Vol. 2 - No. 2
March 2025 E-ISSN: 2584-0657

Hepatitisnet: Analysis and Prediction of Hepatitis using Machine Learning

K. Dinesh Reddy[1], Dr. S. Poongodi[2], Dr. P. Gnanavel[3], K. Jyothika[4]
K. Harish Goud[5], K. Kartheek[6]
[1,4,5,6] Department of Electronics and Communication Engineering
CMR Engineering College, Hyderabad, Telangana, India
[2] Professor, Department of Electronics and Communication Engineering
CMR Engineering College, Hyderabad, Telangana, India
[3] Associate Professor, Department of Computer Science Engineering
Rural Engineering College, Hulkoti, Karnataka, India
DOI: *https://doi.org/10.5281/zenodo.15641565*

Abstract

Hepatitis is a liver infection caused by the HCV virus. In hepatitis, it is hard to detect in the early stage as symptoms arise late. It can predict in advance in order that the patients may not suffer from permanent damage of liver. The main objective of this proposed method is to implement some machine learning techniques to predict this disease utilizing easily available and inexpensive data from blood test to make a diagnosis and take care of the patients in the early phase. The proposed method consists of a comparison of three machine learning techniques, those are support vector machine (SVM), logistic regression (LR), decision tree(DT), have been applied in same dataset. To achieve a appropriate strategy for disease prediction, confusion matrix, precision, recall, F1 score, accuracy, receiver operating characteristics(ROC), performance of all these algorithm results have been compared. Total accuracy of SVM model is 0.92, which is the maximum among the three models.
Keywords: *Deep Learning, HCV Virus*

INTRODUCTION

Hepatitis, a series of viral illnesses that affect the liver, remains a major global health concern, affecting millions of people worldwide. Early identification and accurate forecast of infection progression are crucial for effective management and treatment. Traditional diagnostic methods, although widely used, often rely on clinical evaluations and laboratory tests that can be huge time taken method and possibilities of having human error. The advent of machine learning (ML) provides a promising solution to enhance diagnostic accuracy and provide more timely predictions of disease outcomes.

The proposed method introduces HepatitisNet, a machine learning-based framework designed to analyze and calculate the hepatitis infection outcomes and stages based upon patient data. By leveraging various ML algorithms, the system aims to not only identify the presence of the disease but also predict the probability of complications such as cirrhosis or liver cancer. The integration of such analytical models into healthcare systems has the potential to modernize the way hepatitis is managed, allowing healthcare peoples to make informed decisions and personalize treatment strategies.

The research delves into the relevance of different machine learning techniques, including different categorization and regression models, to assess their performance in hepatitis prediction. With a robust dataset and a variety of performance parameters, this proposed method aims to provide helpful insights into the future of hepatitis diagnosis and prediction, highlighting the potential of AI-driven approaches in the field of medical diagnostics.

Problem Statement

Early detection of hepatitis is difficult with irregular clinical manifestations and restricted availability of diagnostics. The current research recommends the establishment of an effective, affordable, and scalable diagnostic system for correct hepatitis identification. The platform incorporates machine learning models to enhance detection, reduce false positives, and facilitate timely treatment recommendations.

LITERATURE REVIEW

Hepatitis prediction has gained attention due to rising liver disease cases globally. Researchers used clinical, laboratory, and demographic data to apply a variety of machine learning (ML) and deep learning (DL) approaches. Early studies used statistical models like logistic regression and decision trees, which had moderate accuracy. Recent advancements leverage deep learning, ensemble learning, and multi-omics data, though challenges remain in data quality, class imbalance, and feature selection.

Hepatitis C has been predicted and diagnosed in several studies using machine learning algorithms. Ma et al. [39] developed a number of classification models and determined that the XG Boost method achieved the highest accuracy (91.56%). Three machine learning methods were used by Ahammed et al. [40], and the greatest accuracy (94.40%) was obtained using KNN. In their investigation, Nandipati et al. [41] discovered that binary class labels performed better than multiclass labels, and they used the RF model to obtain an accuracy of 54.56%. Among the four machine learning models created by Mamdouh et al. [42], the RF model achieved an accuracy of 94.06% without hyper parameter adjustment and 94.88% with tuning. Using multiple-classifier models, El-Salam et al. [43] obtained accuracy rates between 65.6% and 68.9%. Applying a number of machine learning techniques, Hashem et al. [44] discovered For predicting advanced chronic hepatitis C, the accuracy was between 66.3% and 84.4%. Neural networks had the best accuracy (95.12%) among the algorithms tested by Syafa'ah et al. [45]. With four machine learning methods, Oleiwi et al. [46] determined that the decision tree approach classified and diagnosed hepatitis C with the highest accuracy (93.44%). This work attempts to choose the optimal algorithms using frequent and low-cost blood test data for hepatitis C prediction.

EXISTING METHOD

In the existing method, we employ Logistic Regression, a widely used probabilistic classifier, to predict Hepatitis C infection. Logistic regression calculates the likelihood that a given input belongs to a specific category (for example, Hepatitis C positive or negative). Logistic Regression, as opposed to linear regression, which predicts continuous values, uses the sigmoid function to ensure that output values remain between 0. The logistic regression function is defined as:

$$P(Y=1|X) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n)}} \qquad (1)$$

where

- $P(Y=1|X)$ is the probability of a positive class (Hepatitis C).
- $X_1, X_2, \ldots, X_n$ are the independent variables (selected features)
- $b_0, b_1, \ldots, b_n$ are the model coefficients, learned from the data.

Proposed Method

HepatitisNet is a machine learning-based system for diagnosing and predicting hepatitis. Early detection can improve patient outcomes, and HepatitisNet analyses patient demographics, medical history, and lab results to identify patterns. The system workflow includes:

Data Collection: Acquiring relevant patient data.

Data Preprocessing: Cleaning, handling missing values, normalizing, and encoding data.

Feature Selection: Identifying key attributes for better model efficiency.

Model Training: Using Logistic Regression, SVM, and Random Forest.

Data Preprocessing

Preprocessing ensures data is clean and suitable for machine learning. It involves:

- Handling missing values and encoding categorical data.
- Splitting datasets into training and testing sets.
- Feature scaling and balancing using SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance.

Exploratory Data Analysis (EDA)

EDA helps understand dataset distribution and relationships using:

Visualizations: Histograms (ALP, Age), scatter plots (AST vs. ALT), and heatmaps.

Correlation Analysis: Identifying key relationships between medical indicators.

Dataset Splitting

The dataset is divided into:

Training Set: Used to train the model.

Test Set: Evaluates the model’s performance on unseen data.

Support Vector Machine (SVM)

SVM is used for classification by creating a decision boundary (hyperplane). Steps include:

Model Training: Selecting an appropriate kernel (Linear, RBF, Polynomial).

Hyperparameter Tuning: Optimizing performance using techniques like Grid Search.

Prediction & Decision Making: Classifying new data points based on the learned hyperplane.

Advantages of the Proposed System

- Handles classification and regression problems.
- Reduces overfitting through ensemble techniques.
- Works well with missing/null values.
- Offers parallelization and stability.
- Avoids dimensionality issues by selecting relevant features

RESULTS

The proposed method has developed a data analysis and machine learning workflow for predicting hepatitis. Here's an explanation of the code:

Importing Libraries: The code starts by importing necessary Python libraries for data analysis and machine learning. These include pandas, matplotlib, seaborn for data visualization, imbalanced-learn (imblearn) for handling class imbalance using SMOTE, and various modules from Scikit-learn is a machine learning library that may be used for applications such as logistic regression, random forest, and support vector classification.

Loading the Dataset: The code is designed to taken as a input from given dataset in CSV file format named ('hcvdat.csv') using pandas library and stores it in a DataFrame named ‘df’.

Data Exploration and Visualization: Several data exploration and visualization steps are performed using Seaborn and Matplotlib to understand the dataset. This includes examining the dataset's structure, summary statistics, class distribution, and various plots to visualize relationships between features.

Dataset Description

The dataset contains demographic and medical test results to predict hepatitis categories. Below is a summary of key features:

Target Variable

- **Category:** Represents the type of hepatitis (the label to be predicted).

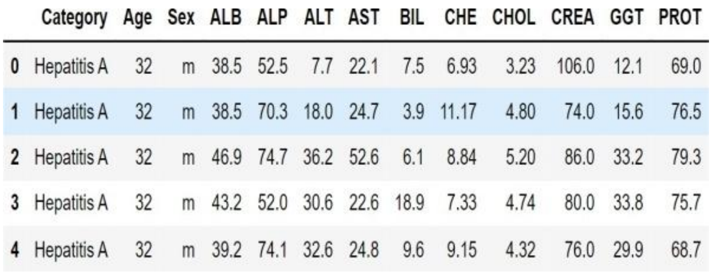
Independent Variables

Demographic Data:

- **Age:** Patient Age.
- **Sex:** Patient Gender (Male/Female).

RESULTS AND DESCRIPTION

Fig 1 represents a sample of the dataset, displaying various attributes such as Age, Sex, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT, and the Hepatitis Category.



	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	Hepatitis A	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	Hepatitis A	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	Hepatitis A	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
3	Hepatitis A	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	Hepatitis A	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7

Fig 1 Sample Dataset image

The Fig 2 shows the distribution of ALP (Alkaline Phosphatase) values on the x- axis and the count of data points on the y-axis. It helps to visualize the frequency of different ALP values in the dataset.

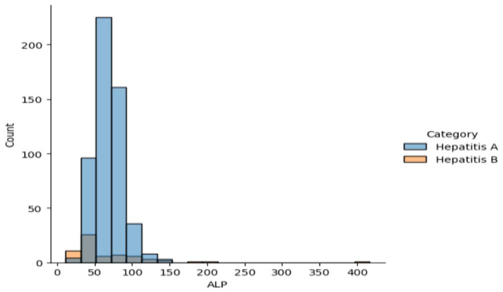


Fig 2 ALP Vs Count distribution of both classes

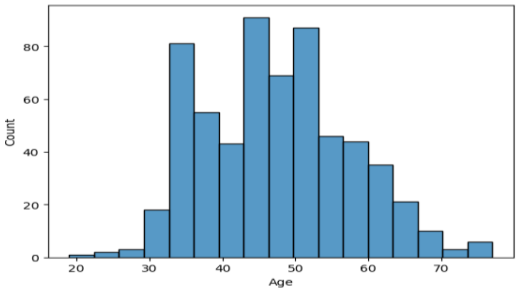


Fig 3 Age Vs Count distribution of cells

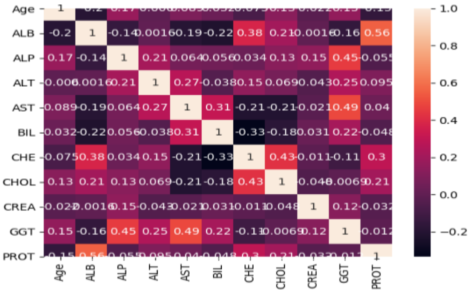


Fig 4 Correlation matrix

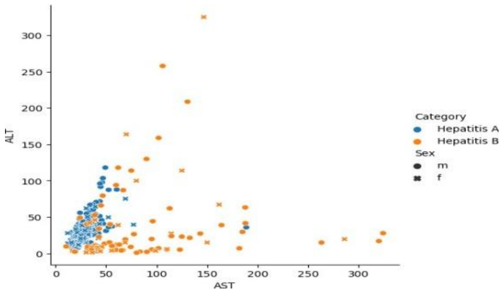
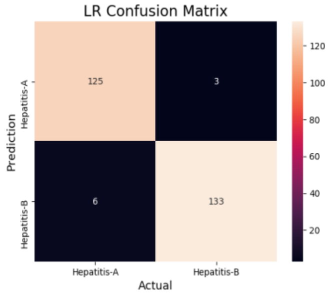


Fig 5 AST Vs ALP distribution of both classes



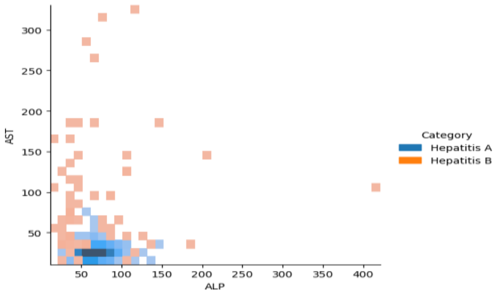


Fig 6 ALP Vs AST distribution of both classes **Fig 7 LR Confusion matrix**

Table 1 is a detailed categorization report for the Logistic Regression model. It provides precision, recall, and F1-score values for each category (Hepatitis-A and Hepatitis-B), along with their weighted average and macro average. These scores help assess the performance of the model for each category.

Table 1 LR Classification report

Class	Accuracy	Precision	Recall	F1-Score	Support
Hepatitis-A	0.955	0.95	0.98	0.97	128
Hepatitis-B	0.966	0.98	0.96	0.97	139
Macro Avg	0.954	0.97	0.97	0.97	267
Weighted Avg	0.966	0.97	0.97	0.97	267

Table 2 is the classification report for the Random Forest model. It provides precision, recall, and F1-score metrics for each category, along with their averages.

Table 2 RF model Classification report

Metric	Hepatitis-A	Hepatitis-B	Macro Avg	Weighted Avg
Accuracy	0.950	0.950	0.950	0.950
Precision	0.940	0.950	0.960	0.950
Recall	0.940	0.960	0.950	0.950
F1-Score	0.950	0.950	0.950	0.950

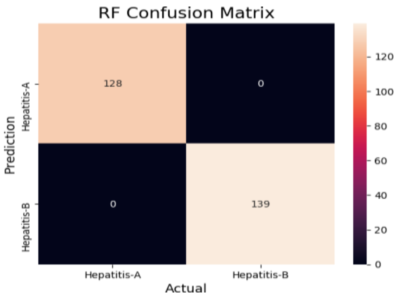


Fig 8 RF Confusion matrix

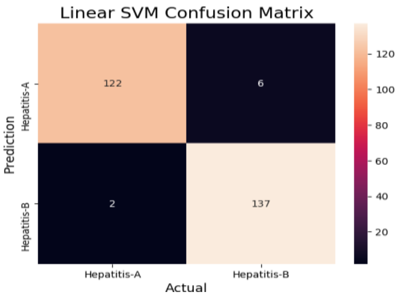


Fig 9 Linear SVM Confusion matrix

Table 3 SVM model Classification report

Metric	Hepatitis-A	Hepatitis-B	Macro Avg	Weighted Avg
Accuracy	0.970	0.970	0.970	0.970
Precision	0.970	0.960	0.970	0.970
Recall	0.950	0.990	0.970	0.970
F1-Score	0.970	0.970	0.970	0.970

Table 3 is the classification report for the Linear SVM model. It offers precision, recall, and F1-score metrics for each category, along with their averages.

Conclusion

Finally, the proposed method attempted to use a variety of machine learning approaches to predict hepatitis C using standard and low-cost blood tests. Our findings showed that SVM and logistic regression approaches are highly accurate in diagnosing hepatitis C in its early stages. However, our study has certain limitations, such as the use of small datasets, a lack of clinical data, and the absence of a trial. In the future, we hope to include more hepatitis C-related features into machine learning techniques to make them more trustworthy and efficient. Furthermore, we advocate performing a clinical trial to test the effectiveness of these strategies in real-world situations. Overall, the suggested approach yields encouraging findings for early detection and diagnosis of hepatitis C using machine learning approaches, potentially improving patient outcomes and saving human lives.

REFERENCES

[1] Ajuwon, Busayo I., et al. "Clinical Validity of a Machine Learning Decision Support System for Early Detection of Hepatitis B Virus: A Binational External Validation Study." *Viruses* 15.8 (2023): 1735.

[2] Alizargar, Azadeh, Yang-Lang Chang, and Tan-Hsu Tan. "Performance comparison of machine learning approaches on Hepatitis C prediction employing data mining techniques." Bioengineering 10.4 (2023): 481.

[3] Alotaibi,Abrar,etal."ExplainableEnsemble-BasedMachineLearningModelsfor Detecting the Presence of Cirrhosis in Hepatitis C Patients." Computation11.6 (2023): 104.

[4] Sachdeva, Ravi Kumar, et al. "A systematic method for diagnosis of hepatitis disease using machine learning." Innovations in Systems and SoftwareEngineering 19.1 (2023): 71-80.

[5] Poongodi S Vani.K, Reddy S Ram Kishore," Cervical Cancer (CC) causes and Identification Techniques", Indian journal of Public Health Research & Development, vol.9, issue 11, 2018,PP.2094-2097.

[6] Swetha, K., et al. "Inflammation of Liver and Hepatitis Disease Prediction using Machine Learning Techniques."2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2023.

[7] Garg, Umang, et al. "Identification and Prediction of Hepatitis Band NAFLD using Machine Learning." 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE, 2023.

[8] Singh, Yuvaraj, et al. "Detection of colorectala denomas using artificial intelligence models in patients with chronic hepatitis C." World Journal of Hepatology15.1 (2023): 107.

[9] Dr.S. Poongodi, Basawakumar,K. Vani, M. Vijay Karthik, "An analytical investigation on artificial Intelligence in various applications", Journal of International Pharmaceutical Research, vol.46, issue 2, 2019,PP.256-260.

[10] Jangiti, Jaydev, et al. "Hepatitis C Severity Prognosis: A Machine Learning Approach." Journal of Electrical Engineering & Technology (2023): 1-12.

[11] Ali, Ali Mohd, et al. "Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection." Machines 11.3 (2023): 391.

[12] Harabor, Valeriu, et al. "Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity." International Journal of Environmental Research and Public Health 20.3 (2023): 2380.

[13] Yağanoğlu, M. (2022). Hepatitis C virus data analysis and prediction using machine learning. Data & Knowledge Engineering, 142, 102087.

[14] Konerman, Monica A., Lauren A. Beste, Tony Van, Boang Liu, Xuefei Zhang, Ji Zhu, Sameer D. Saini et al. "Machine learning models to predict disease progression among veterans with hepatitis C virus." PloS one 14, no. 1 (2019): e0208141.

[15] S Poongodi, B Kalaavathi ,” Comparative study of various transformations in robust watermarking algorithms”, vol.45, issue 11, 2012.

[16] Kaunang, F. J. (2022). A Comparative Study on Hepatitis C Predictions Using Machine Learning Algorithms. 8ISC Proceedings: Technology, 33-42.

[17] Ahammed, K., Satu, M. S., Khan, M. I., & Whaiduzzaman, M. (2020, June). Predicting infectious state of hepatitis c virus affected patient's applying machine learning methods. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 1371-1374). IEEE.

[18] Edeh, M. O., Dalal, S., Dhaou, I. B., Agubosim, C. C., Umoke, C. C., Richard-Nnabu, N. E., & Dahiya, N. (2022). Artificial intelligence-based ensemble learning model for prediction of hepatitis C disease. Frontiers in Public Health, 10, 892371.

[19] Ehsan Bazgir, Ehteshamul Haque, Md. Maniruzzaman and Rahmanul Hoque, “Skin cancer classification using Inception Network”, World Journal of Advanced Research and Reviews, 2024, 21(02), 839–849.

[20] Rahmanul Hoque, Suman Das, Mahmudul Hoque and Ehteshamul Haque, “Breast Cancer Classification using XGBoost”, World Journal of Advanced Research and Reviews, 2024, 21(02), 1985–1994

[21] Fazakis, N.; Kocsis, O.; Dritsas, E.; Alexiou, S.; Fakotakis, N.; Moustakas, K. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access 2021, 9, 103737–103757.

[22] Dritsas, E.; Trigka, M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. Sensors 2022, 22, 5304.

[23] S. Poongodi, Dr.B. Kalaavathi, "Data Hiding in Watermarking Technique with Effective Key Length Using Spread Spectrum Technique” Australian Journal of Basic and Applied Sciences, September 2014, PP.100-105.

[24] S. Poongodi, M. Shanmugapriya, Dr. B. Kalavathi,” Secure Transformation of Data in Encrypted Image Using Reversible Data hiding Technique ” International Journal of Engineering Science and Innovative Technology (IJESIT), July 2013,PP.45-50.