Chapter 13
Hepatitis C Prediction Using Feature Selection by Machine Learning Technique

Jeet Majumder
Brainware University, India

Suman Ghosh
Brainware University, India

Alex Khang
https://orcid.org/0000-0001-8379-4659
Global Research Institute of Technology and Engineering, USA

Tridibesh Debnath
Brainware University, Barasat

Avijit Kumar Chaudhuri
https://orcid.org/0000-0002-5310-3180
Brainware University, India

ABSTRACT

This study suggests a prediction framework for the Hepatitis C virus that is based on machine learning techniques. The authors made use of a dataset available on Kaggle. In this dataset, 564 patients with 12 distinct features are present. They tested two cases, the first one without feature selection and with feature selection based on gain ratio attribute evaluation (GRAE), to guarantee the strength and dependability of the suggested framework. Additionally, an evaluation is conducted on the feature subset that was chosen using the GRAE-generated features. For model evaluation, induction methods and classifiers such as logistic regression (LR), naive bayes (NB), decision tree (DT), support vector machine (SVM), random forest (RF), and multilayer perceptron (MLP) are used. According to the experimental findings, the suggested framework outperformed the others in terms of all accuracy matrices following GRAE selection. According to the experimental findings, the suggested framework outperformed the unfeatured one in terms of accuracy after GRAE selection.

DOI: 10.4018/979-8-3693-2105-8.ch013

1. INTRODUCTION

One of the main viruses that cause liver disease is the hepatitis C virus (HCV), which belongs to the Flaviviridae family. Approximately 175 million individuals globally, or 3% of the global population, are infected with HCV. Although 90% of injectable drug users are most at risk, parental transmission is the primary mode of HCV transmission. Conventional interferon and ribavirin, which have 38–43% sustained virological response rates, are still the gold standard for treating chronic HCV (Munir et al., 2010). Approximately 58 million people worldwide suffer from long-term hepatitis C virus disease, and 1.5 million new cases are reported each year. Nearly 3.2 million children and adolescents suffer from a chronic case of hepatitis C (WHO, 2023). It is asymptomatic at first, but when the infection worsens, it can cause chronic illnesses such as hepatocellular cancer and liver cirrhosis. To diagnose this illness, a few different non-invasive serum biochemical indicators are employed (Nandipati et al., 2020).

To determine the disease's stage, a variety of harmless blood biochemical indicators and patient medical information have been employed. Machine learning techniques have shown to be a helpful alternative for determining the phase of this chronic liver disease, avoiding the drawbacks of a biopsy (Butt et al., 2021). To stop the spread of disease and identify affected areas early on, medical research relies heavily on the forecasting and categorization of diseases. Machine learning (ML) techniques are frequently employed to accurately forecast and categorise diseases, serving as a useful tool for medical professionals (Mamdouh Farghaly et al., 2023). Clinical data contains complex and non-linear correlations that machine learning (ML) techniques are especially good at acquiring and analysing.

Through the identification of HCV-positive individuals, machine learning algorithms, including classification approaches, can be employed to create a model for HCV diagnosis. However, unsuitable attribute set features can degrade the classifier's effectiveness (John et al., 1994). When there are more important and non-redundant attributes in the data, various learning algorithms perform better and produce more accurate results. An effective feature selection strategy is required to extract intriguing aspects relevant to the condition, as clinical datasets contain a huge number of duplicated and irrelevant information (Jain & Singh, 2018).

Table 1. Data set for Hepatitis C virus performance comparison

Author	Method	Performance Matrix	Percentage
Orooji and Kermani (2021)	MLP, Bayesian network, and DT	Specificity	100%
		f-measure	99.90%
		Accuracy	99.90%
Alotaibi et al. (2023)	RF. Gradient Boosting Machine, DT	recall	96.00%
		precision	99.81%
		AUC/ROC	96.00%
		Accuracy	96.92%
Abd El-Salam et al. (2019)	Bayesian Network	Accuracy	68.90%
		ROC	74.80%
KayvanJoo et al. (2014)	NB	Accuracy	89.17%
Eliyahu et al. (2018)	RF	Accuracy	91%
Hashem et al. (2020)	DT, LR	Accuracy	95.60%

Hepatocellular carcinoma (HCC) infection is still an important contributor to liver cirrhosis, liver transplants, and a global health problem today. But because of decades of incredible progress, HCV is now the first chronic viral infection that can be cured (Manns & Maasoumy, 2022) as shown in Table 1.

2. REVIEW OF LITERATURE

Automatic generation of prediction and diagnostic strategies is facilitated by machine learning techniques. Although machine learning assessments have produced fascinating findings, the results do not match the methods and datasets used. To predict the disease, we have experimented with and tested a variety of classifiers, including LR, NB, SVM, DT, RF, and MLP.

For this study's analysis, Rabab Salama & Mohamed M. Ezz used a dataset of 4965 severe C patients. There are twenty-four distinct clinical laboratory variables in the sample. They discovered that the ROC curve's accuracy was between 68.9% and 74.8% (Abd El-Salam et al., 2019).

Chi-Squared was utilised by Amir Hossein KayvanJoo & Mansour Ebrahimi. Using Naïve Bayes, they attained the highest accuracy of 89.17% (KayvanJoo et al., 2014).

Sivan Eliyahu, Oz Sharabi & Shiri Elmedvi are achieved 91% of accuracy. They have received an accuracy of the model by Random Forest to the final two samples. 100 iterations of the sampling and training procedures were conducted to make sure the model was not biased towards any particular samples (Eliyahu et al., 2018).

Somaya Hashem & Mahmoud ElHefnawi achieved an accuracy of 95.6% and 99% of AUCROC. They used a dataset comprising 4423 CHC patients was examined to determine the important variables for predicting the existence of Chronic Hepatitis C (CHC). This work built Hepatocellular Carcinoma (HCC) classification models for the prediction of HCC existence using a variety of machine-learning approaches, including classification and regression trees, alternating decision trees, reduction pruning error trees, and linear regression algorithms (Hashem et al., 2020).

Azam Orooji & Farzaneh Kermani shows the over-sampling approach enhanced the performance metrics of data mining algorithms in predicting diseases, as demonstrated by the results, which revealed that the best technique (random forest) had an accuracy of 99.9% in the O-dataset. Additionally, the random forest for the O-dataset obtained the highest performance metrics in terms of 100% specificity, f-measure (99.9%), accuracy, and sensitivity (Orooji & Kermani, 2021).

Abrar Alotaibi & Lujain Alnajrani show four machine learning algorithms—a Random Forest, a Gradient Boosting Machine, an Extreme Gradient Boosting, and an Extra Trees model—were trained on the dataset to identify cirrhosis in hepatitis C patients. By utilising only 16 of the 28 characteristics, the additional trees model surpassed the other models, attaining accuracy of 96.92%, a recall of 94.00%, a precision of 99.81%, and an area under the receiver operating characteristic curve of 96% (Alotaibi et al., 2023).

3. METHODOLOGY

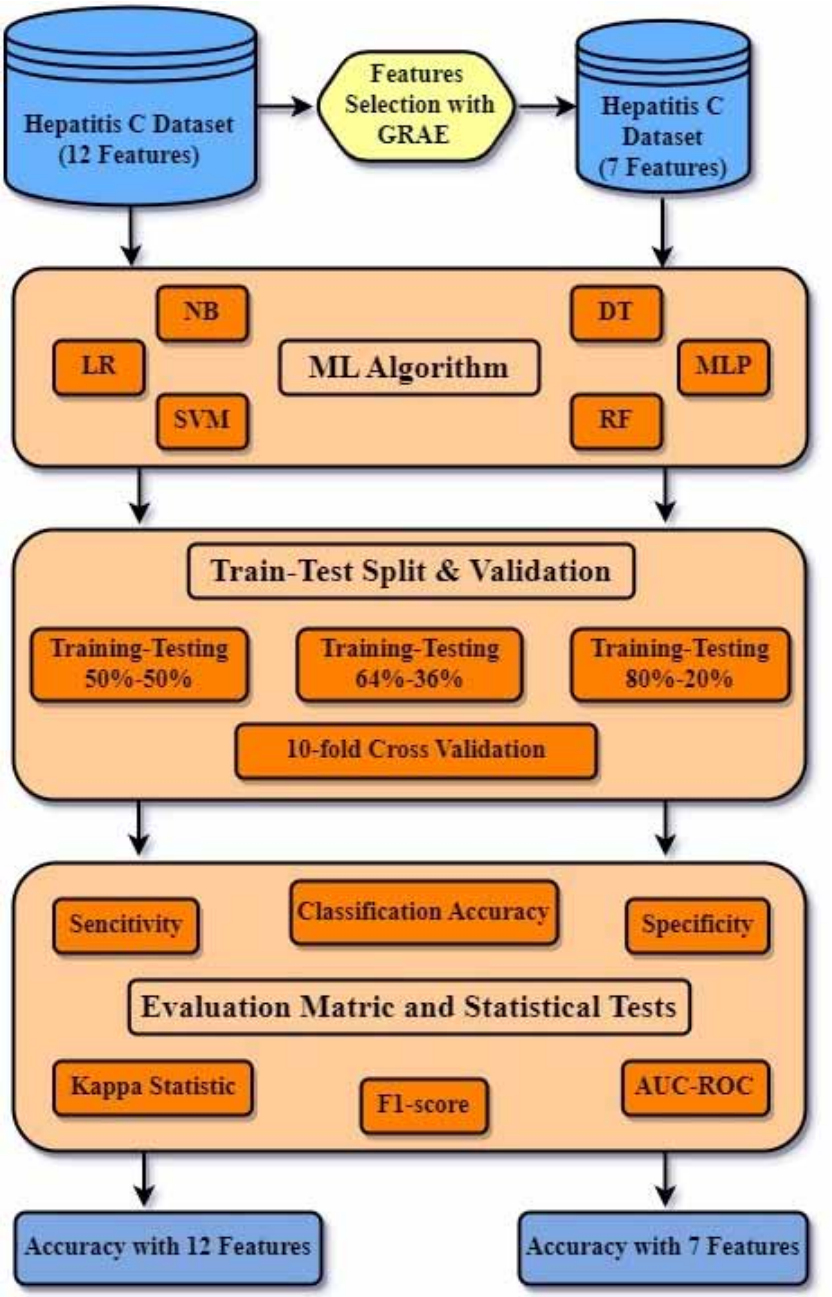
The goal of the proposed model is to distinguish between patients who have Hepatitis C Virus (HCV) infection and those who do not. Our goal has been to increase the precision of machine learning techniques used to classify healthcare workers across the world. To compare the attained accuracy, the classifiers'

performance has been evaluated on both all characteristics and specific features independently (Khang & Medicine, 2023).

3.1. Workflow

Proposed workflow as Figure 1.

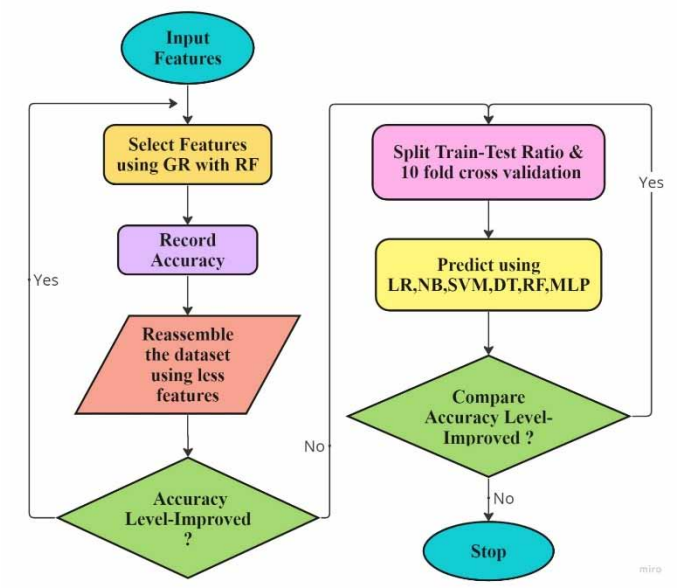
Figure 1. Depicts the proposed workflow to distinguish between patients who have Hepatitis C virus



3.2. Flowchart

Proposed flowchart as Figure 2.

Figure 2. Depicts the proposed flowchart to distinguish between patients who have Hepatitis C virus



4. DESCRIPTION OF DATASET

The authors used the Kaggle (Diagnostic) dataset for the Hepatitis C virus, created by Ralf Lichtinghagen, Frank Klawonn, and Georg Hoffmann (2023). There are 11 missing values present in that dataset. Here we assume 'suspect blood donor' as 'blood donor' and we remove the records who have Fibrosis and Cirrhosis as shown in Table 2.

5. IMPACT OF FEATURE SELECTION ON MACHINE LEARNING MODEL PERFORMANCE

Using the Hepatitis C dataset, we investigated how feature selection affected the classifier's performance using popular classification algorithms such as LR, NB, SVM, DT, RF, and MLP. As indicated in Table 3, classification methods are applied in two scenarios to predict the presence of hepatitis C. (1) with the total feature selection approach and (2) with the GRAE feature selection approach applied.

Table 2. Description of dataset

Sl. No	Attributes	Description	Mean	Standard Deviation
1	Age	Patients age (in year)		
2	Sex	Gender (m,f)		
3	ALB	Albumin level	42.08	5.44
4	ALP	Alkaline Phosphatase	67.86	20.24
5	ALT	Alanine Aminotransferases	27.58	21.03
6	AST	Aspartate Aminotransferase	29.19	21.23
7	BIL	Bilirubin level	8.78	6.69
8	CHE	HCV Core Antigen	8.42	1.96
9	CHOL	Cholesterol level	5.46	1.09
10	CREA	Creatinine	78.54	15.7
11	GGT	Gamma-glutamyl transferase levels.	33.46	40.81
12	PROT	Protein factor	71.99	5.18
13	Category	The Diagnosis (values: '0=Blood Donor', '1=Hepatitis')		

Table 3. Comparison of accuracies

Training-Testing Partition	Features	Accuracy					
		LR	NB	SVM	DT	RF	MLP
50-50	12	0.82	0.81	0.83	0.83	0.84	0.76
	7	0.83	0.82	0.95	1	1	1
66-34	12	0.8	0.79	0.78	0.87	0.86	0.79
	7	0.82	0.81	0.76	1	1	0.96
80-20	12	0.76	0.75	0.77	0.88	0.84	0.75
	7	0.80	0.75	0.96	1	1	1
10 fold cross validation	12	0.81	0.80	0.81	0.85	0.88	0.82
	7	0.90	0.89	0.95	1	1	1

6. RESULTS AND DISCUSSIONS

We used WEKA, a machine learning analysis tool, to design, simulate, and assess a suggested model. The authors of this model conduct a comparative analysis of six sophisticated machine learning algorithms: LR, NB, SVM, DT, RF, and MLP of these six widely used machine learning techniques, some perform better in terms of accuracy, while others perform worse as shown Table 4.

The statistical results of classifications obtained from the examination of several classification schemes are shown in a confusion matrix. The data produced in this matrix is typically used to evaluate the effectiveness of all such systems. The results of the various machine learning methods produced from confusion matrices are displayed in Table 5.

Table 4. Comparison of sensitivity and specificity

Training-Testing Partition	Features	LR		NB		SVM		DT		RF		MLP	
		Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
50-50	12	0.84	0.67	0.83	0.68	0.87	0.66	0.88	0.66	0.86	0.71	0.82	0.38
	7	0.89	0.71	0.94	0.65	1	0.85	1	1	1	1	1	1
66-34	12	0.81	0.73	0.78	0.75	0.8	0.65	0.89	0.8	0.86	0.84	0.83	0.6
	7	0.88	0.71	0.87	0.69	1	0.9	1	1	1	1	1	0.9
80-20	12	0.76	0.73	0.75	0.71	0.77	0.72	0.88	0.85	0.84	0.83	0.79	0.61
	7	0.87	0.68	0.84	0.62	1	0.9	1	1	1	1	1	1
10 fold cross validation	12	0.83	0.67	0.81	0.65	0.83	0.68	0.87	0.72	0.89	0.8	0.87	0.63
	7	0.92	0.84	0.98	0.76	1	0.87	1	1	1	1	1	1

Table 5. Comparison of F1 score

Training-Testing Partition	Features	F1 SCORE					
		LR	NB	SVM	DT	RF	MLP
50-50	12	0.8	0.79	0.82	0.83	0.83	0.74
	7	0.83	0.82	0.94	1	1	1
66-34	12	0.77	0.76	0.76	0.87	0.84	0.77
	7	0.82	0.8	0.96	1	1	0.96
80-20	12	0.72	0.71	0.74	0.87	0.83	0.74
	7	0.8	0.75	0.96	1	1	1
10 fold cross validation	12	0.79	0.76	0.79	0.84	0.87	0.81
	7	0.89	0.89	0.95	1	1	1

Table 6. Comparison of kappa statistic

Training-Testing Partition	Features	Kappa Statistic					
		LR	NB	SVM	DT	RF	MLP
50-50	12	0.39	0.36	0.48	0.51	0.49	0.25
	7	0.62	0.62	0.88	1	1	1
66-34	12	0.36	0.33	0.33	0.65	0.58	0.4
	7	0.6	0.58	0.92	1	1	0.92
80-20	12	0.32	0.29	0.36	0.68	0.58	0.36
	7	0.57	0.48	0.92	1	1	1
10 fold cross validation	12	0.4	0.3	0.38	0.55	0.62	0.48
	7	0.76	0.77	0.89	1	1	1

Table 7. Comparison of AUC / ROC

Training-Testing Partition	Features	AUC / ROC					
		LR	NB	SVM	DT	RF	MLP
50-50	12	0.816	0.81	0.72	0.79	0.92	0.73
	7	0.93	0.9	0.96	1	1	1
66-34	12	0.82	0.82	0.64	0.86	0.93	0.82
	7	0.94	0.91	0.97	1	1	0.92
80-20	12	0.81	0.81	0.66	0.83	0.94	0.79
	7	0.93	0.9	0.97	1	1	1
10 fold cross validation	12	0.81	0.8	0.66	0.83	0.95	0.82
	7	0.94	0.94	0.96	1	1	1

By weighing precision and recall, the F1 score is a statistic that evaluates how well the binary classification system performs. It has a range of 0 to 1, where 1 denotes error-free recall and precision.

The Kappa statistic is a categorization indicator that accounts for random fluctuation. On a scale of -1 to 1, a value of 1 denotes perfect agreement, a value of 0 denotes agreement equal to chance, and a value of negative implies acceptance less than chance as shown in Table 6. When determining the dependability of classification methods is critical, kappa is useful. Because it takes into consideration the likelihood of agreement occurring at random, it offers a more thorough examination than raw agreement (Khang and Abuzarova et al., 2023).

A measure used to assess the effectiveness of binary classification models is the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). The relationship between the true positive rate and with false positive rate for various categorization levels is depicted by the ROC curve. AUC measures the area of the curve; a larger AUC (closer to 1) denotes better discrimination from the model as shown in Table 7.

7. CONCLUSION

This ML framework consisting of LR, NB, SVM, DT, RF, and MLP is proposed in this research to predict and categorise HCV-infected patients based on enrolled features. The most important characteristics of the application dataset are chosen using the GRAE feature selection method. We also tested the dataset by applying all of the characteristics directly to the LR, NB, SVM, DT, RF, and MLP without feature selection to modify the registered data without feature selection. Lastly, we investigated how parameter adjustment affected learning strategies. The findings show that the accuracy attained following choosing characteristics using GRAE using the new parameters has much improved (Khang & Kali et al., 2023).

8. FUTURE WORK

It has been demonstrated that the machine-learning techniques employed in this suggested classifier to forecast early identification of the Hepatitis C virus are very successful. Little datasets with few attri-

butes were used to test the study's suggested model. However, further diverse real-world cancer datasets, including data from multiple different sources, are needed to evaluate the same approach. Medical datasets or non-medical data can be retrieved from particular electronic health record repositories (Khang & Rana et al., 2023).

REFERENCES

Abd El-Salam, S. M., Ezz, M. M., Hashem, S., Elakel, W., Salama, R., ElMakhzangy, H., & ElHefnawi, M. (2019). Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients. *Informatics in Medicine Unlocked*, *17*, 100267. doi:10.1016/j.imu.2019.100267

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 (Basel, Switzerland)*, *11*(6), 104. doi:10.3390/computation11060104

Butt, M. B., Alfayad, M., Saqib, S., Khan, M. A., Ahmad, M., Khan, M. A., & Elmitwally, N. S. (2021). Diagnosing the stage of hepatitis C using machine learning. *Journal of Healthcare Engineering*, *2021*, 2021. doi:10.1155/2021/8062410 PMID:35028114

Eliyahu, S., Sharabi, O., Elmedvi, S., Timor, R., Davidovich, A., Vigneault, F., Clouser, C., Hope, R., Nimer, A., Braun, M., Weiss, Y. Y., Polak, P., Yaari, G., & Gal-Tanamy, M. (2018). Antibody repertoire analysis of hepatitis C virus infections identifies immune signatures associated with spontaneous clearance. *Frontiers in Immunology*, *9*, 3004. doi:10.3389/fimmu.2018.03004 PMID:30622532

Hashem, S., ElHefnawi, M., Habashy, S., El-Adawy, M., Esmat, G., Elakel, W., Abdelazziz, A. O., Nabeel, M. M., Abdelmaksoud, A. H., Elbaz, T. M., & Shousha, H. I. (2020). Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease. *Computer Methods and Programs in Biomedicine*, *196*, 105551. doi:10.1016/j.cmpb.2020.105551 PMID:32580053

Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, *19*(3), 179–189. doi:10.1016/j.eij.2018.03.002

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994* (pp. 121–129). Morgan Kaufmann. doi:10.1016/B978-1-55860-335-6.50023-4

KayvanJoo, A. HEbrahimi, MHaqshenas, G. (2014). Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. *BMC Research Notes*, *7*(1), 1–11. PMID:25150834

Khang, A. (2023). *AI and IoT-Based Technologies for Precision Medicine* (1st ed.). IGI Global Press., doi:10.4018/979-8-3693-0876-9

Khang, A. (2023). Enabling the Future of Manufacturing: Integration of Robotics and IoT to Smart Factory Infrastructure in Industry 4.0. In AI-Based Technologies and Applications in the Era of the Metaverse. IGI Global Press. doi:10.4018/978-1-6684-8851-5.ch002

Khang, A., & Abdullayev, V. A. (2023). *AI-Aided Data Analytics Tools and Applications for the Healthcare Sector. In AI and IoT-Based Technologies for Precision Medicine* (1st ed.). IGI Global Press. doi:10.4018/979-8-3693-0876-9.ch018

Khang, A., Rana, G., Tailor, R. K., & Hajimahmud, V. A. (2023). *Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem* (1st ed.). CRC Press. doi:10.1201/9781003356189

LichtinghagenR.KlawonnF.HoffmannG. (2023). https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset

Mamdouh Farghaly, H., Shams, M. Y., & Abd El-Hafeez, T. (2023). Hepatitis C Virus prediction based on machine learning framework: A real-world case study in Egypt. *Knowledge and Information Systems*, *65*(6), 2595–2617. doi:10.1007/s10115-023-01851-4

Manns, M. P., & Maasoumy, B. (2022). Breakthroughs in hepatitis C research: From discovery to cure. *Nature Reviews. Gastroenterology & Hepatology*, *19*(8), 533–550. doi:10.1038/s41575-022-00608-8 PMID:35595834

Munir, S., Saleem, S., Idrees, M., Tariq, A., Butt, S., Rauff, B., Hussain, A., Badar, S., Naudhani, M., Fatima, Z., Ali, M., Ali, L., Akram, M., Aftab, M., Khubaib, B., & Awan, Z. (2010). Hepatitis C treatment: Current and future perspectives. *Virology Journal*, *7*(1), 1–6. doi:10.1186/1743-422X-7-296 PMID:21040548

Nandipati, S. C., & XinYing, CWah, K. K. (2020). Hepatitis C virus (HCV) prediction by machine learning techniques. *Applications of Modelling and Simulation*, *4*, 89–100.

Orooji, A., & Kermani, F. (2021). Machine learning based methods for handling imbalanced data in hepatitis diagnosis. *Frontiers in Health Informatics*, *10*(1), 57. doi:10.30699/fhi.v10i1.259

WHO. (2023). *Hepatitis C*. https://www.who.int/news-room/fact-sheets/detail/hepatitis-c