"Predicting Hepatitis B Infection in Early Stages: Identifying Risk Factors and Determining Treatment Outcomes Using Artificial Neural Networks and Data Mining"

Safeena C,

Assistant professor , Department of Computer Applications, Safi college of Advanced Studies (Autonomous), Malappuram
India ,673633

Abstract—**This study explores the early prediction of Hepatitis B infection, focusing on identifying key risk factors and determining treatment outcomes. Using artificial neural networks (ANN) and data mining techniques, we analyzed patient data to develop predictive models. Our approach demonstrates the potential of AI in improving early diagnosis and personalized treatment strategies for Hepatitis B, aiming to enhance patient outcomes and reduce disease progression .Hepatitis B is a viral infection that affects the liver. It can lead to either short-term (acute) or long-term (chronic) health issues. The virus is transmitted through contact with the blood, semen, or other bodily fluids of an infected individual. Epidemiology: Explore the worldwide prevalence and effects of Hepatitis B, emphasizing the risk of developing chronic infections, the progression of liver diseases such as cirrhosis and liver cancer, and the difficulties associated with early detection** Diagnosing **hepatitis B** using an **artificial neural network (ANN)** based expert system and **data mining** techniques is an advanced approach that can significantly improve the accuracy, speed, and reliability of diagnosing this viral infection. Below is an outline of how these techniques are applied in the context of hepatitis B diagnosis :Hepatitis B is a liver infection caused by the **hepatitis B virus (HBV)**. It can lead to chronic disease, liver cirrhosis, and liver cancer if not properly managed. Early diagnosis is essential for effective treatment and to prevent further complications. Diagnosing hepatitis B traditionally involves a combination of **clinical symptoms**, **laboratory tests** (like **HBV DNA levels**, **serological markers**), and **imaging studies**. However, the process is time-consuming and often requires expert interpretation

AI ,ANN ,HBV, classification , clustering *(key words)*

1.INTRODUCTION

Hepatitis B is a viral infection that affects the liver. It can lead to either short-term (acute) or long-term (chronic) health issues. The virus is transmitted through contact with the blood, semen,or other bodily fluids of an infected individual . Explore the worldwide prevalence and effects of Hepatitis B, emphasizing the risk of developing chronic infections, the progression of liver diseases such as cirrhosis and liver cancer, and the difficulties associated with early detection Artificial intelligence (AI) is a new discipline that aims to simulate, extend, and expand human intelligence and integrates theory, method, and application research and development[1]. According to the different algorithms of AI, AI is usually divided into traditional AI (*e.g.*, traditional machine learning) and deep learning (based on neural network structure) in the medical field[2]. Compared with traditional AI, deep learning can directly apply the image to the learning process without manual feature extraction, while traditional machine learning needs manual recognition and extraction of different features. Therefore, deep learning has the advantage of no manual extraction of various features, which makes its learning process faster, intelligent, and accurate. In addition, deep learning can iterate and improve from past mistakes, but it needs more big data and more result analysis to fully show its robust and precise efficiency .This systematic review is intended to ascertain the clinical utility in routine practice of AI based on data mining, predict the incidence of hepatitis A, B, C, or E, categorize the different stages of hepatitis B, predict the progression of hepatitis C in veterans, and diagnose or screen for hepatitis B. Particularly, this review focuses on the application of deep learning or radiomics based on radiology to stage hepatic fibrosis, predict the occurrence of HCC, identify microvascular invasion, and predict the risk of liver failure after hepatectomy in HCC patients.

II Artificial Neural Networks in Healthcare:

Artificial Neural Networks (ANN) are computational models that draw inspiration from the structure and functioning of the human brain. Their proficiency in modeling non-linear relationships renders them exceptional, particularly for complicated tasks. Moreover, ANNs handle high-dimensional data very efficiently, like those often encountered in medical datasets, revealing complex patterns and relationships which might not be identified through traditional statistical methods . The use of ANN in research related to Hepatitis B is warranted since it offers the possibility of improving the accuracy of predictions regarding the risk of infection, timely diagnosis, and treatment outcomes. Because it evaluates different types of data from patients, ANN may detect slight features that indicate disease progression and treatment response; therefore, management and care for patients will be better.

Problem Definition and Objectives:

This study aims to:

- You can forecast the prevalence of Hepatitis B infection: Develop an ANN model for discriminating individuals as high risk or low risk for Hepatitis B depending on their demographic, medical and behavioral factors.
- Significant risk factors identification: Use ANN to determine the factors that are most strongly associated with Hepatitis B infection, such as age, comorbidities, and lifestyle.
- Early diagnosis: It allows the early detection of infection, especially in asymptomatic cases using medical history and laboratory results.
- Predict treatment outcomes: Build a model that predicts treatment response, complication probability, or potential progression to chronic Hepatitis B.

III. Data Collection:

Dataset Selection:

Use a comprehensive dataset that includes:

Demographic information: Age, gender, ethnicity, socio-economic status.

Medical history: Past diagnoses, family history, comorbidities (e.g., HIV, diabetes, liver disease).

Behavioral risk factors: Intravenous drug use, unprotected sex, tattooing, blood transfusions, occupational exposure.

Biomarkers: HBV DNA levels, liver function tests (e.g., ALT, AST), hepatitis B surface antigen (HBsAg), viral load.

Treatment Data: Type of treatment, medication adherence, response to therapy, complications (e.g., cirrhosis, liver cancer).

Data Preprocessing:

Handle missing values using imputation techniques or remove incomplete entries.

Normalize continuous features (e.g., age, viral load) to a consistent scale.

One-hot encode categorical variables (e.g., gender, history of alcohol use).

Split the dataset into training (70%), validation (15%), and testing (15%) sets for model evaluation

IV. Data Mining Techniques

Data mining refers to extracting useful patterns and knowledge from large datasets. In the case of hepatitis B, data mining involves analyzing historical patient data, such as:

- ***Demographic data*** (age, gender, etc.).
- ***Clinical data*** (symptoms, medical history).
- ***Laboratory data*** (HBV surface antigen, HBV DNA, liver enzyme levels, etc.).
- ***Imaging data*** (ultrasound, biopsy results, etc.).

Key data mining techniques for hepatitis B diagnosis include:

- ***Classification***: Categorizing patients into different groups such as infected or not infected, or acute vs. chronic. Common algorithms include decision trees, k-nearest neighbors, and support vector machines (SVMs).
- ***Clustering***: Grouping similar patients together based on shared characteristics. This can help identify subgroups of patients who may require different treatment approaches.
- ***Association rule mining***: Identifying relationships between various variables (e.g., what test results or clinical signs often occur together in hepatitis B patients).
- ***Regression analysis***: Predicting continuous outcomes, such as liver enzyme levels or HBV viral load, based on patient features.
- ***decision tree algorithm*** for the early detection of Hepatitis B using Python and the scikit-learn library. This example assumes you have a dataset with various features (such as demographics, clinical, and laboratory data) and that the goal is to predict whether a patient is positive for Hepatitis B.

V Steps:

- ***Data Preprocessing***: Clean and prepare the dataset (handling missing values, encoding categorical variables, etc.).
- ***Feature Selection***: Choose relevant features to improve model performance.
- ***Model Training***: Train a decision tree classifier on the data.
- ***Evaluation***: Evaluate the model's performance using metrics like accuracy.

Data Loading & Preprocessing:

- The dataset (hepatitis_data.csv) should contain patient data with relevant features like demographics, clinical information, and lab results.
- Missing values are handled by replacing them with the column mean (for numerical data). You can choose other imputation methods as needed.
- Categorical variables (like gender) are encoded using Label Encoder.

Feature Selection & Splitting the Data:

- X represents the features (demographic and clinical data), and y is the target variable indicating the presence of Hepatitis B (0 = negative, 1 = positive).
- The data is split into training (70%) and testing (30%) sets.

Model Training:

- *A* ***Decision Tree*** classifier is trained using the training data (X_ train and y train).

Evaluation:

- The accuracy of the model is calculated using accuracy score, and additional performance details are provided using classification report (which includes precision, recall, and F1-score).

***Visualization (Optional)*:**

The decision tree is visualized using plot tree from learn. This can help interpret how the model makes decisions based on the features.

Example Output:

bash

Copy

```
Accuracy: 85.00%


Classification Report:

         precision    recall  f1-score   support



    0       0.84       0.87       0.85       30

    1       0.86       0.83       0.84       20



  accuracy                            0.85       50

 macro avg       0.85       0.85       0.85       50

weighted avg       0.85       0.85       0.85       50
```

In this case, the accuracy is 85%, meaning the model is correctly identifying Hepatitis B status 85% of the time on the test data. The classification report provides a breakdown of precision, recall, and F1-score for both classes (Hepatitis B positive and negative).

Modifications:

You can experiment with other algorithms such as **Random Forest**, **Support Vector Machine**, or **Logistic Regression** to compare performed.

V. Hepatitis detection based on data mining

No.	Task	Algorithms	Sample size (type)
1	Predicting incidence of hepatitis A	ANN; ARIMA	N/A (CDC data)
2	Predicting incidence of hepatitis B	ARIMA; ElmanNN	486983 cases (data from health commission)
3	Forecasting incidence of hepatitis B	Hybrid method (combing GM and BP-ANN)	10486959 cases (data from health ministry)
4	Prediction of incidence of hepatitis E	ARIMA; SVM; LSTM	N/A (CDC data)
5	Automated classification of the different stages of hepatitis B	ADHB-ML-MFIS expert system	52 patients (serological data)
6	Analyzing HBV infection from normal blood samples	Polynomial function; RBF	119 serum samples from HBV infected patients (Raman spectroscopy data)
7	Rapidly screening hepatitis B from non-hepatitis B	LSTM	1134 blood samples (Raman spectroscopy data)
8	Finding undiagnosed patients with hepatitis C infection	Logistic regression; Gradient boosting trees; Gradient boosting trees with temporal variables; Stacked ensemble; Random forest	9721923 patients (data from the patient's medical history)
9	Predicting hepatitis C virus progression among veterans	CS Cox model longitudinal Cox model; CS boosting model Longitudinal-boosting model	72683 CHC individuals (VHA data)

Classifying different stages of hepatitis

According to the serological level of transaminase and serological titer of hepatitis B antigen or antibody, hepatitis B can be categorized into five stages: I, HBeAg (+) chronic infection; II, HBeAg (+) chronic hepatitis; III, HBeAg (-) chronic infection; IV, HBeAg (-) chronic hepatitis; V, HBsAg (-) stage. For young doctors, it can be difficult to accurately classify the clinical types of hepatitis. Therefore, with 52 patients included, an automated diagnosis of hepatitis B using multilayer mamdani fuzzy inference system expert system (ADHB-ML-MFIS) was established by Ahmad *et al*[16] to categorize the different stages of hepatitis B. In the ADHB-ML-MFIS expert system, there were two levels of input variables. The input variables of the first layer were transaminase, and the input variables of the second layer were serological markers of hepatitis B. The diagnostic accuracy of the ADHB-ML-MFIS expert system for differentiating stages of hepatitis B was 92.2% with the detailed results shown in Table 1.

4. ANN Model Development:

Architecture:

Input Layer: Corresponds to the features of the dataset (e.g., age, viral load, comorbidities).

Hidden Layers: One or more hidden layers, where each layer can have 10–50 neurons depending on data complexity. The number of neurons in each layer can be adjusted based on experimentation.

Output Layer:

For classification (infection prediction): A single neuron with a sigmoid activation function for binary classification (infected vs. non-infected).

For regression (treatment outcomes): A single neuron with a linear activation function for predicting continuous values (e.g., viral load, chances of cirrhosis).

Training the ANN:

Use **ReLU** (Rectified Linear Unit) as the activation function for hidden layers.

Select the **Adam optimizer** for faster convergence.

Choose an appropriate **loss function**:

Binary cross-entropy for classification tasks.

Mean squared error (MSE) for regression tasks.

Train using ***cross-validation*** to ensure that the model generalizes well on unseen data.

Hyperparameter Tuning: Experiment with the number of neurons, learning rate, batch size, and epochs to optimize the model's performance.

VII.Artificial Neural Networks (ANN) to predict early-stage Hepatitis B from a dataset. We'll *use **Keras***, a popular deep learning library built on top of TensorFlow, to create a neural network model.

Steps:

Preprocessing the Data: Handle missing values, encode categorical variables, and scale the features.

Building the ANN Model: Construct the ANN architecture (input layer, hidden layers, and output layer).

Model Compilation: Choose a loss function and optimizer.

Model Training: Train the model on the training data

Evaluation: Assess the model's performance on the test data.

VIII. Data Preprocessing

Missing values: The missing values are filled with the mean of the column.

Categorical encoding: If you have categorical variables like gender, they are encoded into numerical values using LabelEncoder.

Feature scaling: Neural networks perform better when features are on a similar scale, so StandardScaler is used to standardize the features (zero mean, unit variance).

IX Building the ANN:

Input layer: The first layer receives the input features. The input_dim is the number of features (columns) in your dataset.

Hidden layers: You can experiment with the number of hidden layers and neurons in each layer. In this case, we use one hidden layer with 64 neurons and ReLU (Rectified Linear Unit) as the activation function.

Output layer: The output layer has a single neuron because we are performing binary classification (Hepatitis B: Positive/Negative). The sigmoid activation function is used to output probabilities.

***Model Compilation*:**

We use **Adam** optimizer, which is adaptive and works well for many tasks.

Binary cross-entropy is used as the loss function since this is a binary classification problem (predicting presence or absence of Hepatitis B).

***Training the Model*:**

The model is trained for 50 epochs (you can increase or decrease based on performance) with a batch size of 32.

The validation set is provided to evaluate the model during training to ensure it's not overfitting.

X Evaluation:

After training, we make predictions on the test data.

We use accuracy score to compute the accuracy of the model and print a detailed classification report showing precision, recall, F1-score, and support.

Model Evaluation:
Metrics for Classification (infection prediction)*:*

Accuracy: Proportion of correct predictions.

Precision, Recall, and F1-Score: To handle imbalanced datasets (important if one class is underrepresented).

AUC-ROC curve: To evaluate the model's ability to discriminate between the infected and non-infected groups.

Metrics for Regression (treatment outcomes)*:*

Mean Absolute Error (MAE) *and* ***Mean Squared Error (MSE)*** to evaluate the prediction accuracy.

R-squared to assess how well the model explains the variance in the data.

Confusion Matrix: Visualize true positives, false positives, false negatives, and true negatives for classification task

XII. Results and Discussion:
Risk Factors: Interpret the importance of each feature in predicting Hepatitis B infection. For example, age, comorbidities, or specific lab results (e.g., HBsAg status) might emerge as significant predictors.

Model Performance: Compare the ANN model's performance against traditional machine learning models like logistic regression, decision trees, or random forests.

XIII Clinical Implications: Discuss how the ANN model could help healthcare providers:

Early detection of Hepatitis B in asymptomatic patients.

Personalized treatment decisions based on predicted outcomes.

Resource allocation: Target high-risk individuals for further testing and treatment.

Limitations: Discuss potential limitations, such as the quality and size of the dataset, model overfitting, or lack of real-world validation.

XIV Limitations and future perspectives At present, all reviewed studies using AI are mostly single-center research, and the amount of data in the training sets may be insufficient. In the future, people should build an internet database or hospital, and disease-related information from various countries or regions could be uploaded to the internet database or hospital so that AI researchers can obtain more disease-related information with different demographics, geographic areas, *etc.*, to carry out multicenter research, and establish a more robust and inclusive AI model for clinical use.
XV Conclusion:
Summary of Findings: Recap the major results, emphasizing how ANN can enhance the prediction of Hepatitis B infection risk and treatment outcomes.

Implications for Healthcare: Suggest how your ANN-based tool could be integrated into clinical practice for better management of Hepatitis B.

Future Work: Propose future directions, such as:

Improving the model with larger and more diverse datasets.

Exploring the integration of genomic or longitudinal data to refine predictions.

Validating the model in real-world clinical settings.

XVI Journal Selection for Submission:

Some potential journals for publishing your work:
Journal of Hepatology: Focuses on clinical hepatology and liver diseases.

Hepatology: Covers all aspects of liver disease, including new technologies in diagnosis and treatment.

Journal of Medical Systems: Focuses on the application of AI and machine learning in healthcare.

Artificial Intelligence in Medicine: Focuses on the application of AI in medical and healthcare problems.

Computers in Biology and Medi**cine**: Covers computational biology and the integration of AI techniques in healthcare.

XVII Final Thoughts:

This approach should allow you to comprehensively address your research question on Hepatitis B prediction, early diagnosis, and treatment outcomes using ANN. If you need assistance with implementing the ANN model in Python or any other technical details, let me know! I can guide you through the coding process or any specific challenges.

XVIII References:

.Khan S, Ullah R, Khan A, Ashraf R, Ali H, Bilal M, Saleem M. Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. Photodiagnosis Photodyn Ther. 2018;23:89–93. doi: 10.1016/j.pdpdt.2018.05.010. [DOI] [PubMed]

Zheng Y, Zhang L, Zhu X, Guo G. A comparative study of two methods to predict the incidence of hepatitis B in Guangxi, China. PLoS One. 2020;15:e0234660. doi: 10.1371/journal.pone.0234660. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 1.Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12:878. doi: 10.15252/msb.20156651. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 2.Zhou LQ, Wang JY, Yu SY, Wu GG, Wei Q, Deng YB, Wu XL, Cui XW, Dietrich CF. Artificial intelligence in medical imaging of the liver. World J Gastroenterol. 2019;25:672–682. doi: 10.3748/wjg.v25.i6.672. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 3.Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep Learning in Medical Image Analysis. Adv Exp Med Biol. 2020;1213:3–21. doi: 10.1007/978-3-030-33128-3_1. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 4.Gore JC. Artificial intelligence in medical imaging. Magn Reson Imaging. 2020;68:A1–A4. doi: 10.1016/j.mri.2019.12.006. [DOI] [PubMed] [Google Scholar]

- 5.Liang X, Yu J, Liao J, Chen Z. Convolutional Neural Network for Breast and Thyroid Nodules Diagnosis in Ultrasound Imaging. Biomed Res Int. 2020;2020:1763803. doi: 10.1155/2020/1763803. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 6.Wang F, Liu X, Yuan N, Qian B, Ruan L, Yin C, Jin C. Study on automatic detection and classification of breast nodule using deep convolutional neural network system. J Thorac Dis. 2020;12:4690–4701. doi: 10.21037/jtd-19-3013. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 7.Dong F, She R, Cui C, Shi S, Hu X, Zeng J, Wu H, Xu J, Zhang Y. One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound. Eur Radiol. 2021;31:4991–5000. doi: 10.1007/s00330-020-07561-7. [DOI] [PubMed] [Google Scholar]

- 8.Khan S, Ullah R, Khan A, Ashraf R, Ali H, Bilal M, Saleem M. Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. Photodiagnosis Photodyn Ther. 2018;23:89–93. doi: 10.1016/j.pdpdt.2018.05.010. [DOI] [PubMed] [Google Scholar]

- 9.Li W, Huang Y, Zhuang BW, Liu GJ, Hu HT, Li X, Liang JY, Wang Z, Huang XW, Zhang CQ, Ruan SM, Xie XY, Kuang M, Lu MD, Chen LD, Wang W. Multiparametric ultrasomics of significant liver fibrosis: A machine learning-based analysis. Eur Radiol. 2019;29:1496–1506. doi: 10.1007/s00330-018-5680-z. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 10.Zhen S, Cai X. Letter to the Editor: Predicting Survival After Hepatocellular Carcinoma Resection Using Deep-Learning on Histological Slides. Hepatology. 2021;73:2077–2078. doi: 10.1002/hep.31543. [DOI] [PubMed] [Google Scholar]

- 11.Lavanchy D. Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. J Viral Hepat. 2004;11:97–107. doi: 10.1046/j.1365-2893.2003.00487.x. [DOI] [PubMed] [Google Scholar]

- 12.Guan P, Huang DS, Zhou BS. Forecasting model for the incidence of hepatitis A based on artificial neural network. World J Gastroenterol. 2004;10:3579–3582. doi: 10.3748/wjg.v10.i24.3579. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 13.Zheng Y, Zhang L, Zhu X, Guo G. A comparative study of two methods to predict the incidence of hepatitis B in Guangxi, China. PLoS One. 2020;15:e0234660. doi: 10.1371/journal.pone.0234660. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 14.Gan R, Chen X, Yan Y, Huang D. Application of a hybrid method combining grey model and back propagation artificial neural networks to forecast hepatitis B in china. Comput Math Methods Med. 2015;2015:328273. doi: 10.1155/2015/328273. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 15.Guo Y, Feng Y, Qu F, Zhang L, Yan B, Lv J. Prediction of hepatitis E using machine learning models. PLoS One. 2020;15:e0237750. doi: 10.1371/journal.pone.0237750. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 16.Ahmad G, Khan MA, Abbas S, Athar A, Khan BS, Aslam MS. Automated Diagnosis of Hepatitis B Using Multilayer Mamdani Fuzzy Inference System. J Healthc Eng. 2019;2019:6361318. doi: 10.1155/2019/6361318. [DOI] [PMC free article] [PubMed] [Google Scholar]

- 17.Wang X, Tian S, Yu L, Lv X, Zhang Z. Rapid screening of hepatitis B using Raman spectroscopy and long short-term memory neural network. Lasers Med Sci. 2020;35:1791–1799. doi: 10.1007/s10103-020-03003-4. [DOI] [PubMed] [Google Scholar]