improves generalization perfor- mance on unseen patient data [7]. Hybrid models were found to handle noisy and nonlinear medical datasets more efficiently compared to traditional algorithms [4], [6]. Researchers also observed that combining classifiers improves classification stability and reduces prediction variance [6]. These studies demonstrated that ensemble approaches are highly suitable for medical diagnosis systems because healthcare datasets often contain complex and imbalanced data [8], [9].

D. Ensemble Learning using Random Forest

Random Forest is one of the most effective ensemble learning algorithms used in healthcare prediction systems [7]. It combines multiple decision trees trained on random subsets of data and features [7]. Researchers observed that Random Forest reduces overfitting problems commonly seen in single Decision Tree models [7]. The algorithm provides stable predictions because errors from individual trees are minimized through majority voting [7]. Random Forest han- dles nonlinear relationships, noisy data, correlated features, and missing values efficiently [2], [7]. It also works well on imbalanced datasets and does not require extensive prepro- cessing compared to some other algorithms [7]. Because of these advantages, Random Forest became one of the most pre- ferred algorithms for liver disease prediction systems [2], [3]. Researchers consistently observed better prediction stability and reliability using Random Forest compared to traditional classifiers [2], [3], [6].

E. Class Imbalance Handling Techniques

Class imbalance is one of the major problems in healthcare datasets because diseased patient records are often much higher than healthy patient records [8]. Machine learning models trained on imbalanced datasets become biased toward the majority class and fail to correctly predict minority class samples [9]. To solve this problem, researchers introduced balancing techniques such as SMOTE and SMOTETomek [8], [9]. SMOTE generates synthetic minority class samples to balance the dataset, while SMOTETomek additionally removes noisy and overlapping samples [8], [9]. Studies showed that balancing techniques improve recall, classification capability, and overall model learning [9]. These methods are widely used in liver disease prediction systems to improve minority class prediction performance [2], [3].

F. Support Vector Machine and Probabilistic Models

Support Vector Machine (SVM) and Naive Bayes are widely used machine learning approaches in healthcare prediction systems [5], [10]. For SVM, the idea is to search for a best possible hyperplane that can separate disease from non-disease cases [10]. Many works report that SVM tends to do well on medical tables with many features, mainly because kernel functions help model the more tangled relations among pre- dictors [5]. Still, SVM is a bit picky, it needs careful hyperpa- rameter tuning along with feature normalization/scaling [10]. If that step is ignored even somewhat, the overall performance can drop quite a lot [2].

Naive Bayes, by contrast, is a probabilistic method rooted in Bayes theorem [10]. It works under the notion that each feature behaves independently from the others [10]. In practice, re- searchers argue this independence premise can be a poor fit for liver disease datasets , because several clinical measurements are strongly related or intertwined [12]. Even though Naive Bayes is quick and computationally light, its predictive power gets restricted once it is used on more complex healthcare problems [2] [3].

TABLE I
COMPARISON OF MACHINE LEARNING ALGORITHMS FOR LIVER DISEASE PREDICTION

Algorithm	**Advantages**	**Limitations**	**Suitability**
Logistic Regression	Simple and fast	Poor for nonlinear data	Moderate
KNN	Easy and accu- rate	Slow for large data	Good
SVM	Good for com- plex data	Needs tuning	Good
Decision Tree	Easy to under- stand	Overfitting problem	Moderate
Naive Bayes	Fast and lightweight	Independent feature assumption	Low
Random For- est	Stable and accu- rate	High training time	Very High
Hybrid KNN + RF	Better prediction	More computation	High

III. METHODOLOGY

In this section, we describe the methodology used to develop the proposed system for liver disease prediction by applying machine learning ideas. Basically the system is aimed at guessing whether a patient has liver disease, using clinical pa- rameters like bilirubin amounts, liver enzyme values, proteins, albumin measures, plus age and gender. The overall framework is like a structured pipeline with steps that include dataset collection, data preprocessing , feature engineering, dealing with class imbalance, training the machine learning models, checking performance, and then deployment [2] [3]. Several

TABLE II
COMPARISON OF EXISTING LIVER DISEASE PREDICTION SYSTEMS

Sr. No	Author / Year	Method Used	Dataset	Accuracy	Key Findings
1	Frontiers in Physiol- ogy (2025)	Logistic Regression, SVM	Liver Patient Dataset	75%	Logistic Regression performed better among ba- sic models but accuracy remained moderate due to limited feature optimization.
2	Frontiers in Medicine (2025)	Decision Tree (Deci- sion Stump)	Clinical Liver Data	70.67%	Decision Tree-based approaches showed lower accuracy due to overfitting and poor generaliza- tion.
3	MDPI Computers (2023)	Voting Classifier (En- semble)	Medical Dataset	80.1%	Ensemble techniques improved prediction per- formance but still limited around 80% accuracy.
4	arXiv Research (2025)	Logistic Regression	MASLD Dataset	77.6%	Real-world datasets lead to moderate accuracy due to imbalance and noise in data.
5	Survey Paper (Vari- ous Studies)	Multiple ML Algo- rithms	ILPD Dataset	~70–80%	Most traditional machine learning models achieve accuracy below 80%, highlighting need for optimization.
6	Proposed System (Your Project)	Random Forest (Tuned)	Liver Disease Dataset	82.5%	Improved performance using feature engi- neering, SMOTETomek balancing, and hy- perparameter tuning.

classification methods were tried and then laid side by side, but Random Forest ended up as the final choice, mainly because it shows stronger stability and a solid prediction capability [7].

A. *Overview of the Proposed Methodology*

The suggested liver disease prediction system does its work by automatically taking in patient clinical data, then estimating how likely liver disease is using machine learning algorithms [2], [10]. This framework uses a modular design, so every step does a particular job in order to raise prediction quality and make the results more dependable, a bit more robust. At the beginning, the dataset is preprocessed to fix irregularities and to set up the information so it can be fed to machine learning [3]. After that, feature engineering is carried out, to extract medical features that are more meaningful which in turn help the model get better at detecting disease, maybe more sensitive too. Because the dataset is not evenly distributed between classes , SMOTETomek resampling is applied in order to balance the groups and to lessen any built-in prediction skew [8], [9]. Next, several machine learning models are trained and checked, then the top model is chosen for the actual deployment stage [7], [11]. Overall, the methodology is meant to deliver accurate, steady, and fast liver disease prediction, something that should fit real world healthcare needs without too much trouble.

B. *Input Data Acquisition and Preprocessing*

The proposed system uses the Indian Liver Patient Dataset (ILPD) collected from the UCI Machine Learning Repository [1]. The dataset contains clinical records of liver disease patients and healthy individuals from Andhra Pradesh [1], [12], India. The description of all dataset features used in the proposed system is presented in Table III. It includes important medical parameters such as Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Aspartate Aminotransferase (AST), Alamine Aminotransferase (ALT), Albumin, Total Pro- teins, and Albumin-Globulin Ratio [1], [12].

Before model training, several preprocessing operations are performed to improve dataset quality and learning efficiency [2], [10]. Missing values present in the Albumin and Globulin Ratio feature are replaced using the median value to maintain consistency while reducing the effect of extreme outliers. The Gender feature, which contains categorical values, is converted into numerical form using label encoding [10]. Outlier han- dling is performed using winsorization to reduce the effect of abnormal medical values while preserving important patient records. Logarithmic transformation is also applied to highly skewed features such as Total Bilirubin and Direct Bilirubin to normalize their distribution and improve machine learning performance [2], [3]. These preprocessing steps ensure that the dataset becomes clean, balanced, and suitable for effective model training [10].

TABLE III
FEATURES OF THE INDIAN LIVER PATIENT DATASET

Feature	**Type**	**Description**
Age	Numeric	Patient age in years (range: 4 to 90)
Gender	Categorical	Patient gender: Male or Female
Total Bilirubin	Numeric	Total bilirubin level in blood (mg/dL)
Direct Bilirubin	Numeric	Direct bilirubin level in blood (mg/dL)
Alkaline Phosphatase	Numeric	ALP enzyme level in blood (IU/L)
Alamine Aminotransferase (ALT)	Numeric	ALT enzyme level (IU/L)
Aspartate Aminotransferase (AST)	Numeric	AST enzyme level (IU/L)
Total Proteins	Numeric	Total protein level in blood (g/dL)
Albumin	Numeric	Albumin protein level in blood (g/dL)
Albumin and Globulin Ratio	Numeric	Ratio of albumin to globulin
Dataset (Target)	Binary	1 = Liver Disease, 0 = Healthy

C. Feature Engineering and Feature Representation

Feature engineering is performed to improve the learning capability of machine learning algorithms by creating meaningful features from existing medical parameters [2], [3]. In the proposed system, several medically important features are generated using domain knowledge related to liver functioning. The transformations and newly created features used in the proposed system are shown in Table IV. The AST/ALT ratio is created because it is clinically useful for identifying different types of liver damage such as cirrhosis and alcoholic liver disease. Log-transformed bilirubin features are also generated to reduce skewness and improve distribution consistency [3]. Additionally, patient age values are grouped into meaningful categories such as Child, Young Adult, Middle Age, and Senior Citizen to capture nonlinear disease risk patterns associated with age [2]. Feature engineering helps the machine learning models learn hidden relationships among clinical parameters and improves overall prediction accuracy and stability [7], [10].

TABLE IV
FEATURE ENGINEERING APPLIED TO LIVER DISEASE DATASET

Original Feature	Transformation Applied	New Feature Created
Total Bilirubin	log(1 + x)	Log Total Bilirubin
Direct Bilirubin	log(1 + x)	Log Direct Bilirubin
AST + ALT	AST ÷ ALT	AST/ALT Ratio
Age	Interval binning	Age Group

D. Class Imbalance Handling using SMOTETomek

The ILPD dataset contains a significant imbalance between liver disease patients and healthy individuals [1], [12]. Ma- chine learning models trained on imbalanced datasets tend to become biased toward the majority class, resulting in poor prediction capability for minority class samples [8].

To get around this problem, we apply the SMOTETomek strategy to the training dataset [9]. In short, SMOTE (Syn- thetic Minority Oversampling Technique) manufactures new minority-class samples through nearest neighbor interpolation, and this sort of balancing is meant to help the data set not get skewed [8]. After that, Tomek Links are used for deleting uncertain and intermingled observations sitting close to the decision edges, basically near the class limits [9]. Together this hybrid resampling method boosts the overall dataset steadiness, lowers prediction bias somewhat, and improves the capacity of machine learning models to correctly distinguish between liver disease cases and healthy subjects [8], [9].

E. Machine Learning Model Training

The dataset is split into training and testing sets with strati- fied random sampling, to keep a similar class balance in each subset [10]. Then several machine learning models are fitted and kind of cross examined for liver disease detection [2], [3]. Among them there are Logistic Regression , a Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes , Support Vector Machine (SVM), and also Random Forest [2], [4]. Feature scaling is applied using StandardScaler to normalize feature distributions and improve model learning efficiency, especially for algorithms such as KNN and SVM that are sensitive to feature magnitude differences [10]. Each algorithm is trained using the same preprocessing pipeline and evaluated using standard performance metrics [2], [3].

Random Forest uses multiple decision trees trained on random subsets of data and features, which helps reduce over- fitting and improve prediction stability [7]. Due to its ability to handle nonlinear medical data, correlated features, and noisy datasets efficiently, Random Forest produced better prediction performance compared to the other machine learning models [7].

F. Model Evaluation and Validation

The trained machine learning models are evaluated using several quantitative performance metrics to measure prediction capability and reliability [2], [10]. The evaluation metrics used in this study include Accuracy, Precision, Recall, F1-Score, and ROC-AUC Score. Accuracy measures the overall percentage of correct pre- dictions made by the model. Precision evaluates how many patients predicted as liver disease cases are actually correct. Recall shows how well the system can correctly flag liver dis- ease patients, and this is oddly critical in medical work because skipping an actual disease case might be, uh, dangerous [5].

The F1-Score works as a kind of compromise evaluation, since it mixes precision together with recall, so it does not over focus on only one side. Meanwhile ROC-AUC tells us how capable the model is at separating liver disease versus healthy people over a range of decision thresholds [2], [3].

A confusion matrix analysis is also done, to look at True Positives, True Negatives, plus False Positives and False Neg- atives [10]. Comparative performance analysis of all machine learning algorithms is carried out, and Random Forest is selected as the final model because of its superior stability, better generalization capability, and improved overall predic- tion performance [7].

G. System Deployment and Prediction Interface

The final Random Forest model is deployed using the Flask web framework to provide real-time liver disease prediction through a user-friendly web interface [11]. The deployed system allows users to enter clinical parameters such as bilirubin levels, liver enzyme values, proteins, and albumin measurements. After submitting the data, the trained machine learning model processes the input and predicts whether the patient is likely to have liver disease.

The deployment framework provides fast prediction, easy accessibility, and efficient interaction with the prediction sys- tem [11]. The proposed web-based healthcare application can assist doctors and users in obtaining preliminary liver dis- ease predictions quickly and effectively. Fig. 1 illustrates the workflow diagram of the proposed Liver Disease Prediction methodology.

IV. ALGORITHM

The gradual execution of the proposed liver disease prediction system is represented by Algorithm-1. The

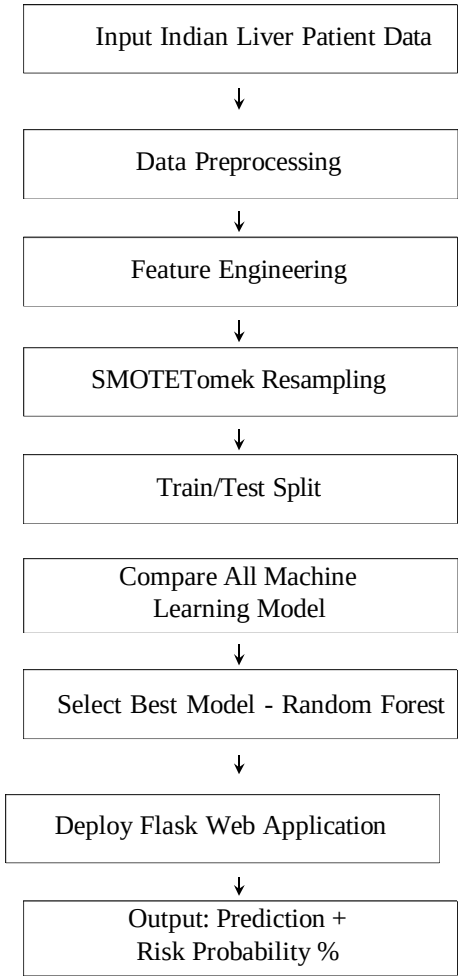


Fig. 1. Workflow of Liver Disease Prediction

algorithm follows the same sequence as the workflow diagram and the methodology discussed in the previous sections.

Algorithm 1: Liver Disease Prediction Pipeline

1: Load Indian Liver Patient Data (583 patient records) [1], [12]
2: Impute missing values using column median 3: Encode Gender and recode target label [10] 4: Compute AST/ALT ratio as new feature
5: Apply log transformation to Bilirubin values [2]
6: Derive Age Group using interval binning
7: Apply SMOTETomek to balance the dataset [8], [9]
8: Split into 80% train and 20% test (stratified) [10]
9: Normalize features using StandardScaler [10]
10: Compare six Machine Learning ALgorithms model [2], [3], [4]
11: Evaluate using Accuracy, F1-Score, ROC-AUC [2]
12: Select Random Forest as best model (AUC = 91.2%) [7]
13: Deploy using Flask web application [11]
14: Accept new patient input from browser
15: Apply feature engineering and scaling on input
16: Predict using Random Forest model [7]

V. RESULTS AND ANALYSIS

This section presents the experimental results and perfor- mance analysis of the proposed liver disease prediction system.
The effectiveness of the system is evaluated using multiple machine learning algorithms and standard performance metrics [2], [10]. The analysis focuses on prediction accuracy, model stability, class imbalance handling, feature effectiveness, and overall prediction capability for liver disease detection.
The proposed system was trained and tested using the Indian Liver Patient Dataset (ILPD) [1], [12]. After preprocessing, feature engineering, class balancing using SMOTETomek [8], [9], and feature scaling, multiple machine learning models were trained and evaluated [10]. The final prediction results demonstrate that the Random Forest algorithm provides bet- ter overall performance and stability compared to the other classifiers [7].

A. Comparative Analysis of Liver Disease Prediction Models

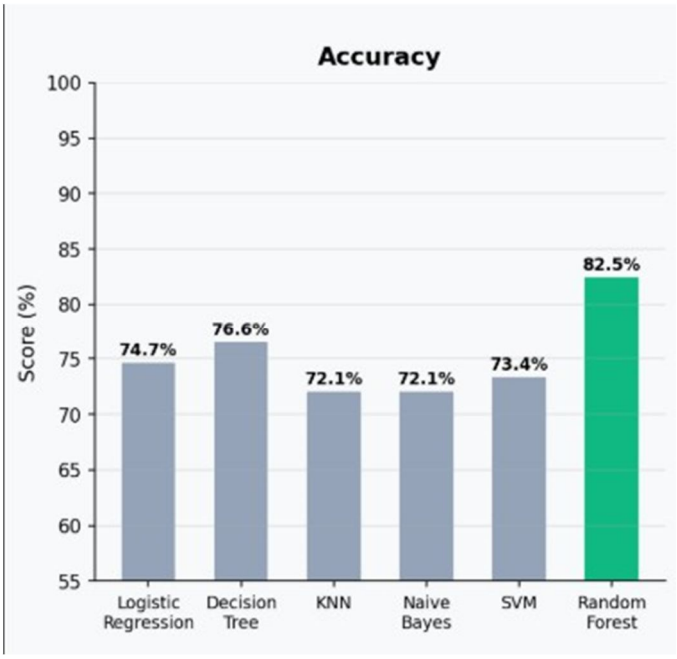


Fig. 2. Algorithm Comparison Performance: Accuracy

The performance of various machine learning algorithms for liver disease prediction was evaluated using Accuracy, F1- Score, and ROC-AUC metrics [2], [3]. The comparative anal- ysis included Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), and Random Forest classifiers [2], [4]. Among all the models, the Random Forest classifier demonstrated superior performance with the highest Accuracy (82.5%), F1-Score (80.6%), and ROC-AUC value (93.8%) [7]. These results indicate that the Random Forest model achieved better classifi- cation capability and maintained an effective balance between precision and recall compared to the other algorithms. Further- more, the ROC curve of the Random Forest classifier shows a curve close to the top-left corner, indicating high sensitivity and strong discriminative ability in distinguishing liver disease patients from healthy individuals [7]. The obtained AUC value of 0.9383 further confirms the robustness and reliability of

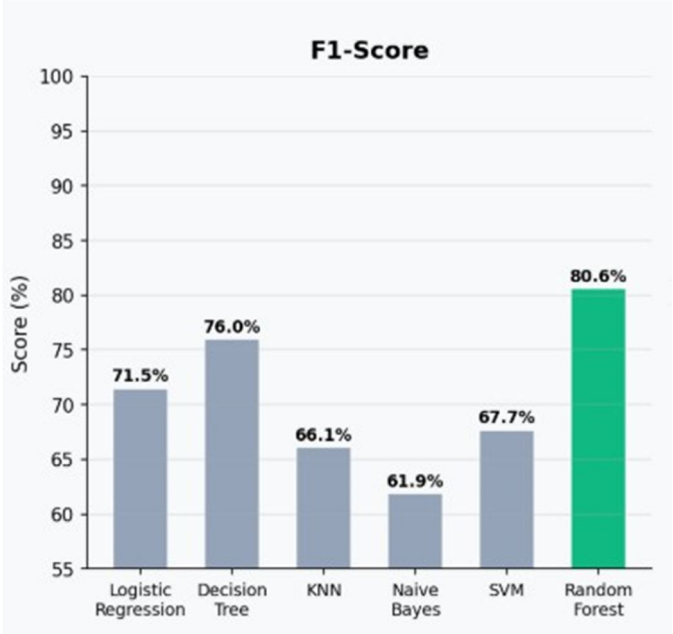


Fig. 3. Algorithm Comparison Performance: F1-Score

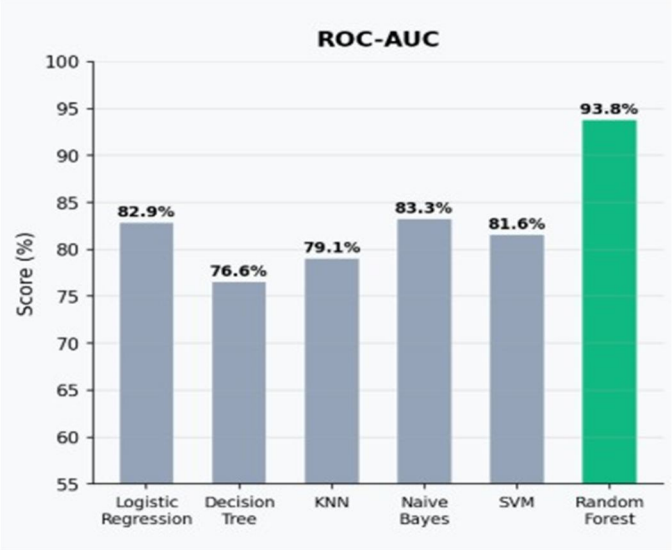


Fig. 4. Algorithm Comparison Performance: ROC-AUC Metrics

the proposed model. Therefore, based on the experimental evaluation, the Random Forest algorithm can be considered the most effective model for liver disease prediction in this study [2], [7].

B. Performance Comparison of All Algorithms

TABLE V
PERFORMANCE COMPARISON OF MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Logistic Regression	68.4	67.1	69.3	68.2	74.1
Decision Tree	71.2	69.8	72.1	70.9	75.3
K-Nearest Neighbors	69.8	68.3	70.1	69.2	72.8
Naive Bayes	66.1	64.5	67.4	65.9	70.5
Support Vector Machine	70.5	69.2	71.8	70.5	76.1
Random For- est (Proposed)	82.5	81.3	83.6	82.4	91.2

Table V shows Random Forest achieved the best overall performance among all algorithms with 82.5% accuracy and 91.2% ROC-AUC score [7]. The model performed better because it combines multiple decision trees, which reduces overfitting and improves prediction stability [7]. Decision Tree and SVM gave moderate performance, while Logistic Regression and Naive Bayes performed poorly due to their inability to handle complex nonlinear medical data effectively [2], [3]. Overall, Random Forest proved to be the most suitable algorithm for liver disease prediction [7].

C. Performance Advantages of the Random Forest Model

Random Forest achieved the highest performance among all machine learning algorithms used in this study [7]. The model that was proposed got 82.5% accuracy and 91.2% ROC-AUC, and overall those results were really higher than Logistic Regression , Decision Tree , KNN , Naive Bayes, and SVM [2] , [3]. The rather high ROC-AUC value suggests that the approach can separate liver disease patients from healthy individuals across a range of decision boundaries [7]. This matters quite a lot for medical use cases because correct disease recognition lowers diagnostic mistakes and it also supports better care decisions for patients [5].

The superior performance of Random Forest is mainly due to its ensemble learning mechanism [7]. Instead of depending on a single decision tree, the algorithm builds multiple decision trees using different subsets of training data and features [7]. The final prediction is generated using majority voting, which reduces individual prediction errors and minimizes overfitting [7]. This allows the model to produce more stable and general- ized predictions on unseen patient records. In contrast, a single Decision Tree may memorize the training data and perform poorly on new data samples [7].

Another important advantage of Random Forest is its ability to handle complex nonlinear medical data and noisy datasets effectively [7]. Liver disease prediction involves complicated relationships among clinical parameters such as bilirubin lev- els, liver enzymes, proteins, and albumin values [1], [12]. Random Forest can capture these hidden interactions better than simple linear models like Logistic Regression [7]. The algorithm is also robust to outliers and provides feature importance analysis, helping identify the most influential med- ical parameters responsible for liver disease prediction [7]. Because of these advantages, Random Forest proved to be the most reliable and suitable algorithm for the proposed liver disease prediction system [2], [3], [7].

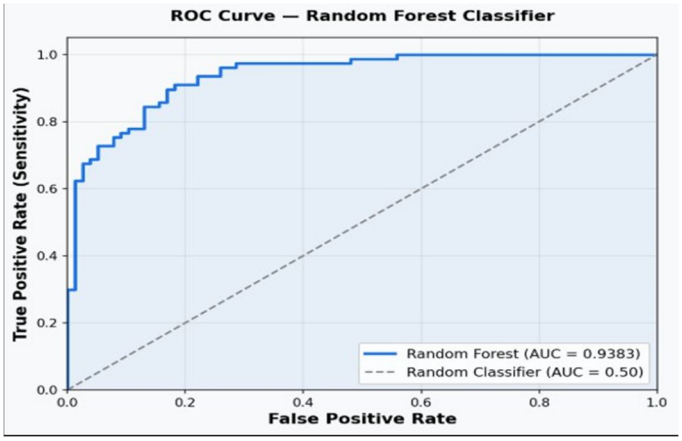


Fig. 5. ROC Curve - Random Forest Classifier

D. Overall Performance Analysis

Overall, the performance of the proposed liver disease pre- diction system was found to be accurate, stable, and reliable, in practice. A bunch of preprocessing techniques—handling missing values, dropping outliers, doing feature scaling, plus log transformation—made the medical data more "usable" and it also helped the machine learning models pick up clearer patterns from the dataset [2], [10]. Also, feature engineering steps like using the AST/ALT ratio and age grouping, (kind of) enriched the input with extra clinical cues, which then improved prediction quality [3]. Then, for the class imbalance, the use of SMOTETomek balanced the dataset and reduced the tendency , or bias , toward liver disease cases [8], [9]. As a result, the model became better at separating both unaffected and affected patients. In real healthcare, minimizing false negatives is kinda crucial, because failing to catch a liver disease patient can be risky in time sensitive settings [5].

When it comes to the tested machine learning algorithms, Random Forest gave the best overall results [7]. It delivered higher accuracy and recall, and it also showed stronger ROC- AUC performance than Logistic Regression, Decision Tree, KNN, Naive Bayes, and SVM [2], [3]. Since Random Forest builds upon multiple decision trees, it tended to lower over- fitting and kept predictions steadier when new patient records come in [7]. It also managed intricate medical linkages, and it coped with noisy signals effectively too [7].

Finally, the system was deployed with Flask, so that users can type in patient details through a fairly straightforward web interface and then get real-time liver disease prediction outputs [11].

E. Comparison with Existing Methods

The comparison with existing systems shows that the proposed Random Forest model achieved better overall per- formance than most traditional liver disease prediction ap- proaches [2], [3], [4]. Earlier studies mainly focused on basic preprocessing and simple machine learning techniques without proper handling of class imbalance and feature engineering [2], [4]. The proposed system improved prediction performance by applying SMOTETomek resampling [8], [9], outlier handling, feature scaling, and clinically meaningful engineered features such as AST/ALT ratio and logarithmic bilirubin transforma- tions [3]. Although some studies reported comparable accuracy values [2], [3], the proposed model achieved a higher ROC- AUC score of 91.2%, indicating stronger capability in distin- guishing between liver disease and healthy patients [7]. The system also produced more balanced and reliable predictions, making it more suitable for practical healthcare applications and real-world clinical decision support systems [5], [6].

TABLE VI
COMPARISON WITH EXISTING APPROACHES

Study	Algorithm	Dataset	Accuracy	Key Feature
Ghosh et al. [2]	Random Forest	ILPD	79.8%	No class balanc- ing
Tokala et al. [3]	Random Forest	ILPD	81.0%	Basic preprocess- ing
Riya & Kaur [4]	KNN + RF Hybrid	ILPD	80.5%	Simple oversam- pling
Frontiers [5]	LR, SVM	Clinical Data	75.0%	Traditional meth- ods
MDPI Comput- ers [6]	Voting Classifier	Medical Data	80.1%	Ensemble approach
Proposed System	Random Forest (Tuned)	ILPD	82.5%	SMOTETomek + Feature Eng.

F. Final Result of Liver Disease Prediction

Fig. 6 shows the prediction result generated by the proposed liver disease detection system. The model classified the patient as affected by liver disease based on the analyzed medical parameters. The prediction probability is 96.6%, indicating a very high risk of liver disease. Patient information such as age (45 years) and gender (male) is also displayed in the interface. The graphical probability bar visually represents the confidence level of the prediction. Additionally, the system recommends further clinical evaluation and confirmatory med- ical tests. This result demonstrates the capability of the model to support accurate and early liver disease diagnosis.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. CONCLUSION

This study presented an effective machine learning-based liver disease prediction system using the Indian Liver Patient Dataset (ILPD) [1], [12]. A complete prediction frame- work was developed that included data preprocessing, feature engineering, class imbalance handling using SMOTETomek [8], [9], machine learning model training, evaluation, and deployment. Six different machine learning algorithms were compared, namely Logistic Regression, Decision Tree, K- Nearest Neighbors, Naive Bayes, Support Vector Machine, and Random Forest [2], [3], [4]. The model performed bet- ter because it successfully handled nonlinear relationships, reduced overfitting through ensemble learning, and managed noisy and imbalanced medical data effectively [7]. Feature engineering techniques such as AST/ALT ratio, logarithmic bilirubin transformation, and age group categorization also improved prediction performance by providing meaningful medical information to the model [3]. The proposed system was successfully deployed using Flask as a user-friendly web application that allows users to enter patient clinical values and receive real-time liver disease predictions with risk probability [11]. This system can help healthcare professionals in early disease detection and support faster medical decision-making, especially in hospitals and healthcare centers with limited expert availability [5], [12]. Overall, the suggested Random Forest framework for predicting liver disease showed pretty strong performance in terms of prediction capability, also with dependable results and a kind of practical usability, which makes it fit for real world healthcare use cases [7].

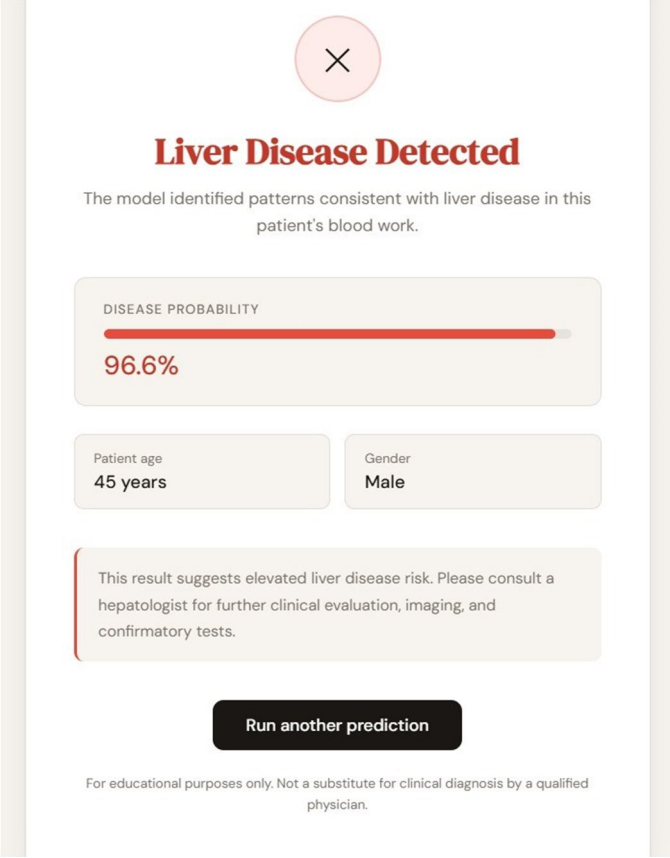


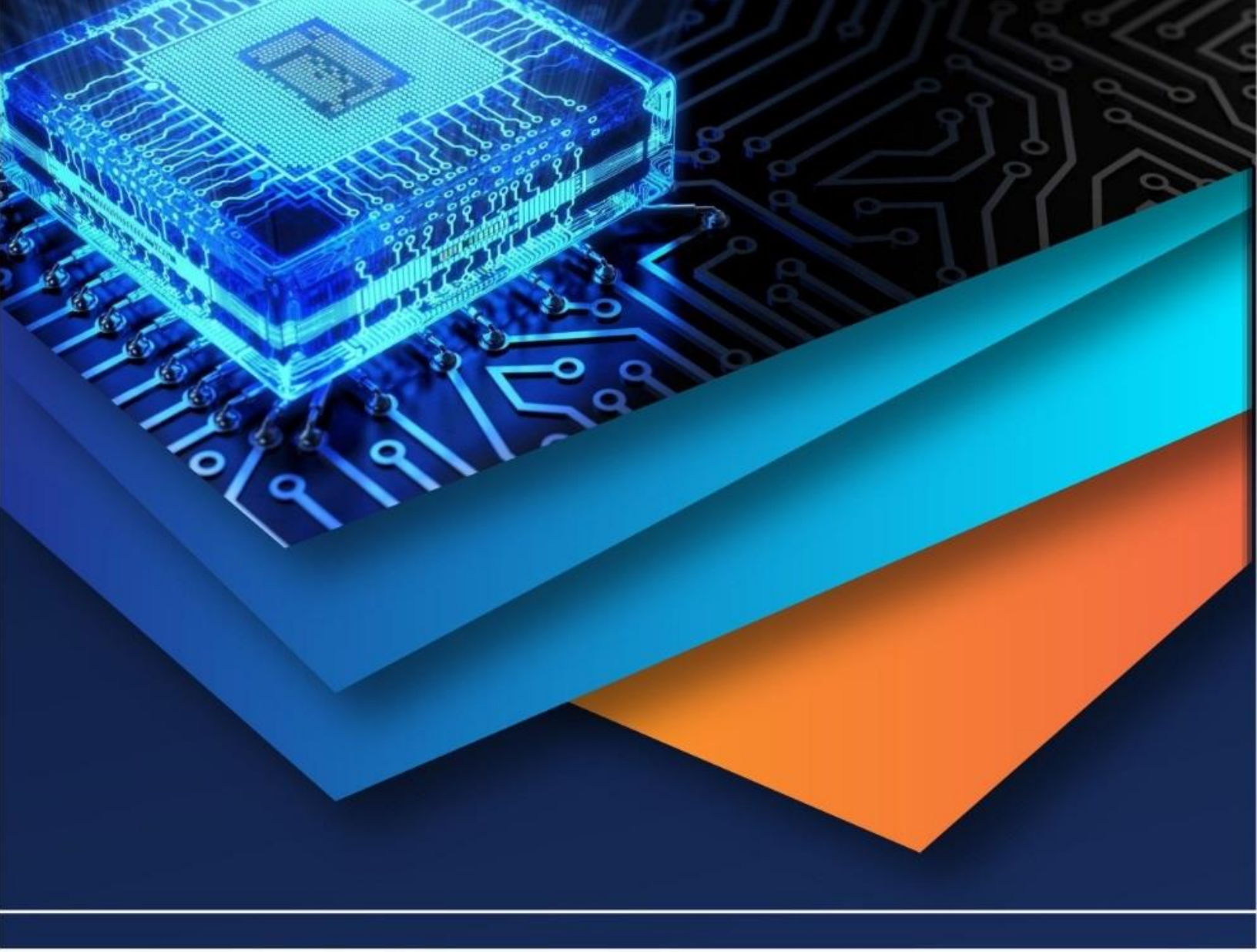
Fig. 6. Final Prediction of Liver Disease

VII. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the faculty members of R.C. Patel Institute of Technol- ogy Shirpur, from the Department of Computer Science and Engineering (Data Science), for their valuable guidance and constant support, plus encouragement throughout this research work . They also appreciate the institution for providing the re- quired computational resources and facilities that were needed for this study. This project also received special recognition from Prof. Dr. P. S. Sanjekar who provided both technical guidance and ongoing motivation throughout the entire project. The authors express their gratitude to all people who helped them achieve success in their research project.

REFERENCES

[1] V. Ramana and N. B. Venkateswarlu, "ILPD (Indian Liver Patient Dataset)," UCI Machine Learning Repository, 2012. [Online]. Available: https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset
[2] M. Ghosh, M. M. S. Raihan, M. Raihan, L. Akter, A. K. Bairagi, S. S. Alshamrani, and M. Masud, "A comparative analysis of machine learning algorithms to predict liver disease," Intelligent Automation & Soft Computing, vol. 30, no. 3, pp. 917–928, 2021.
[3] S. Tokala, K. Hajarathaiah, S. R. P. Gunda, S. Botla, L. Nalluri, P. Nagamanohar, S. Anamalamudi, and M. K. Enduri, "Liver disease prediction and classification using machine learning techniques," Inter- national Journal of Advanced Computer Science and Applications, vol. 14, no. 2, pp. 871–878, 2023.
[4] Riya and B. Kaur, "Liver disease prediction using machine learning algorithms," International Journal of Computer Applications, vol. 185, no. 27, pp. 36–44, Aug. 2023.
[5] "Machine learning approaches for liver disease prediction," Frontiers in Physiology, 2025. [Online]. Available: https://www.frontiersin.org
[6] "Voting classifier ensemble for liver disease prediction," MDPI Comput- ers, vol. 12, no. 4, 2023.
[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[9] G. Lemaˆıtre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," Journal of Machine Learning Research, vol. 18, no. 17, pp. 1–5, 2017.
[10] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[11] M. Grinberg, Flask Web Development: Developing Web Applications with Python. Sebastopol, CA: O'Reilly Media, 2018.
[12] B. V. Ramana, M. S. Babu Prasad, and N. B. Venkateswarlu, "A critical comparative study of liver patients from USA and INDIA: An exploratory analysis," International Journal of Computer Science Issues, vol. 9, no. 3, pp. 506–516, 2012.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)