OPEN ACCESS

International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 2 (2025) https://doi.org/10.58723/ijaaiml.v2i2.419

Check for updates

Research Article

Hepatitis Disease Prediction Using Convolutional Neural Network Algorithm in Machine Learning Technology

Received: May 08, 2025	Revised: June 20, 2025	Accepted: July 04, 2025	Publish: July 04, 2025

Ranga Swamy Sirisati*, B. Jayasri, A. Avanthi, A. Ramyasri, K. Sowmya

Abstract:

Background of Study: Hepatitis is a significant viral infection causing liver inflammation, potentially leading to hepatocyte death and impaired liver function. Types B (HBV) and C (HCV) can cause chronic hepatitis, cirrhosis, and cancer. Globally, around 257 million people are infected with HBV and 71 million with HCV. Early detection of chronic Hepatitis B is crucial for effective management.

Aims and Scope of Paper: This study aims to predict hepatitis progression in patients from their medical histories. It seeks to enhance prediction accuracy by addressing challenges like noise and inefficiency caused by similar aspect values and distributions within datasets.

Methods: Machine learning, a branch of AI, is employed for chronic disease prediction. The study primarily utilizes the K-Nearest Neighbour (KNN) algorithm to predict and eliminate redundant data and noise. Other models evaluated include Logistic Regression, Random Forest, and Convolutional Neural Networks (CNN), with SMOTE used for dataset balancing.

Result: KNN achieved 0.970 accuracy, Logistic Regression 0.966, and Random Forest 0.95. The CNN model demonstrated exceptional performance, reaching 1.0 accuracy with perfect precision, recall, and F1-score for Hepatitis A and B.

Conclusion: While KNN performed well among traditional methods, deep learning models like CNN show superior accuracy and generalizability, offering a robust framework for hepatitis prediction.

Keywords: Liver diseases Instruments, Machine learning algorithms, Machine learning Tools, Neural networks, Support vector machine classification.

1. INTRODUCTION

Hepatitis, which is mostly brought on by viral infections like Hepatitis A, B, C, D, and E, is a serious worldwide health issue that is defined by liver inflammation (Mancinelli et al., 2020). The most well-known of these are Hepatitis B (HBV) and Hepatitis C (HCV), which can result in chronic liver disease and serious side effects such cirrhosis and hepatocellular cancer (Abdelhamed & El-Kassas, 2024). According to estimates from the World Health Organization (WHO), there are around 71 million people with chronic HCV infection (Vo Quang et al., 2021) and over 350 million people with chronic HBV infection (Gautam, 2018).

Effective management of many diseases depends on early identification and prompt action; yet, standard diagnostic techniques may be constrained by issues including speed, accuracy, and accessibility.

a. Types of Hepatitis

Each of the five primary forms of viral hepatitis is brought on by a distinct virus:

1. Hepatitis A (HAV): Primarily transmitted through contaminated food and water, it usually results in acute illness and does not lead to chronic infection (Migueres et al., 2021).
2. Hepatitis B (HBV): This form of the virus, which is contracted by coming into touch with infected bodily fluids (e.g., blood, semen), can develop into a chronic illness and cause serious side effects like liver cancer and cirrhosis (Pattyn et al., 2021).
3. Hepatitis C (HCV): Mostly spread via blood-to-blood contact, this disease is a major cause of liver transplants and frequently results in chronic infections (Morozov & Lagaye, 2018).
4. Hepatitis D (HDV): This virus only infects those already infected with HBV and can exacerbate the severity of the disease. Co-infection with HBV and HDV often leads to more severe outcomes (Mathur et al., 2024).

5. Hepatitis E (HEV): Typically spread through contaminated water, it is more common in developing countries and usually causes acute illness, though chronic cases can occur in immunocompromised individuals (Castagna et al., 2024).

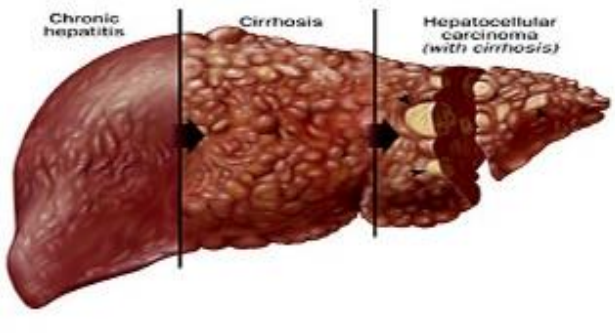

Figure 1. Hepatitis Disease

Machine learning (ML), a subset of Artificial Intelligence (AI), has emerged as a robust tool in health informatics for the prediction and analysis of chronic diseases. They offer innovative solutions for disease prediction and diagnosis. By leveraging vast amounts of clinical data, ML algorithms can identify complex patterns and correlations that may not be apparent through conventional statistical approaches. This capability is particularly beneficial in the context of hepatitis, where early identification of at-risk individuals can lead to improved treatment outcomes and reduced transmission rates.

b. Recent Work

Nonetheless, challenges are encountered owing to the presence of a substantial number of samples with comparable aspect values and distributions within the dataset, resulting in noise and inefficiency. These similar aspect values result in a chaotic and ineffective record set, a lot of undesired issues, and noise. Based on the supplied data, to find the literature gap, we may focus on these areas: Numerous machine learning approaches, including Support Vector Machines (SVM), ensemble methods, and deep learning models (CNN), have been the focus of research on the prediction of hepatitis illness. Performance measures such as accuracy, sensitivity, specificity, F1-score, and AUC have been published for each research.

Recent research has explored various machine learning approaches for hepatitis disease prediction. (Tun et al., 2024) developed a hybrid model that combined random forest and SVM classifiers. The study's hepatitis detection accuracy was 98%. (Prakash et al., 2023) putting forward a convolutional neural network (CNN)-based deep learning method that produced 98.34% accuracy, 99.72% sensitivity, and 97.84% recall. (Modhugu, 2023) applied transfer learning using a pre-trained ResNet50 model that has been refined on an actual hepatitis dataset, reporting 80.7% accuracy. Furthermore, (Ajuwon et al., 2023) investigated the clinical validity of a machine learning decision support system for early detection of Hepatitis B Virus. (Alizargar et al., 2023) conducted performance comparisons of machine learning approaches on Hepatitis C prediction employing data mining techniques. (Alotaibi et al., 2023) developed explainable ensemble-based machine learning models for identifying cirrhosis in patients with Hepatitis C.

Despite these advancements, several limitations remain in the current research.

1. Underexplored Techniques or Models: While traditional ML and CNNs dominate, advanced models like transformers or federated learning for hepatitis prediction remain underexplored.
2. Dataset Limitations: Most studies rely on limited datasets such as UCI Hepatitis Dataset, lacking diverse and large-scale real-world data for better generalization.
3. Hardware Constraints: There is limited work on optimizing models for resource-constrained environments or real-time predictions on edge devices.
4. Evaluation Metrics: Few studies explore comprehensive metrics like F1-score, Matthews correlation coefficient (MCC), or AUC-ROC for a holistic assessment.
5. Practical Deployment: Studies focusing on deployment in real-world clinical settings or low-resource environments remain sparse.
6. Comparative Studies: Lack of head-to-head comparisons of different ML models under similar conditions limits benchmarking and standardization.

2. MATERIAL AND METHOD

Algorithm 1: Enhanced Data Preprocessing for Hepatitis Disease Prediction

Objective: Improve the quality of input data for more accurate hepatitis prediction using ML models.

Step-by-Step Procedure:

Data Collection: Data will be obtained from prominent public datasets, such as the UCI Hepatitis Dataset, and considered for integration with additional clinical data from hospitals, subject to applicable ethical approvals.

To ensure model validity and generalization, cross-validation will be systematically applied. Furthermore, the potential use of larger and more diverse datasets will be explored to enhance model robustness.

Data Cleaning and Imputation: Missing values will be handled using a combination of regression-based imputation (e.g., *MissForest* or *MICE*) for numerical data and mode imputation for categorical data. This approach is chosen to minimize bias and preserve the original data distribution. Outlier detection techniques such as the Z-score or IQR will be employed, and outliers will be analyzed to determine whether to remove, transform, or handle them with other robust methods (Mello-Román & Martínez-Amarilla, 2025).

Feature Selection: Feature selection will involve extensive analysis using *Recursive Feature Elimination (RFE)* combined with filter-based methods such as Chi-square tests or *mutual information* to identify the most relevant features and reduce data dimensionality. This hybrid approach aims to improve model performance and reduce *overfitting* (Priyatno & Widiyaningtyas, 2024).

Normalization: Numerical features will be normalized using *Min-Max scaling* to transform the data into a [0, 1] range, which is crucial for distance-based models like KNN and Neural Networks (Protić et al., 2023).

$$X' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

For data with non-normal distributions, transformations such as *log transformation* or *Box-Cox transformation* will be considered to meet model assumptions and enhance performance.

Outlier Detection and Removal: Identify and remove outliers using z-score or IQR methods.

Feature Engineering: Derivation of meaningful features, such as liver enzyme ratios (e.g., ALT/AST), and incorporation of domain knowledge for new feature creation will be applied.

Data Augmentation (for imbalanced datasets): To address the issue of *imbalanced datasets* common in medical data, oversampling techniques like SMOTE (*Synthetic Minority Over-sampling Technique*) will be applied to generate synthetic samples for the minority class. Additionally, undersampling techniques such as *NearMiss* may also be explored to effectively balance class distribution, ensuring the model is not biased towards the majority class.

Output Processed Data: The cleaned and enhanced dataset will be saved for input into *machine learning* models.

Algorithm 2: Optimization of AI Models for Hepatitis Disease Prediction

Objective: Enhance model accuracy through hyperparameter tuning and architecture refinement.

Step-by-Step Procedure:

Model Selection: The *Machine Learning* models to be evaluated include *Logistic Regression*, *Random Forest*, *K-Nearest Neighbors (KNN)*, and *Convolutional Neural Network (CNN)*. The selection of these models is based on their proven performance in disease prediction studies, considering their ability to handle structured and (potentially) image data. A focus will be placed on *Ensemble Models* (such as *Gradient Boosting* or *AdaBoost*) and *Deep Learning* (CNN), as recent studies demonstrate superior accuracy in hepatitis prediction tasks.

Hyperparameter Tuning: Comprehensive *hyperparameter tuning* will be performed using *Grid Search* or *Random Search* for classical models and *Bayesian Optimization* for *deep learning* models to find the optimal configuration. This includes critical parameters such as learning rate, tree depth, number of estimators, and regularization parameters. The use of *k-fold cross-validation* will be integrated during the *tuning* process to ensure the robustness and generalization of the discovered models.

Model Training: The dataset will be split into 80% for training and 20% for validation. *Gradient Descent* will be used for model optimization.

Evaluation Metrics: Model performance will be evaluated using a comprehensive set of metrics including *Accuracy*, *Precision*, *Recall* (*Sensitivity*), *Specificity*, *F1-Score*, and *Area Under the Receiver Operating Characteristic Curve (AUC-ROC)*. These metrics will provide a holistic view of the model's ability to predict hepatitis, especially considering the potential for class imbalance in the dataset.

Model Comparison: Performance comparisons between different ML models will be conducted to select the best-performing model based on evaluation metrics.

Explainability: To enhance model interpretability, *explainable AI (XAI)* techniques such as SHAP (*SHapley Additive exPlanations*) or LIME (*Local Interpretable Model-agnostic Explanations*) will be employed. This will help in understanding the contribution of each feature to the model's predictions, which is crucial in a medical context to gain trust from clinicians.

Deployable Model: The final model will be optimized for computational efficiency and scalability, enabling implementation in resource-constrained environments or *edge devices*. A Flask-based web application will be developed for demonstration, allowing users to upload data and receive predictions in *real-time*. This process will include converting categorical columns to numerical and handling missing values as described in the *preprocessing* step.

Convolutional Neural Networks (CNNs)

One specific kind of deep learning model that is frequently utilized for tasks involving picture data is the Convolutional Neural Network (CNN) (Taye, 2023). A CNN model could help by automatically learning complex patterns in patient data, such as lab test results

or even medical imaging (e.g., liver scans), to predict the likelihood of the disease (Salehi et al., 2023). CNNs are typically composed of layers that automatically learn hierarchical feature representations from the input data (AYENI, 2022).

Technologies Involved in CNN for Hepatitis Prediction

1. Deep Learning Frameworks: TensorFlow or PyTorch are used to build, train, and deploy CNN models.
2. Data Preprocessing Tools: OpenCV or PIL are used for processing medical imaging data (like liver scans) before feeding them into the CNN.
3. Transfer Learning: Pre-trained CNN models (e.g., ResNet, VGG, or Inception) trained on large datasets (like ImageNet) can be fine-tuned for medical applications, including hepatitis prediction, especially for imaging data.
4. GPU Acceleration: NVIDIA CUDA and cuDNN libraries enable faster training of CNNs on large datasets by utilizing GPU power for parallel processing.
5. Activation Functions: The most popular activation function in CNNs is called ReLU (Rectified Linear Unit), which adds non-linearity to the model so that it can recognize intricate patterns.

Experimental Setup

An experimental setup provides a systematic and reproducible methodology for evaluating ML models in predicting hepatitis disease. Below is a comprehensive plan:

a. Dataset Preparation
   1. Data Source: Use publicly available datasets, such as the UCI Hepatitis Dataset or hospital-specific records (subject to ethical approval). Include structured patient data (e.g., liver enzyme levels, demographics, viral loads) or medical imaging data (e.g., liver scans).
   2. Data Description: Features: Age, gender, liver function test results (e.g., ALT, AST, bilirubin levels), viral load, and clinical observations. Target Variable: Hepatitis status (e.g., healthy, infected, stage progression).
   3. Data Preprocessing: Handle missing values using imputation techniques (mean/mode, predictive modeling), Normalize numerical data using Min-Max scaling or Z-score normalization.

b. Experimental Design
   1. Machine Learning Models: Evaluate the performance of multiple ML models: Baseline Models: Logistic Regression, Decision Trees, SVM. Deep Learning Models: CNN for imaging or structured data.
   2. Train-Test Splits. 80% of the data is in the training set. Test Set: 20% of the data (reserved for the last assessment).

c. Model Training
   1. Feature Selection: Use Recursive Feature Elimination (RFE) or statistical tests (e.g., Chi-square) to select relevant features.
   2. Hyperparameter Tuning: Apply Grid Search or Random Search for models like SVM (kernel selection, C, gamma), Random Forest (number of trees, max depth), and deep learning models (learning rate, number of layers).
   3. Training Protocols: Use Gradient Descent for model optimization.

d. Evaluation Metrics
   1. Accuracy: The percentage of cases (patients with or without hepatitis) that were accurately predicted relative to all occurrences.
   2. Sensitivity (Recall): The model's capacity to accurately detect true positives, or those having a hepatitis diagnosis
   3. Precision: The percentage of actual positive results among all cases that were anticipated to be positive. Precision gauges the accuracy of the model's optimistic forecasts.
   4. F1-Score: The harmonic mean of precision and sensitivity. It provides a balanced measure of the model's performance, particularly in cases of imbalanced datasets.

3. RESULTS AND DISCUSSION

3.1 Result

The comprehensive data analysis and machine learning workflow designed for predicting hepatitis based on a dataset containing various health-related attributes and lab test results. The project begins by importing essential Frameworks. The dataset, stored in 'hcvdat.csv', is loaded into a DataFrame, followed by exploratory data analysis (EDA) to understand its structure, class distribution, and feature relationships through visualizations like plots and summary statistics.

Data preprocessing steps include converting categorical columns like 'Category' and 'Sex' into numeric values and handling missing values by filling them with zeros. Feature engineering involves binning continuous variables such as 'Age', 'ALB', and 'ALT' into discrete categories to enhance model performance.

Three machine learning models Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN) are trained and evaluated based on accuracy and confusion matrix metrics. The Random Forest model also includes a feature importance analysis to identify key predictors. Finally, a Flask web application is set up to deploy the model, allowing users to upload CSV files for predictions. The app processes the data, generates predictions, and displays results via an HTML template.

The dataset features include health metrics like Albumin (ALB), Bilirubin (BIL), and Cholesterol (CHOL), with 'Category' as the target variable indicating hepatitis classification. This end-to-end pipeline demonstrates a structured approach to predictive modeling, from data

preprocessing and EDA to model deployment, ensuring accurate and interpretable results for hepatitis prediction.

Figure 2 represents the confusion matrix for the Logistic Regression (LR) model. It breaks down the predictions made by the model, comparing them to the actual Figure 2 represents the confusion matrix for the Logistic Regression (LR) model. It breaks down the predictions made by the model, comparing them to the actual

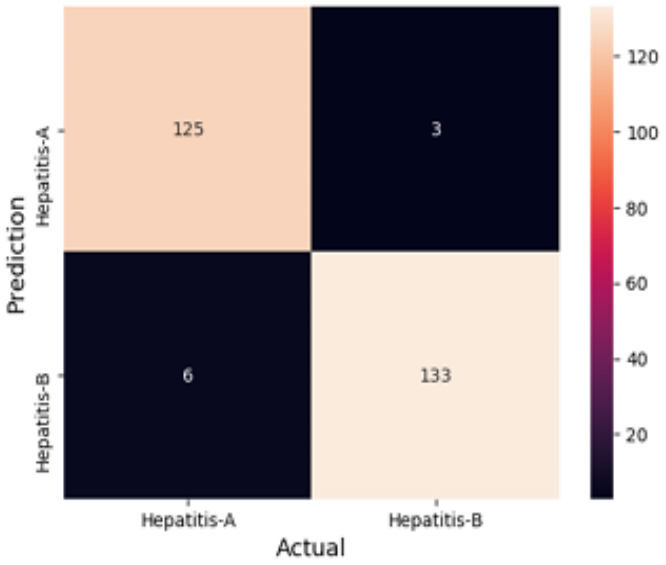

Figure 2. LR Confusion Matrix

Table 1 is a detailed classification report for the LR model. It provides Metrices 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. Classification Report for the LR Model.

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

Similar to the LR confusion matrix, this Figure 10.8 presents the confusion matrix for the Random Forest (RF) model. It includes accuracy as 0.95, indicating a high accuracy level. The classification report provides further details about precision, recall, and F1-score.

Figure 2 represents the confusion matrix. The accuracy of the KNN model is given as 0.970, and the subsequent classificatio n report.

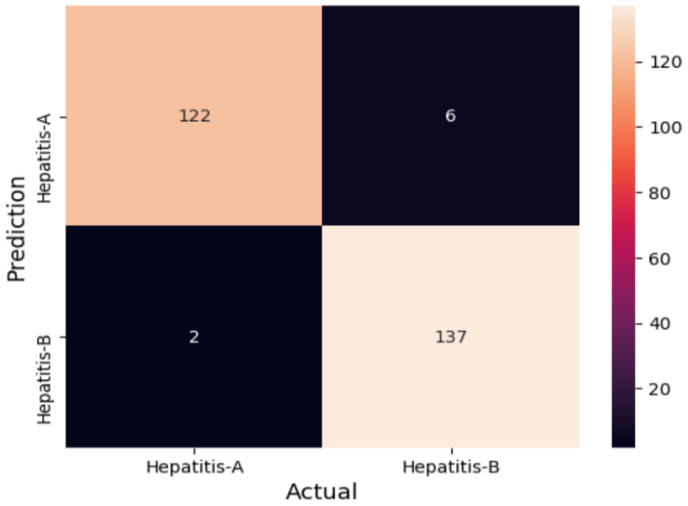

Figure 3. KNN Model Classification Report

Table 2 is the classification report for the Linear KNN model. It offers precision, rec all, and F1-score metrics for each category, along with their averages.

Table 2. Classification Report For The Linear KNN Model

Metrics	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

Figure 3. represents the confusion matrix for the Convolution Neural Network (CNN) model. It breaks down the predictions made by the model, comparing them to the actual Figure 3 represents the confusion matrix for the Convolution Neural Network (CNN) model. It breaks down the predictions made by the model, comparing them to the actual.

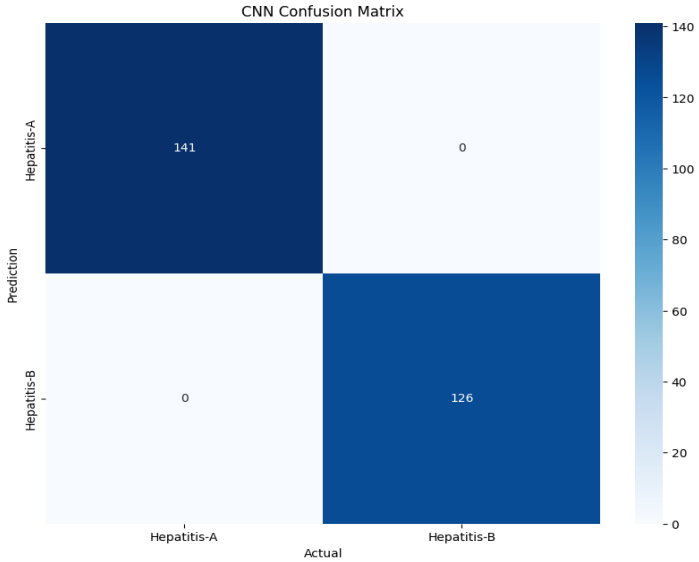

Figure 4. CNN Model Classification Report

Table 3 is a detailed classification report for the Convolution Neural Network model. They give for each category (Hepatitis-A and Hepatitis-B), along with their weighted average and macro average.

Table 3. Classification Report For The Convolution Neural Network Model

Class	Accuracy	Precision	Recall	F1-Score	Support
Hepatitis A	1.0	0.99	1.0	1.0	141
Hepatitis B	1.0	1.0	0.99	1.0	126
Macro Avg	1.0	1.0	1.0	1.0	267
Weighted Avg	1.0	1.0	1.0	1.0	267

3.2 Discusion

This study embarked on a comprehensive exploratory data analysis (EDA) of a hepatitis dataset, employing the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and ensure fair representation across classes. Three machine learning models—Logistic Regression (LR), Random Forest (RF), and K-Nearest Neighbors (KNN)—were developed and evaluated for their effectiveness in predicting hepatitis. The performance of these models was assessed using key metrics such as accuracy and confusion matrices.

The results indicate that the K-Nearest Neighbors (KNN) classifier demonstrated the highest accuracy among the evaluated models on the test dataset, achieving an accuracy of 0.970, with corresponding precision, recall, and F1-scores of 0.970 for both Hepatitis A and Hepatitis B (Table 2). This is further supported by its confusion matrix (Figure 2), which shows 122 correct predictions for Hepatitis-A and 137 for Hepatitis-B, with only minimal misclassifications.

In comparison, the Logistic Regression (LR) model yielded an overall weighted average accuracy of 0.966 (Table 1). Its confusion matrix (Figure 1) shows 125 correct predictions for Hepatitis-A and 133 for Hepatitis-B. The Random Forest model achieved an accuracy of 0.95, as indicated by its confusion matrix (Figure 10.8). While these models also performed well, KNN exhibited a slight edge in overall accuracy.

Furthermore, a Convolutional Neural Network (CNN) model was also evaluated, demonstrating exceptional performance. The CNN model achieved an accuracy of 1.0, with precision, recall, and F1-scores of 1.0 for both Hepatitis A and Hepatitis B (Table 3). Its confusion matrix (Figure 3) shows perfect classification, with 141 correct predictions for Hepatitis-A and 126 for Hepatitis-B, and no misclassifications. This suggests that while KNN performed best among the traditional machine learning models, deep learning models like CNN can offer even higher accuracy and generalizability.

Feature importance analysis was conducted for the KNN model to identify the most influential attributes in predicting hepatitis, providing valuable insights into the key factors driving the predictions. This comprehensive framework for hepatitis prediction provides a strong foundation, though further model tuning and validation may be necessary for seamless clinical deployment. Integrating domain-specific expertise can further enhance the interpretability and practical utility of these predictive models

3.2.1 Implications

The study's findings have significant implications for the early detection and management of hepatitis. The high accuracy achieved by the K-Nearest Neighbors (KNN) classifier and especially the Convolutional Neural Network (CNN) model suggests that machine learning can be a powerful tool in clinical settings for predicting hepatitis. Early identification of at-risk individuals can lead to improved treatment outcomes and reduced transmission rates, addressing a critical global health issue. The use of techniques like SMOTE for addressing class imbalance ensures that the models are reliable even with skewed medical datasets. Furthermore, the framework's comprehensive approach, from data preprocessing to model deployment via a Flask web application, demonstrates a pathway for practical integration into healthcare systems for real-time predictions. The inclusion of explainable AI (XAI) techniques like SHAP or LIME also contributes to building trust among clinicians by providing insights into feature contributions to predictions.

3.2.2 Research contribution

This research contributes to the field of health informatics by presenting a comprehensive machine learning workflow for hepatitis prediction, addressing challenges such as noisy datasets and class imbalance. The study systematically evaluates the performance of

Logistic Regression, Random Forest, and K-Nearest Neighbors, demonstrating KNN's superior accuracy among traditional models with a 0.970 accuracy, and highlighting the exceptional performance of a Convolutional Neural Network (CNN) model, which achieved perfect classification with an accuracy of 1.0. The integration of feature engineering, data augmentation techniques like SMOTE, and robust evaluation metrics provides a structured approach to predictive modeling for medical data. Additionally, the work outlines a plan for model optimization, hyperparameter tuning using k-fold cross-validation, and the development of a deployable model with explainability features, advancing the readiness of such systems for real-world clinical application.

3.2.3 Limitations

Despite the advancements, the current research faces several limitations. Most studies, including this one, tend to rely on limited datasets, such as the UCI Hepatitis Dataset, which may lack the diversity and large scale of real-world data necessary for better generalization. There is also an underexplored area for advanced models like transformers or federated learning in hepatitis prediction. Furthermore, the optimization of models for resource-constrained environments or real-time predictions on edge devices is not extensively covered. While various evaluation metrics are used, some studies may still lack comprehensive metrics like F1-score, Matthews correlation coefficient (MCC), or AUC-ROC for a holistic assessment across all models. Finally, studies focusing on practical deployment in real-world clinical settings or low-resource environments remain sparse, and there's a general lack of head-to-head comparisons of different ML models under similar conditions, limiting benchmarking and standardization.

3.2.4 Suggestions

To further enhance the predictive models for hepatitis, several suggestions can be considered. Future research should prioritize acquiring and integrating larger and more diverse real-world datasets from various geographical regions and demographics to improve model robustness and generalization. Exploring advanced machine learning techniques, such as transformer networks or federated learning, could potentially yield even higher accuracies and address privacy concerns with distributed data. Moreover, optimizing models for resource-constrained environments and edge devices is crucial for practical deployment in diverse clinical settings. Implementing and evaluating a wider array of comprehensive evaluation metrics, including F1-score, MCC, and AUC-ROC across all models, would provide a more complete assessment of their performance. Finally, greater emphasis should be placed on conducting comparative studies of different machine learning models under standardized conditions and focusing on the seamless integration and validation of these models within actual clinical workflows, potentially incorporating domain-specific expertise to enhance interpretability and practical utility.

4. CONCLUSIONS

This research conducted a comprehensive exploratory data analysis (EDA) on a hepatitis dataset and utilized the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. Three machine learning models Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN) were developed and evaluated for their effectiveness in predicting hepatitis. Model performance metrics, including accuracy and confusion matrices, were calculated and examined.

The KNN classifier demonstrated the highest accuracy among the evaluated models on the test dataset. It achieved an accuracy of 0.970, with corresponding precision, recall, and F1-scores of 0.970 for both Hepatitis A and Hepatitis B. The KNN confusion matrix showed 122 correct predictions for Hepatitis-A and 137 for Hepatitis-B, with only minimal misclassifications.

In comparison, the Logistic Regression (LR) model yielded an overall weighted average accuracy of 0.966. Its confusion matrix showed 125 correct predictions for Hepatitis-A and 133 for Hepatitis-B. The Random Forest model achieved an accuracy of 0.95. While these models performed well, KNN exhibited a slight edge in overall accuracy.

Furthermore, a Convolutional Neural Network (CNN) model was also evaluated, demonstrating exceptional performance. The CNN model achieved an accuracy of 1.0, with precision, recall, and F1-scores of 1.0 for both Hepatitis A and Hepatitis B. Its confusion matrix showed perfect classification, with 141 correct predictions for Hepatitis-A and 126 for Hepatitis-B, and no misclassifications. This suggests that while KNN performed best among traditional machine learning models, deep learning models like CNN can offer even higher accuracy and generalizability.

Feature importance analysis was conducted for the KNN model to identify the most influential attributes in predicting hepatitis, providing valuable insights into the key factors driving the predictions. This comprehensive framework for hepatitis prediction provides a strong foundation, though further model tuning and validation may be necessary for seamless clinical deployment. Integrating domain-specific expertise can further enhance the interpretability and practical utility of these predictive models.

5. ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to everyone who supported and guided us throughout the completion of our project titled "Hepatitis Disease Prediction using Machine Learning Technology." We are especially thankful for the expert guidance, continuous encouragement, and valuable feedback, which were instrumental in shaping this work. We also extend our

sincere thanks to our co-authors for their collaboration, support, and dedication throughout the project. This work explores the application of advanced machine learning models, including Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), for the early detection and prediction of hepatitis. We acknowledge the immense contribution of these algorithms in improving diagnostic accuracy and supporting effective clinical decision-making.

6. AUTHOR CONTRIBUTION STATEMENT

RS, BJ, AA, AR, and KS conceptualized the study and its methodology. RS, BJ, AA, AR, and KS contributed to the investigation and data analysis. RS supervised the project. All authors contributed to the writing, review, and editing of the manuscript.

AUTHOR INFORMATION

Corresponding Authors

Ranga Swamy Sirisati, Department of Information Technology, Vignan's Institute of Management and Technology for Women, Ghatkesar, Medchal-Malkajgiri, Telangana , India
https://orcid.org/0000-0002-3104-6672
Email: sirisatiranga@gmail.com

Authors

B. Jayasri, Department of Information Technology, Vignan's Institute of Management and Technology for Women, Ghatkesar, Medchal-Malkajgiri, Telangana, India
https://orcid.org/0009-0009-9042-4133
Email: jayasrikrishna2003@gmail.com

A. Avanthi, Department of Information Technology, Vignan's Institute of Management and Technology for Women, Ghatkesar, Medchal-Malkajgiri, Telangana, India
https://orcid.org/0009-0001-3747-1872
Email: avaduthwaravanthi@gmail.com

A. Ramyasri, Department of Information Technology, Vignan's Institute of Management and Technology for Women, Ghatkesar, Medchal-Malkajgiri, Telangana, India.
https://orcid.org/0009-0003-0599-5924
Email: ayitiramyasri@gmail.com

K. Sowmya, Department of Information Technology, Vignan's Institute of Management and Technology for Women, Ghatkesar, Medchal-Malkajgiri, Telangana, India.
https://orcid.org/0009-0009-4521-2043
Email: sowmyakasparaju@gmail.com

REFFERENCE

Abdelhamed, W., & El-Kassas, M. (2024). Hepatitis B virus as a risk factor for hepatocellular carcinoma: There is still much work to do. *Liver Research*, *8*(2), 83–90. https://doi.org/10.1016/j.livres.2024.05.004

Ajuwon, B. I., Richardson, A., Roper, K., & Lidbury, B. A. (2023). Clinical Validity of a Machine Learning Decision Support System for Early Detection of Hepatitis B Virus: A Binational External Validation Study. *Viruses*, *15*(8). https://doi.org/10.3390/v15081735

Alizargar, A., Chang, Y. L., & Tan, T. H. (2023). Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques. *Bioengineering*, *10*(4). https://doi.org/10.3390/bioengineering10040481

Alotaibi, A., Alnajrani, L., Alsheikh, N., Alanazy, A., Alshammasi, S., Almusairii, M., Alrassan, S., & Alansari, A. (2023). Explainable Ensemble-Based Machine Learning Models for Detecting the Presence of Cirrhosis in Hepatitis C Patients. *Computation*, *11*(6). https://doi.org/10.3390/computation11060104

AYENI, J. A. (2022). Convolutional Neural Network (CNN): The architecture and applications. *Applied Journal of Physical Science*, *4*(4), 42–50. https://doi.org/10.31248/ajps2022.085

Castagna, F., Liguori, G., Lombardi, R., Bava, R., Costagliola, A., Giordano, A., Quintiliani, M., Giacomini, D., Albergo, F., Gigliotti, A., Lupia, C., Ceniti, C., Tilocca, B., Palma, E., Roncada, P., & Britti, D. (2024). Hepatitis E and Potential Public Health Implications from a One-Health Perspective: Special Focus on the European Wild Boar (Sus scrofa). *Pathogens*, *13*(10). https://doi.org/10.3390/pathogens13100840

Gautam, P. K. (2018). Senerio of Sero-Prevalence of Hepatitis B Infection in Rular Area in East Uttar Pradesh: A Hospital Based Study. *Journal of Medical Science And Clinical Research*, *6*(11), 311–315. https://doi.org/10.18535/jmscr/v6i11.55

Mancinelli, R., Rosa, L., Cutone, A., Lepanto, M. S., Franchitto, A., Onori, P., Gaudio, E., & Valenti, P. (2020). Viral hepatitis and iron dysregulation: Molecular pathways and the role of lactoferrin. *Molecules*, *25*(8), 1–21. https://doi.org/10.3390/molecules25081997

Mathur, P., Khanam, A., & Kottilil, S. (2024). Chronic Hepatitis D Virus Infection and Its Treatment: A Narrative Review. *Microorganisms*, *12*(11). https://doi.org/10.3390/microorganisms12112177

Mello-Román, J. D., & Martínez-Amarilla, A. (2025). COVID-19 Data Analysis: The Impact of Missing Data Imputation on Supervised Learning Model Performance. *Computation*, *13*(3), 2–23. https://doi.org/10.3390/computation13030070

Migueres, M., Lhomme, S., & Izopet, J. (2021). Hepatitis A: Epidemiology, high-risk groups, prevention and research on antiviral treatment. *Viruses*, *13*(10), 1–12. https://doi.org/10.3390/v13101900

Modhugu, V. R. (2023). Efficient Hybrid CNN Method to Classify the Liver Diseases. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *14*(3), 36–47. https://doi.org/10.58346/JOWUA.2023.I3.004

Morozov, V. A., & Lagaye, S. (2018). Hepatitis C virus: Morphogenesis, infection and therapy. *World Journal of Hepatology*, *10*(2), 186–212. https://doi.org/10.4254/wjh.v10.i2.186

Pattyn, J., Hendrickx, G., Vorsters, A., & Van Damme, P. (2021). Hepatitis B Vaccines. *Journal of Infectious Diseases*, *224*(Suppl 4), S343–S351. https://doi.org/10.1093/infdis/jiaa668

Prakash, N. N., Rajesh, V., Namakhwa, D. L., Dwarkanath Pande, S., & Ahammad, S. H. (2023). A DenseNet CNN-based liver lesion prediction and classification for future medical diagnosis. *Scientific African*, *20*. https://doi.org/10.1016/j.sciaf.2023.e01629

Priyatno, A. M., & Widiyaningtyas, T. (2024). a Systematic Literature Review: Recursive Feature Elimination Algorithms. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *9*(2), 196–207. https://doi.org/10.33480/jitk.v9i2.5015

Protić, D., Stanković, M., Prodanović, R., Vulić, I., Stojanović, G. M., Simić, M., Ostojić, G., & Stankovski, S. (2023). Numerical Feature Selection and Hyperbolic Tangent Feature Scaling in Machine Learning-Based Detection of Anomalies in the Computer Network Behavior. *Electronics (Switzerland)*, *12*(19). https://doi.org/10.3390/electronics12194158

Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability (Switzerland)*, *15*(7). https://doi.org/10.3390/su15075930

Taye, M. M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. *Computation*, *11*(3). https://doi.org/10.3390/computation11030052

Tun, W., Wong, J. K.-W., & Ling, S.-H. (2024). Hybrid Random Forest and Support Vector Machine Modeling for HVAC Fault Detection and Diagnosis. *Sensors*, *24*(14), 1–15. https://doi.org/10.3390/s21248163

Vo Quang, E., Shimakawa, Y., & Nahon, P. (2021). Epidemiological projections of viral-induced hepatocellular carcinoma in the perspective of WHO global hepatitis elimination. *Liver International*, *41*(5), 915–927. https://doi.org/10.1111/liv.14843