Eng. Cov & Innov. An Int. J. Volume 01, Issue 02, May – August, 2026, ISSN 3108-1843 (online)

Universal Research Forum
Engineering Convergence and Innovation (ECI) An International Journal
www.universalresearchforum.com

Liver Disease Prediction Using Machine Learning Techniques

Bhuvan K Pujar[1*], Bhaskar S R[1], Charan K[1], Ajeedahmad Nadaf[1], Darshan B R[1], Gajanan M Naik[2]

[1]Department of Computer Science and Engineering, RV Institute of Technology and Management, JP Nagar-560076 Bengaluru, Karnataka, India
[2]Department of Mechanical Engineering, RV Institute of Technology and Management, JP Nagar-560076 Bengaluru, Karnataka, India

*Email: bhuvanpujar2004@gmail.com

Published Date: 17 July 2026.

ABSTRACT

The problem of liver disease is considered a crucial one in global public health which is usually diagnosed only at advanced stages causing serious health issues and even death. The key problem under consideration in this project is the insufficient efficiency of liver disease diagnosis methods. The aim of this project is to create a predictive system based on machine learning methods which will be able to classify the patients as either having liver disease or not depending on the provided clinical information. In order to reach the stated aim, we have used the dataset that contains information about 583 patients with the biochemical features. Data preprocessing operations like dealing with missing data, feature encoding and normalization were conducted to increase data quality. We have developed several classifiers like Random Forest, AdaBoost, Support Vector Machine and Naive Bayes Classifier. The best results among them were obtained by Random Forest algorithm with the accuracy of 77.8%.

Keywords - *Machine Learning, Liver Disease Prediction, Healthcare Analytics, Classification Algorithms, Random Forest.*

1. INTRODUCTION

The suggested approach involves a systematic process for machine learning which includes stages of data collection, data preprocessing, model training and model evaluation. In this particular project, the dataset under analysis involves 583 records for patients with 11 features, namely: age, gender, bilirubin level, enzyme level and protein level. Data preprocessing is an important step that contributes to enhancing the effectiveness of models. The missing data in the dataset has been imputed using the median method. The categorical feature, such as gender, was converted into numerical format. Normalization was applied to scale features uniformly. The dataset has been split into two parts, namely, a training dataset (70%) and testing dataset (30%). Random Forest, AdaBoost, Support Vector Machine (SVM) and Naive Bayes machine learning algorithms have been used.

2. METHODOLOGY

The designed system consists of data collection, data preprocessing, model building, and performance assessment stages according to a standard machine learning procedure. The dataset used in this paper includes 583 patient records and has 11 features: age, gender, bilirubin levels, enzymes, and proteins.

The stage of data preprocessing has an important role in improving the performance of the models. Missing values in the dataset were replaced by the median imputation method, whereas categorical variables such as gender were converted into numbers. Feature scaling methods were applied for normalizing the data.

The training dataset included 70% of the records and the remaining 30% was used as a test set. The four machine learning algorithms which were used were Random Forest, AdaBoost, SVM, and Naive Bayes.

3. RESULTS AND DISCUSSION

Performance analysis of the machine learning models that were developed was carried out using classification accuracy and other measures. Random Forest classifier provided the highest accuracy score of 77.8%, thus proving to be the best classifier for this particular data set. The accuracy scores of AdaBoost, Naive Bayes, and SVM classifiers were 70.4%, 69.5%, and 67.8%, respectively. Superiority of Random Forest model in performance is explained by the fact that it is based on an ensemble learning methodology and uses several decision trees. Analysis of the results obtained from confusion matrices shows that Random Forest is able to classify both positive and negative cases with relatively high accuracy rates; however, SVM is not as good in dealing with class imbalance. Thus, a comparative analysis of all machine learning models used proves that the models based on ensemble learning algorithms are better than one-model algorithms for medical predictions. Testing of the system was also conducted with input real-time data, thus proving that disease outcomes could be predicted successfully.

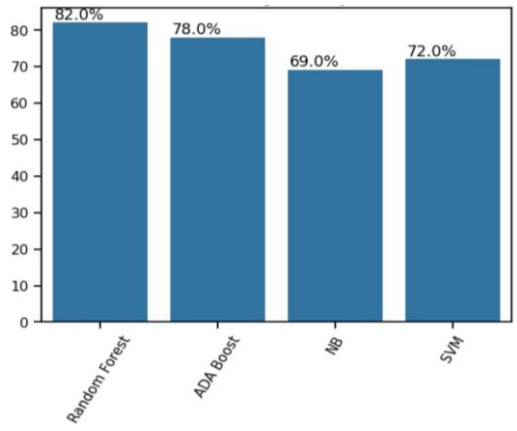

Fig 1. Classification Accuracy Comparison of Models

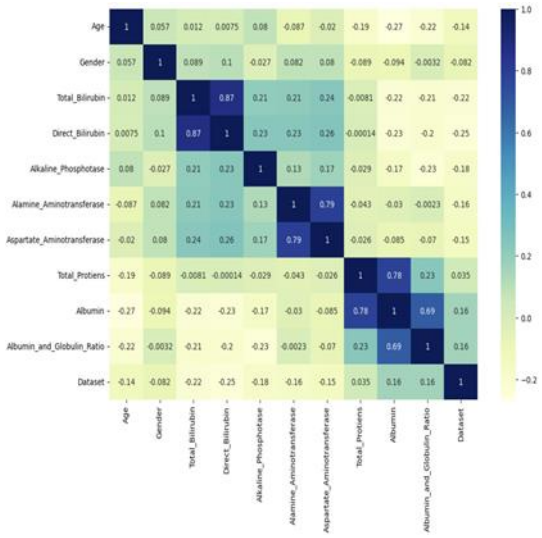

Fig 2. Correlation of data using Heatmap

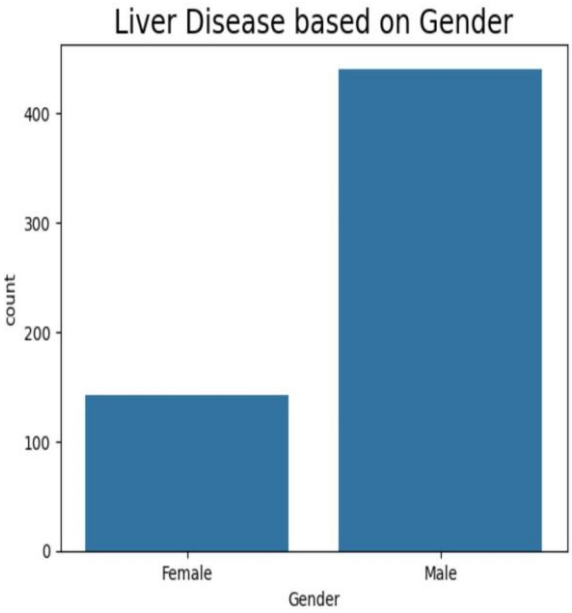

Fig 3. Liver Disease based on Gender

Table 1. Experimental Results

Model	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	77.8	0.78	0.78	0.78
AdaBoost	70.4	0.70	0.70	0.70
Naive Bayes	69.5	0.73	0.70	0.68
SVM	67.8	0.71	0.68	0.66

Age	Gender	Bilirubin	Enzymes	Prediction
38	Male	2.6	High	Positive
30	Male	0.5	Normal	Negative

4. CONCLUSION

The study suggests a machine learning-based method for liver disease prediction through clinical information. The presented approach proves to be effective in demonstrating the use of machine learning in healthcare diagnostics.

- The Random Forest method yielded the best results with accuracy rate of 77.8% The use of machine learning allows for early liver disease prediction.
- The system decreases reliance on conventional diagnostic tools Allows for quick and accurate predictions.
- It can help healthcare specialists in making decisions It is applicable for future healthcare purposes

REFERENCE

[1] Seneviratne, S., Hu, Y., Nguyen, T., Lan, G., Khalifa, S., Thilakarathna, K., ... & Seneviratne, A. (2017). A survey of wearable devices and challenges. IEEE Communications Surveys & Tutorials, 19(4), 2573-2620.

[2] Chen, X., Kim, D. H., & Lu, N. (2024). Introduction: wearable devices. Chemical Reviews, 124(10), 6145-6147.

[3] Dua, D., & Graff, C. (2019). UCI machine learning repository. University of California, School of Information and Computer Science. Irvine, CA, available at: http://archive. ics. uci. edu/ml (last accessed 07.11. 2024).

[4] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[6] Durai, V., Ramesh, S., & Kalthireddy, D. (2019). Liver disease prediction using machine learning. Int. J. Adv. Res. Ideas Innov. Technol, 5(2), 1584-1588.

[7] Ghazal, T. M., Rehman, A. U., Saleem, M., Ahmad, M., Ahmad, S., & Mehmood, F. (2022, February). Intelligent model to predict early liver disease using machine learning technique. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-5). IEEE.

[8] Dritsas, E., & Trigka, M. (2023). Supervised machine learning models for liver disease risk prediction. Computers, 12(1), 19.

[9] Sontakke, S., Lohokare, J., & Dani, R. (2017, February). Diagnosis of liver diseases using machine learning. In 2017 international conference on emerging trends & innovation in ICT (ICEI) (pp. 129-133). IEEE.

[10] Rahman, A. S., Shamrat, F. J. M., Tasnim, Z., Roy, J., & Hossain, S. A. (2019). A comparative study on liver disease prediction using supervised machine learning algorithms. International Journal of Scientific & Technology Research, 8(11), 419-422.