Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260501021
Volume 5, Issue1 (January-April 2026), PP: 151-155.
www.indjcst.com
ISSN No: 2583-5300

Liver disease diagnosis using predictive analytics-based machine learning models

Garima Rathi[1], Shipra Tripathi[2], Rahul Singh[3]
[1]*Assistant Professor, Department of Computing Science, Uttaranchal University, Dehradun, Uttarakhand, India.*
[2]*Assistant Professor, Department of Computer Science, Institute of Technology and Management, Dehradun, Uttarakhand, India.*
[3]*Assistant Professor, Department of Computer Science, Sardar Bhagwan Singh University, Dehradun, Uttarakhand, India.*

To Cite this Article: *Garima Rathi[1], Shipra Tripathi[2], Rahul Singh[3], "Liver disease diagnosis using predictive analytics-based machine learning models", Indian Journal of Computer Science and Technology, Volume 05, Issue 01 (January-April 2026), PP: 151-155.*

Abstract: *In the world, liver diseases cause roughly one million deaths. Several traditional methods can be used to diagnose liver issues, but they are expensive. Early prediction and treatment of liver disease may benefit everyone at risk. Due to its early illness prediction capabilities, machine learning has a significant impact on healthcare as technology develops. This study assesses how well machine learning predicts liver disease. This article introduces the liver disease prediction (LDP) approach, which enables researchers, stakeholders, students, and medical professionals to predict liver illness. However, classification machine learning models such as logistic regression (LR), support vector machines (SVM), decision trees (DT), and random forests (RF) were used in this study. Python uses accuracy comparison to forecast the outcome. The random forest algorithm predicts liver diseases with the highest accuracy, per the results. Above the permitted accuracy threshold and could be taken into account when determining the prognosis of liver disease.*
It adjusts the learning process's cost function to account for the unequal distribution of classes, improving the model's performance. To predict liver disease, we first acquire an evenly distributed dataset and then train a machine learning model (specifically, logistic regression, support vector classifier, and gradient boosting classifier). A different test dataset is used to evaluate the effectiveness of the suggested model using a variety of criteria, including accuracy, precision, recall, and F1-score. This model is expected to significantly aid medical professionals in the field by effectively addressing class imbalance through data balancing algorithms.

Key Words: *Liver, Machine learning, LDP, SVM, DT, LDA, Predictive, Data balancing, and algorithms.*

I. INTRODUCTION

The human body contains a number of vital organs, each of which serves a very advantageous purpose. On the right side of the body, beneath the ribs, is a large, solid organ called the liver? It is situated above the right kidney, below the diaphragm, and above the stomach. The liver has numerous roles to play. [1] Among its many essential functions are "toxin removal, energy conversion from digested food, vitamin and mineral storage, and regulation of the amount of fat and sugar returned to the rest of the body."

The advancements in machine learning and its integration with health data science have significant implications for the identification of liver-based diseases, which rank among the leading causes of death worldwide [2]. Treatments for liver diseases like cirrhosis, fatty liver, hepatitis, and liver cancer are limited when they are identified in their advanced stages [2]. Early detection significantly improves patient outcomes while reducing medical costs.

There is an additional opportunity to use technology and data wisdom to improve patient care delivery and embody health care in the evolving landscape of information technology and health care. Fundamentally, machine learning (ML) uses artificial intelligence to find predictive patterns in massive data sets, thereby inducing predictive models more effectively and efficiently than traditional styles [3]. In light of this, these approaches can be used in a variety of hepatological contexts.

II.LITERATURE REVIEW

Infectious complications that arise following liver transplantation (LT) are quite prevalent and contribute to increased mortality rates and prolonged hospital stays [5]. By using a two-factor binary regression model to analyze the effects of bilirubin and INR levels on the fifth postoperative day following pediatric LT, it may be possible to prevent negative outcomes and aid in the early diagnosis of infections[4]. According to this study, laboratory and computed bilirubin and INR levels on the fifth postoperative day can be used to predict with considerable accuracy the occurrence of infectious complications in the early postoperative phase following LT. INR and total bilirubin levels can rise as a result of liver function being compromised by infections[3].

Liver diseases are becoming more common over time, making it challenging to detect these conditions early on. Researchers have implemented numerous data mining models and machine learning techniques to identify such diseases in their early stages [6]. However, in this area of liver disease prediction, it has been observed through experimental results that CHIRP is effective in

reducing the error rate in evaluation metrics compared to other models used[7]. When comparing performance, RF and MLP exhibit better accuracy than CHIRP. Nevertheless, the differences in accuracy between RF, MLP, and CHIRP are not significant when contrasted with the higher error rates mentioned [8].

The suggested convolutional neural network was developed and executed. It was designed for the classification of uninfected liver images and images of metastasized (infected) lesions using TensorFlow. For image sizes of 65×65, 60×60, and 55×55, the highest classification accuracy achieved was 99% for the 65×65 image size. The proposed network was assessed against previous research and showed a significant improvement in classification accuracy. This enhancement was achieved through the implementation of a regularization technique to reduce the overfitting issue. Observations indicated that the F1 score is nearly one, demonstrating a "balance between recall and precision. Therefore, it can be concluded that the proposed CNN model is the most effective for the binary classification of infected and uninfected liver CT images".

Because liver disease is difficult to diagnose because its symptoms are so subtle, this study is essential to determining which algorithms are most accurate at predicting this dangerous condition. Following that, five different supervised learning strategies are put into practice using R and evaluated using metrics from the confusion matrix [8]. The results show that K-NN is the most accurate method for predicting liver disease, with an accuracy of 91 percent. The superior ability of autoencoders to recognize overlapping features over traditional K-NNs results in slightly better performance than K-NNs. Seventy-five percent is the acceptable accuracy threshold that most algorithms surpass [9].

Liver Cirrhosis, which can be a life-threatening condition, demands urgent care to avert serious health complications. The implementation of machine learning models could greatly improve the early detection of cirrhosis, potentially minimizing its long-term harmful effects on health. A range of machine learning algorithms has been evaluated for their capacity to forecast liver infections based on various physiological markers, showing potential for future advancements in medical systems [9]. These advancements are anticipated to enhance the precision and effectiveness of these tools. Furthermore, machine learning solutions could aid the public in evaluating the risk of severe conditions such as stroke in adults. Ideally, individuals with liver disease (LD) would gain from early identification and treatment, giving them a better opportunity to manage and recuperate from their illness [10].

Description of the dataset

The Indian Liver Patient Dataset (ILPD) dataset contains databases with 585 records/entries that are extracted in order to address the challenge of this paper. The source of this dataset was the UCI Machine Learning Repository. Details about 585 liver patients from India are included in the entire ILPD dataset [2]. 417 records pertaining to liver patients and 169 data unrelated to liver patients are included in this. The data collection was conducted in Andhra Pradesh, India's northeast. One type of class label that divides individuals into groups according to whether or not they have liver disease is called a selector [11].

Name of Attributes	Data Type	Description
Gender	Boolean	Male=0 , Female=1
Age	Integer	Patient age in Year
Total Bilirubin	Float	Total bilirubin in the blood(mg/dl) High levels indicate liver problem
Direct Bilirubin	Float	Bilirubin level in the blood (mg/dL.) indicate bile flow
Alkaline Phosphatase (ALP)	Integer/float	Enzyme level in blood (U/L). High level indicates liver bile duct damage
Alanine Aminotransferase (ALT)	Integer/float	Liver enzyme level suggest liver cell damage
Aspartate Aminotransferase (AST)	Integer/float	It is also an enzyme, that indicates potential liver damage.
Total Proteins	Float	Amount of proteins in blood(g/dL). Low levels indicate liver dysfunction.
Albumin	Float	Show poor liver function
Albumin-Globulin Ratio	Float	Low ratio indicates liver damage
Liver Disease	Binary	1= liver disease present 0=No liver disease

Table: 2.1 Dataset Description

III.PERFORMANCE EVALUATION MATRICES

Evaluation of the model is a crucial part of any research project. Using a few standard assessment metrics to evaluate your model may yield results that satisfy your requirements. The proposed model in this study is evaluated in relation to other models using the following metrics.

3.1 Accuracy: In classification problems, it is the most widely used evaluation metric. Out of all the cases, it calculates the number of correctly predicted cases.

$$\textbf{Accuracy} = \frac{\text{True Positive (TP)} + True\ Negative\ (TN)}{\text{Total Number of Predictions}}$$

3.2 Recall (True Positive rate)

$$Recall = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + False\ Negative\ (FN)}$$

It measures how well the model identifies actual positives

3.3 Precision:

$$Precision = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + False\ Positive(FP)}$$

3.4 F1-Score:

$$F1\ Score = 2 * \frac{\text{Precision . Recall}}{\text{Precision} + \text{Recall}}$$

Provide a balance between precision and Recall

3.5 Confusion Matrix:

The positive Rate (Recall/Sensitivity)

False positive Rate:

$$FPR = \frac{FP}{FP + TN}$$

False Negative Rate:

$$FNR = \frac{FN}{FN + TP}$$

IV.MACHINE LEARNING ALGORITHMS

The planning function that converts input variables (X) into output variables (Y) is learned using labeled training data by supervised learning algorithms in order to make the best prediction. In other words, it solves for f in the equation that follows: Giving us new inputs enables us to produce outputs that are accurate [12].

4.1 Classification: - When the production variable is in the form of categories, the prediction is made about the result of a specific sample [9]. Labels such as "having liver disease" or "not having liver disease" may be predicted by a classification model based on the input data. A supervised classification model is used in this study to forecast liver diseases [12].

4.1.1 Logistic Regression: The logistic regression method is used to predict the results of a categorical dependent variable. Consequently, the outcome must be a category or discrete value. True or false, yes or no, 0 or 1, etc., but probabilistic values between 0 and 1 can be utilized [13].

4.1.2 Support Vector Machine (SVM): Due to its popularity, SVM is currently being tested for use in a wide range of fields. Regression, ranking, and classification function learning are the primary uses for SVMs. SVMs look for the hyperplane, or location of decision boundaries, that results in the best possible class separation. Structural risk minimization and statistical learning theory serve as their foundations [12].

4.1.3 Decision Tree: When categorizing problems, a decision tree produces a double tree. This strategy is important when it comes to classification problems. Using a tree to carry out the classification process is also applicable to a single record in the dataset and the item being classified for that specific record. Based on how closely the categorization value for different records is interpreted, for instance, the value for that element can be predicted. The J48 algorithm simulates the lost values during this process. According to the item's quality standards, the data should be divided into runs [14].

4.1.4 Random Forest: Being a very versatile classifier, this supervised classification approach is thought to be among the most advanced ensemble learning strategies out there[8]. This technique, as its name suggests, produces a forest with several trees. Even though a lot of trees are involved, more trees in the forest improve accuracy in RF research.

V.MACHINE LEARNING MODEL PERFORMANCE ANALYSIS BASED ON GIVEN MATRICES

5.1 Comparisons between different Matrices

The following Table 1 and Fig. 1 describe the matrix's performance using a machine learning model, logistic regression, and using a support vector machine (SVM)

Matrices	Logistic Regression	Support Vector Machine
Accuracy	81.0	85.1
Recall	82.3	86.3
Precision	79.2	84.1
F1 Score	81.1	85.2

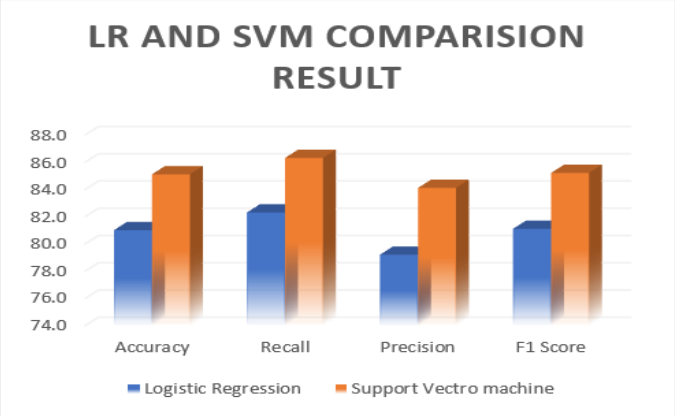

Figure 1: Comparison result between LR and SVM

According to LR and SVM, SVM Recall provides better result to predict liver diseases.

The following Table 2 and Fig. 2 describe the matrices' performance using a machine learning model, a decision tree, and a random forest.

Matrices	Decision Tree	Random Forest
Accuracy	78.1	89.2
Recall	80	91.1
Precision	76.2	87.4
F1 Score	78	89.3

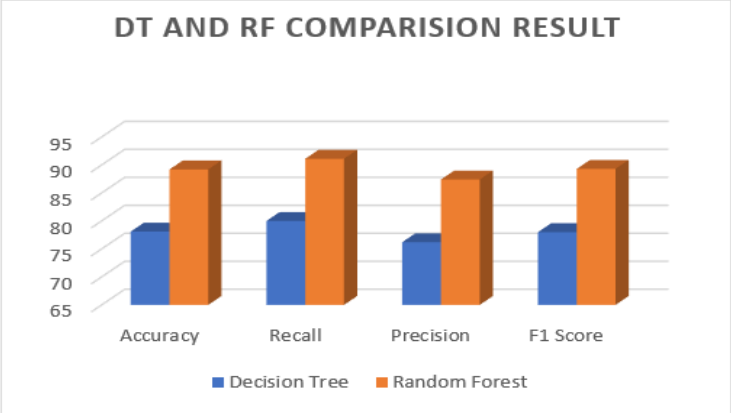

Figure 2: Comparison result between DT and RF

VI.CONFUSION MATRIX

The following table :3 and Fig 3 evaluate the machine learning model using confusion matrices using parameters TN, FP, FN, and TP.

ML Models	TN (True Negative)	FP (False Positive)	FN (False Negative)	TP (True Positive)
LR	90	22	15	77
DT	85	23	20	70
SVM	92	18	14	78
RF	95	10	10	82

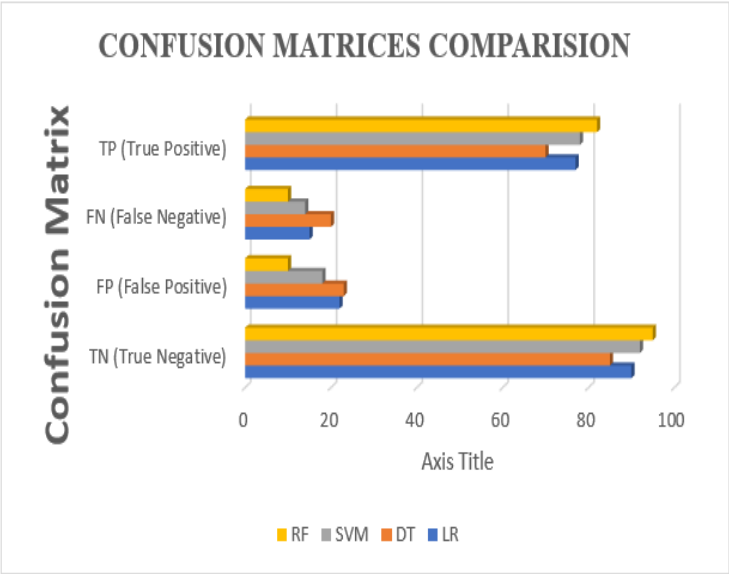

Figure 3: Confusion matrices Comparison result

VII.ANALYSIS OF RESULT

All models were evaluated based on the matrices test dataset. The result summarizes all the graphs provided below. Table 4 and Figure 4.

ML Models	Accuracy	Recall	Precision	F1 Score
Logistic Regression	81.1	83.3	79.5	81.3
Decision Tree	78.2	80.0	76.2	78.0
Support Vector Machine	85	86.5	84.0	85.6
Random Forest	**89**	**91.3**	**87.5**	**89.6**

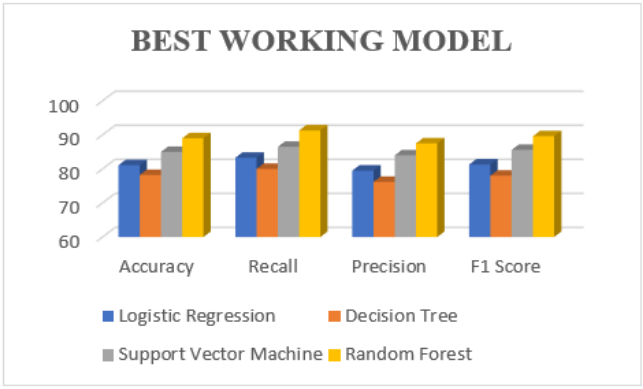

Figure 4: Best working model

The best analysis and prediction results for liver disease are obtained using a random forest.

VIII.DISCUSSION

Predictive machine learning algorithms. liver diseases have the potential to enhance early diagnosis and improve patient outcomes. In this research paper, various studies have evaluated different machine learning algorithms to determine their effectiveness in this field. To predict the best result, compare different learning models.

IX.CONCLUSION

The main conclusion of this research paper is that, while using the machine learning models, mainly Random Forest, it considers dataset quality, interpretability, and clinical integration for their successful results in healthcare fields. Random forest showed high accuracy in identifying liver diseases. SVM and Decision Tree can be a secondary choice because of low performance with this dataset.

REFERENCE

1. Ganie, S. M., Dutta Pramanik, P. K., & Zhao, Z. (2024). Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches. BMC Medical Informatics and Decision Making, 24(1), 160.
2. Al Ahad, A., Das, B., Khan, M. R., Saha, N., Zahid, A., & Ahmad, M. (2024). Multiclass liver disease prediction with adaptive data preprocessing and ensemble modeling. Results in Engineering, 22, 102059.
3. Islam, R., Sultana, A., & Tuhin, M. N. (2024). A comparative analysis of machine learning algorithms with tree-structured parzen estimator for liver disease prediction. Healthcare Analytics, 6, 100358.
4. Zhang, Z., Wang, S., Zhu, Z., & Nie, B. (2023). Identification of potential feature genes in non-alcoholic fatty liver disease using bioinformatics analysis and machine learning strategies. Computers in biology and medicine, 157, 106724.
5. Lanjewar, M. G., Parab, J. S., Shaikh, A. Y., & Sequeira, M. (2023). CNN with machine learning approaches using ExtraTreesClassifier and MRMR feature selection techniques to detect liver diseases on cloud. Cluster Computing, 26(6), 3657-3672.
6. Takahashi, Y., Dungubat, E., Kusano, H., & Fukusato, T. (2023). Artificial intelligence and deep learning: New tools for histopathological diagnosis of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. Computational and Structural Biotechnology Journal, 21, 2495-2501.
7. Aslam, M. H., Hussain, S. F., & Ali, R. H. (2022, November). Predictive analysis on severity of non-alcoholic fatty liver disease (nafld) using machine learning algorithms. In 2022 17th International Conference on Emerging Technologies (ICET) (pp. 95-100). IEEE.
8. Dalal, S., Onyema, E. M., & Malik, A. (2022). Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy. World Journal of Gastroenterology, 28(46), 6551.
9. Che, H., Brown, L. G., Foran, D. J., Nosher, J. L., & Hacihaliloglu, I. (2021). Liver disease classification from ultrasound using multi-scale CNN. International Journal of Computer Assisted Radiology and Surgery, 16(9), 1537-1548.
10. Okanoue, T., Shima, T., Mitsumoto, Y., Umemura, A., Yamaguchi, K., Itoh, Y., ... & Harada, K. (2021). Artificial intelligence/neural network system for the screening of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Hepatology Research, 51(5), 554-569.
11. Su, T. H., Wu, C. H., & Kao, J. H. (2021). Artificial intelligence in precision medicine in hepatology. Journal of Gastroenterology and Hepatology, 36(3), 569-580.
12. Nahar, N., Ara, F., Neloy, M. A. I., Barua, V., Hossain, M. S., & Andersson, K. (2019, December). A comparative analysis of the ensemble method for liver disease prediction. In 2019 2nd international conference on innovation in engineering and technology (ICIET) (pp. 1-6). IEEE.
13. Arbain, A. N., & Balakrishnan, B. Y. P. (2019). A comparison of data mining algorithms for liver disease prediction on imbalanced data. International Journal of Data Science and Advanced Analytics, 1(1), 1-11
14. LaPierre, N., Ju, C. J. T., Zhou, G., & Wang, W. (2019). MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. Methods, 166, 74-82.