

Machine Learning Framework for Accurate Prediction of Hepatitis C Virus Infection

G.Tulasi¹, B.Sai Deepthi²,

¹Final year MCA Student, Department of Computer Applications, GIET Engineering College, Andhra Pradesh, India

² Assistance Professor, GIET Engineering College, Andhra Pradesh, India

Abstract – The Hepatitis C Virus (HCV) is a serious problem in public health that leads to various life-threatening conditions, including liver fibrosis, cirrhosis, and hepatocellular carcinoma. It is crucial to detect HCV in order to timely avoid further complications. The conventional ways of diagnosing HCV include laboratory tests and professional analysis conducted by medical experts. However, these procedures require both time and money. Thus, considering the growing capabilities of machine learning algorithms, there is a possibility to develop a tool capable of supporting healthcare specialists in their decision-making process. The present study focuses on developing a machine learning algorithm for predicting HCV based on patients' data. First, it is necessary to pre-process the obtained data and balance it using the Synthetic Minority Oversampling Technique. In addition, it is important to perform the feature selection for building a Random Forest classifier, which will allow detecting HCV infection. Using a public database of HCV patients, one can train the proposed model that will have the accuracy of nearly 96.6%.

Key Words: Hepatitis C Virus (HCV), Machine Learning, Disease Prediction, Healthcare Analytics, Artificial Intelligence (AI), Random Forest Classifier, Synthetic Minority Oversampling Technique (SMOTE), Data Preprocessing, Feature Extraction, Medical Data Analysis, Predictive Modeling, Classification Algorithms, Healthcare Decision Support System, Liver Disease Detection, Clinical Data Analysis, Data Mining, Supervised Learning, Model Training, Accuracy Evaluation, Medical Diagnosis, Patient Health Monitoring, Data Visualization, Binary Classification, Health Informatics, Biomedical Data Processing, Early Disease Detection, Pattern Recognition, Python Programming, Scikit-Learn, Machine Learning Pipeline, Healthcare Technology.

1. INTRODUCTION

The Hepatitis C virus is a viral infection that targets the liver. The virus, which causes the disease, is called Hepatitis C Virus (HCV). One reason why HCV is referred to as the silent infection is that there are no clear signs of infection among many patients during the early stages of the illness. However, when left untreated, HCV can cause several complications such as chronic liver disease, liver cirrhosis, liver failure, and even liver cancer. Millions of individuals

have been affected by HCV annually, according to global health organizations.

The conventional diagnosis for Hepatitis C involves conducting a series of tests on the patient's blood in a laboratory set-up. Although the procedure yields accurate results, it may take some amount of time, effort, and skills from health experts. Over the past decade, several machine learning approaches have been introduced in the healthcare industry to diagnose diseases using the information gathered from patients' records.

Machine learning models can detect trends in patient records and predict their status accurately through historical analysis. Using variables such as age, sex, albumin, bilirubin, cholesterol, and biochemical measurements among others, the algorithm can diagnose patients for the infection.

2. Methodology

The Hepatitis C prediction process consists of a systematic approach to machine learning comprising several processes.

2.1 Data Acquisition

The dataset utilized in the current research consists of biochemical and medical data taken from the patients. The set comprises such attributes as Age, Gender, Albumin (ALB), Alkaline phosphatase (ALP), Alanine transaminase (ALT), Aspartate transaminase (AST), Bilirubin (BIL), Cholinesterase (CHE), Cholesterol (CHOL), Creatinine (CREA), Gamma-glutamyl transferase (GGT), and Protein (PROT). These attributes provide key indicators of liver functionality.

2.2 Data Preprocessing

Data preprocessing is an essential phase aimed at preparing data for the next processing stage. Such preprocessing activities as removal of unnecessary columns, handling of missing or incomplete data, data transformation, and other actions are implemented during the data preprocessing stage.

2.3 Addressing Class Imbalance Problem

Class imbalance is one of the problems associated with medical datasets. The HCV dataset consists of many healthy patients as compared to the patients who are infected. Class imbalance can adversely affect the prediction result. The Synthetic Minority Oversampling Technique (SMOTE) is used to tackle this problem by generating synthetic instances for the underrepresented class so that the datasets become balanced.

2.4 Feature Extraction

Feature extraction refers to choosing features or input attributes that play a role in predicting the status of patients. The extracted features are separated from the target variable that denotes healthy or infection of the patients.

2.5 Model Training

The dataset is partitioned into training and testing sets. A Random Forest classifier is used as a machine learning technique to build a disease prediction model. Random Forest is preferred as it ensures accurate predictions as well as is robust and resistant to overfitting.

The algorithm learns how to predict the status of patients according to their attributes.

2.6 Prediction and Model Evaluation

After training, the model is tested to predict its accuracy rate. For prediction purposes, testing is conducted to know how well the algorithm predicts outcomes according to the attributes of patients

Results:

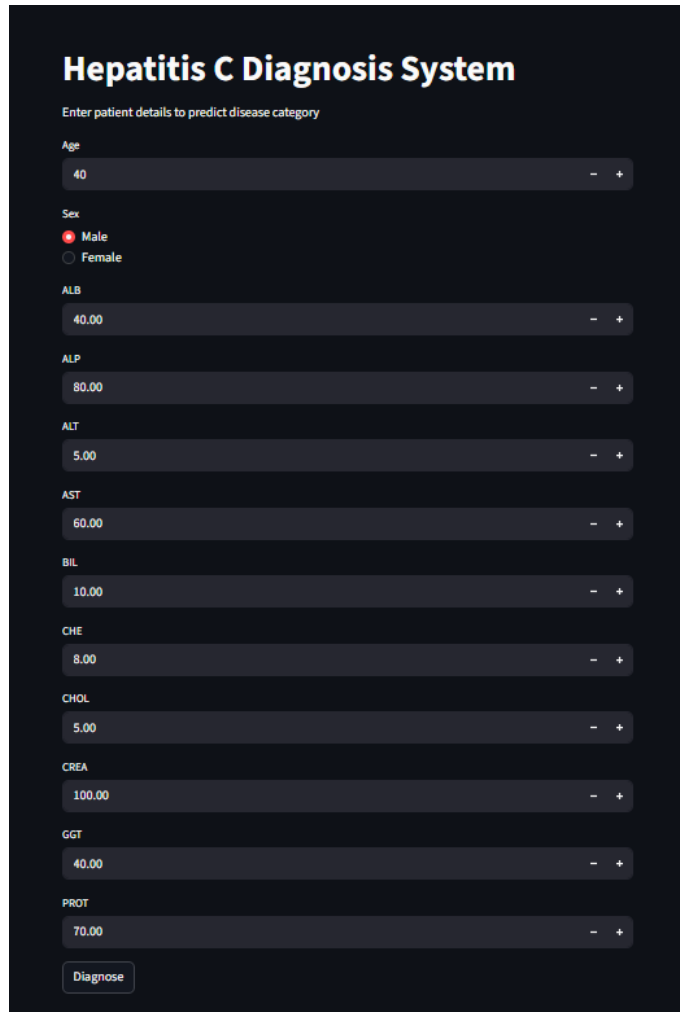


Fig-1: System Interface

Hepatitis C Diagnosis System

Enter patient details to predict disease category

Age: 40

Sex: ☒ Male ☐ Female

ALB: 40.00

ALP: 80.00

ALT: 5.00

AST: 60.00

BIL: 10.00

CHIE: 8.00

CHOL: 5.00

CREA: 100.00

GGT: 40.00

PROT: 70.00

Diagnose

Predicted Category: Blood Donor

Model Accuracy: 93.55%

Prediction Confidence

Blood Donor: 71.62%

Hepatitis: 12.50%

Fibrosis: 14.14%

Cirrhosis: 1.74%

Fig-2: Prediction Normal Case

Hepatitis C Diagnosis System

Enter patient details to predict disease category

Age: 60

Sex: ☐ Male ☒ Female

ALB: 40.00

ALP: 80.00

ALT: 5.00

AST: 60.00

BIL: 10.00

CHIE: 8.00

CHOL: 5.00

CREA: 100.00

GGT: 40.00

PROT: 70.00

Diagnose

Predicted Category: Blood Donor

Model Accuracy: 93.55%

Prediction Confidence

Blood Donor: 61.77%

Hepatitis: 4.07%

Fibrosis: 31.14%

Cirrhosis: 3.02%

⚠ Risk of Fibrosis

Fig-3: Prediction Risk Case

3. DISCUSSION

The system was successfully developed by employing Python along with machine learning libraries. The initial dataset had a severe problem with class imbalance, as it had 526 patients' records representing the "no-disease" category and 63 patients' records from the "with-disease" category. After SMOTE implementation, the balanced dataset had 427 patients' records in each class.

In terms of prediction power, the Random Forest classifier was able to provide good results. The overall accuracy was about 96.6%. This shows that the model is capable of accurately classifying the vast majority of patients' records. Experiments were conducted on normal and abnormal patient datasets.

When presented with records of normal patients containing no abnormalities in liver biochemical function values, the system would make the prediction "HCV Negative." On the other hand, when provided with records of abnormal patients containing liver function abnormalities, the system would predict "HCV Positive." Overall, the obtained high accuracy of prediction proves that the Random Forest model can be employed effectively in predicting various diseases in patients. In addition, the application of SMOTE helped solve the problem with class imbalance. Thus, the developed model can become a useful tool for medical practice.

4. CONCLUSIONS

The current paper managed to develop an efficient machine learning algorithm for HCV disease predictions. The designed framework considers the implementation of data pre-processing, feature extraction, class balancing with SMOTE approach, and the use of the Random Forest algorithm for the development of the final classification model. It was possible to reach an accuracy level of approximately 96.6% in the process. Several advantages are associated with the usage of the proposed framework. Among such benefits, there should be mentioned faster diagnostics, reduced labor costs, increased prediction accuracy, as well as assistance in clinical decisions. As a result, the early identification of HCV patients is going to be possible. As for future developments, it is necessary to mention that in the future it will be possible to integrate big data approaches, the use of deep learning algorithms, cloud deployment opportunities, as well as real-time healthcare management systems.

5. REFERENCES

- [1] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: Management, analysis and future prospects," *J. Big Data*, vol. 6, no. 1, p. 54, Dec. 2019.
- [2] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2012.
- [3] B. Hajarizadeh, J. Grebely, and G. J. Dore, "Epidemiology and natural history of HCV infection," *Nature Rev. Gastroenterol. Hepatol.*, vol. 10, no. 9, pp. 553–562, Sep. 2013.
- [4] J. S. Bajaj, "Alcohol, liver disease and the gut microbiota," *Nature Rev. Gastroenterol. Hepatol.*, vol. 16, no. 4, pp. 235–246, Apr. 2019.
- [5] J.-H. Wang, C.-H. Chen, C.-M. Chang, W.-C. Feng, C.-Y. Lee, and S.-N. Lu, "Hepatitis C virus core antigen is cost-effective in community based screening of active hepatitis C infection in Taiwan," *J. Formosan Med. Assoc.*, vol. 119, no. 1, pp. 504–508, Jan. 2020.
- [6] T. Boettler, P. N. Newsome, M. U. Mondelli, M. Maticic, E. Cordero, M. Cornberg, and T. Berg, "Care of patients with liver disease during the COVID19pandemic:EASLSCMIDpositionpaper," *JHEPRep.*, vol. 2, no. 3, Jun. 2020, Art. no. 100113.
- [7] J. H. Hoofnagle, "Hepatitis C: The clinical spectrum of disease," *Hepatology*, vol. 26, no. S3, pp. 15S–20S, Dec. 1997.
- [8] A. M. Anter, S. Bhattacharyya, and Z. Zhang, "Multi-stage fuzzy swarm intelligence for automatic hepatic lesion segmentation from CT scans," *Appl. Soft Comput.*, vol. 96, Nov. 2020, Art. no. 106677.
- [9] R. T. Ribeiro, R. T. Marinho, and J. M. Sanches, "Classification and staging of chronic liver disease from multimodal data," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 5, pp. 1336–1344, May 2013.
- [10] G. Chandrasekaran, P. R. Karthikeyan, N. S. Kumar, and V. Kumarasamy, "Test scheduling of system-on-chip using dragonfly and ant lion optimization algorithms," *J. Intell. Fuzzy Syst.*, vol. 40, no. 3, pp. 4905–4917, Mar. 2021, doi: 10.3233/JIFS-201691.
- [11] G. Chandrasekaran, S. Periyasamy, and P. R. Karthikeyan, "Test scheduling for system on chip using modified firefly and modified ABC algorithms," *Social Netw. Appl. Sci.*, vol. 1, no. 9, p. 1079, Sep. 2019, doi: 10.1007/S42452-019-1116-X.
- [12] G. Chandrasekaran, V. Kumarasamy, and G. Chinraj, "Test scheduling of core based system-on-chip using modified ant colony optimization," *J. Européen Systèmes Automatisés*, vol. 52, no. 6, pp. 599–605, Dec. 2019, doi: 10.18280/JESA.520607.
- [13] G. Chandrasekaran, G. Singaram, R. Duraisamy, A. S. Ghodake, and P. K. Ganesan, "Test scheduling and test time reduction for SoC by using enhanced firefly algorithm," *Revue d'Intell. Artificielle*, vol. 35, no. 3, pp. 265–271, Jun. 2021, doi: 10.18280/RIA.350310.
- [14] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, "Computer-aided decision-making for

predicting liver disease using PSO based optimized SVM with feature selection,” *Informat. Med. Unlocked*, vol. 17, Jan. 2019, Art. no. 100255.

[15] C.-C. Wu, W.-C. Yeh, W.-D. Hsu, M. M. Islam, P. A. A.

Nguyen, T. N. Poly, Y.-C. Wang, H.-C. Yang, and Y.-C. J. Li, “Prediction of fatty liver disease using machine learning algorithms,” *Comput. Methods Programs Biomed.*, vol. 170, pp. 23–29, Mar. 2019.

[16] J. Y. Nakayama, J. Ho, E. Cartwright, R. Simpson, and V.

S. Hertzberg, “Predictors of progression through the cascade of care to a cure for hepatitis C patients using decision trees and random forests,” *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104461.

[17] O. Kadioglu, M. Saeed, H. J. Greten, and T. Efferth, “Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning,” *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104359.

[18] R. Kumar, V. Kumar, and K. W. Lee, “A computational drug repurposing approach in identifying the cephalosporin antibiotic and anti-hepatitis C drug derivatives for COVID-19 treatment,” *Comput. Biol. Med.*, vol. 130, Mar. 2021, Art. no. 104186.

[19] D. Chicco and G. Jurman, “An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis,” *IEEE Access*, vol. 9, pp. 24485–24498, 2021.

[20] A. Fabijanska and S. Grabowski, “Viral genome deep classifier,” *IEEE Access*, vol. 7, pp. 81297–81307, 2019.

[21] H.-J. Chiu, T.-H.-S. Li, and P.-H. Kuo, “Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine,” *IEEE Access*, vol. 8, pp. 204309–204324, 2020.

[22] A. Basher, B. C. Kim, K. H. Lee, and H. Y. Jung, “Volumetric feature based Alzheimer’s disease diagnosis from sMRI data using a convolutional neural network and a deep neural network,” *IEEE Access*, vol. 9, pp. 29870–29882, 2021, doi: 10.1109/ACCESS.2021.3059658.

[23] P. Khan, M. F. Kader, S. M. R. Islam, A. B. Rahman, M. S. Kamal, M. U. Toha, and K.-S. Kwak, “Machine learning and deep learning approaches for brain disease diagnosis: Principles and recent advances,” *IEEE Access*, vol. 9, pp. 37622–37655, 2021, doi: 10.1109/ACCESS.2021.3062484.

[24] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2001, pp. 1–8.

[25] G. Heitz, S. Gould, A. Saxena, and D. Koller, “Cascaded classification models: Combining models for holistic scene understanding,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2009, pp. 641–648.

[26] Z. Zhang and P. H. S. Torr, “Object proposal generation using two-stage cascade SVMs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 102–115, Jan. 2016.

[27] Z. Cai, M. Saberian, and N. Vasconcelos, “Learning complexity-aware cascades for pedestrian detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2195–2211, Sep. 2020.

[28] X. Liang, D. Li, M. Song, A. Madden, Y. Ding, and Y. Bu, “Predicting biomedical relationships using the knowledge and graph embedding cascade model,” *PLoS ONE*, vol. 14, no. 6, Jun. 2019, Art. no. e0218264.

[29] N. Michielli, U. R. Acharya, and F. Molinari, “Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals,” *Comput. Biol. Med.*, vol. 106, pp. 71–81, Mar. 2019.

[30] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.