ISBN 978-623-99026-6-7

A Comparative Study on Hepatitis C Predictions Using Machine Learning Algorithms

Fergie Joanda Kaunang

Universitas Advent Indonesia, Indonesia

fergie.kaunang@unai.edu

Abstract

Hepatitis C virus (HCV) is known to be the major cause of chronic liver disease. Based on research, HCV has caused more than 100.000 cases of liver cancer per year. This virus has become the cause of at least 280.000 deaths. To diagnose HCV, it takes at least two different tests, namely serological assays and molecular tests, which are quite costly and complex. With Machine Learning technology, the diagnosis of any disease or virus can be made by detecting different patterns or relationships. Therefore, this study aims to predict the Hepatitis C virus using different machine learning algorithms and find out the best model for the classification of Hepatitis C disease. Furthermore, this study shows some visualizations to find out the relationships between attributes. We used different machine learning algorithms, namely K-Nearest Neighbour, Support Vector Machine, Random Forest, Neural Network, Naïve Bayes, and Logistic Regression. The performance of those different machine learning algorithms was evaluated using four different metrics, which are classification accuracy, precision, recall, and F-1 score. The classification accuracy results are 96.5%, 96.7%, 97.3%, 97.1%, 96%, 97.9% each for k-NN, SVM, RandomFores, Neural Network, Naïve Bayes and Logistic Regression. Based on the results, each model showed high performance, but Logistic Regression performs the best result. With the results conducted by this study, it is hoped that it can help the diagnosis process of HCV based on laboratory data. However, it is important to communicate the shortcomings and some possible improvements for each model.

Keywords: Machine Learning, Predictions, Hepatitis C Virus

INTRODUCTION

Hepatitis C is known as an infectious disease that attacks the liver. This disease is caused by an RNA virus called Hepatitis C virus (HCV) that can lead to acute and chronic hepatitis (Lauer & Walker, 2001). HCV is also found to be the cause of liver cirrhosis, as well as hepatocellular carcinoma, and can last for a few weeks or a lifetime (Shepard et al., 2005). Stomach pain, dark urine, Jaundice, grey-colored feces, loss of appetite, fatigue are a few symptoms of Hepatitis C disease (Bailey Jr et al., 2009). The HCV itself is a bloodborne virus; in other words, it spreads through direct contacts, such as through unsafe health care, unsafe injection practices, including transfusion of unscreened blood products (Bréchot, 1996). Based on research, HCV has infected millions of people worldwide (Alter, 2007; Lauer & Walker, 2001).

In most cases, the laboratory investigation of HCV starts with detecting antibodies to HCV through serological assays. This process is followed by the detection of HCV RNA through molecular assays. For low-risk individuals, determining anti HCV antibodies by immunoblot assays or simple, rapid immunoassays has proven to significantly reduce the risk of HCV (Clemens et al., 1992; Somi et al., 2014). However, there are groups of molecular detection methods that need to be done to diagnose HCV for high-risk individuals (Firdaus et al., 2015). One of the methods that have proven to be a useful way to detect liver disease is liver biopsy. This method provides an important clue for prognosis and for the management of patients with hepatitis C (Saadeh et al., 2001).

However, this biopsy method has some potential risks that could lead patients to hospitalization after being tested using this method because of its invasive nature (Nandipati et al., 2020). Therefore, this study aims to explore and predict the Hepatitis C virus using different machine learning algorithms and find out the best model for the classification of Hepatitis C disease. Furthermore, this study shows some visualizations to find out the relationships between attributes.

Literature Review

Nowadays, machine learning has been significantly used in the field of healthcare and medicine. Areas like pathology, radiology, oncology, protein functions, post-translational modification, and cardiology have applied the use of machine learning (Kao et al., 2017; Sandag & Kaunang, 2019; Uttreshwar & Ghatol, 2009; Weng et al., 2017). Machine learning itself is a computational method to make predictions using past information or data (Mohri et al., 2018).

In the previous section, we have seen that using an invasive method like biopsy for detecting and staging the Hepatitis C disease is risky. Therefore, there are some non-invasive methods that have been used as an alternative for detecting and staging liver disease to overcome the deficiencies of the liver biopsy method. (Parkes et al., 2006) showed that Serum markers of liver fibrosis are a less invasive alternative to liver biopsy because of its simplicity and no risk of complications. Another alternative is Liver Stiffness Measurement (LSM) (Ziol et al., 2005). This method was performed by transient elastography and appeared to be a reliable tool to detect hepatitis C disease.

During the last few decades, data mining and machine learning algorithms have been applied by researchers and clinicians as non-invasive methods for the detection and staging of hepatitis C disease. This method can be another non-invasive alternative to automatically diagnose any disease, including hepatitis C disease. (Hashem et al., 2017) has carried out a study to predict advanced liver fibrosis in hepatitis C patients using machine learning approaches. They conducted an evaluation of the group of 39,567 chronic HCV patients in Egypt using four different classification algorithms, namely decision tree algorithm, genetic algorithm, particle swarm optimization, and multi-linear regression algorithm.

Another previous study was conducted by (Agarwal et al., 2019) to determine risk factors of HCV on HIV-infected patients in India. The dataset of 350 observations with 90 attributes was run using the Random Forest algorithm. They found that attributes like Jaundice, Depression, Injected Drug Users, and HIV are important risk factors predictors to HCV. The accuracy of their developed

model is found to be 98.3% which concludes that Random Forest is applicable to predict HCV and can be used as a non-invasive method to determine the risk factors of a disease. (AbuSharekh & Abu-Naser, 2018) used another approach to develop a model for the diagnosis of HCV. Using an ANN-based approach, they aimed to identify factors that play important roles in a diagnosis of HCV as well as to build a prediction model to diagnose the HCV based on some predetermined date. Their ANN-based model evaluation has shown that the ANN algorithm is able to predict the diagnosis of HCV with 98.44% of accuracy.

Data Collection

The Hepatitis C dataset that is used in this study has been referenced from (Hoffmann et al., 2018), which are available at the UCI machine learning repository. The dataset consists of 615 instances with 14 attributes, including the class attributes. The dataset contains the demographic values and laboratory values of blood donors and hepatitis C patients. The attributes information and details are depicted in Table 1.

Table 1: Attributes Information

Attributes	**Description**	**Value**
Patient ID	Patient ID number	Numerical (1, 2, 3, etc)
Category	The diagnosis	0=Blood Donor, 0s=Suspect Blood Donor, 1=Hepatitis, 2=Fibrosis, 3=Cirrhosis
Age	Patient's age (in years)	Numerical (within the range of 23 – 77)
Sex	Gender	m=male, f=female
ALB	Albumin	Numerical (14.9 – 82.2)
ALP	Alkaline phophatase	Numerical (11.3 – 416.6)
ALT	Alanine amino-transferase	Numerical (0.9 – 325.3)
AST	Asparte amino-transferase	Numerical (10.6 – 324)
BIL	Bilirubin	Numerical (0.8 - 209)
CHE	Choline esterase	Numerical (1.42 – 16.41)
CHOL	Total Cholesterol in liver	Numerical (1.43 – 9.67)
CREA	Creatinine	Numerical (8 – 1079.1)
GGT	Gamma-glutamyl transferase	Numerical (4.5 – 650.9)
PROT	Total Protein in liver	Numerical (44.8 – 86.5)

ISBN 978-623-99026-6-7

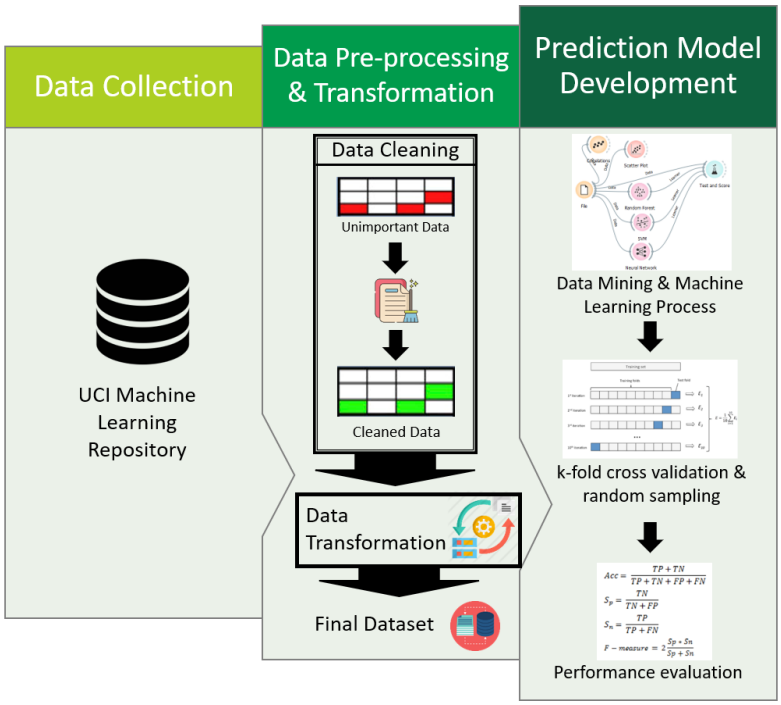
METHODOLOGY

The analytical flowchart of this study is described in this section using stages in KDD. Figure 1 shows the steps involved in developing the prediction model for this study.

Data Collection, Preprocessing, and Transformation

The dataset used in this study has been elaborated on in the previous section. After some investigations, of all 615 instances, there are 26 rows containing missing values. Therefore, all the rows have been eliminated to make sure the data is clean and consistent for further use. This cleaning process yields a total of 589 instances. Since the patient ID attribute only states the serial number of the data and does not contain any important information, therefore the attribute is not used in this study. The category attribute contains five different categories, and to make it easier to read, the attribute is then transformed into two categories which are hepatitis (hepatitis, fibrosis, cirrhosis) and non-hepatitis (blood donors, suspect blood donors).

Figure 1: Model Development Flow Chart



The final dataset contains 589 instances with 13 attributes which are then used as training and testing datasets to develop the model. The overview of the final dataset used in this study can be seen in Figure 2.

Figure 2: Overview of the first ten lines of the dataset

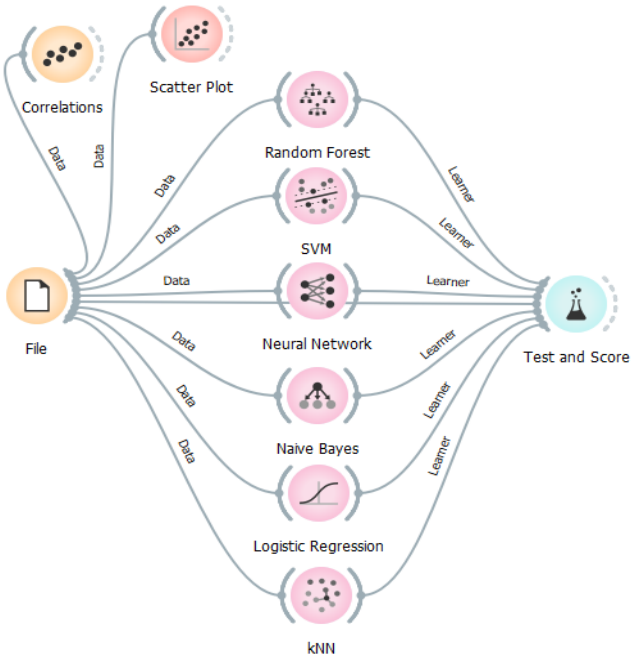
	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	Non-Hepatitis	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
2	Non-Hepatitis	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
3	Non-Hepatitis	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
4	Non-Hepatitis	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
5	Non-Hepatitis	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
6	Non-Hepatitis	32	m	41.6	43.3	18.5	19.7	12.3	9.92	6.05	111.0	91.0	74.0
7	Non-Hepatitis	32	m	46.3	41.3	17.5	17.8	8.5	7.01	4.79	70.0	16.9	74.5
8	Non-Hepatitis	32	m	42.2	41.9	35.8	31.1	16.1	5.82	4.60	109.0	21.5	67.1
9	Non-Hepatitis	32	m	50.9	65.5	23.2	21.2	6.9	8.69	4.10	83.0	13.7	71.3
10	Non-Hepatitis	32	m	42.4	86.3	20.3	20.0	35.2	5.46	4.45	81.0	15.9	69.9

Model Development

Data Mining and Machine Learning Process

The next stage in developing the prediction model is applying the machine learning algorithms to the final dataset. The data mining tool used in this study is the Orange tool. Orange is a data mining and machine learning suite used for data analysis through Python scripting as well as visual programming (Demšar et al., 2013). The classification process was conducted using Support Vector Machine, K-Nearest Neighbour, Random Forest, Logistic Regression, Naïve Bayes, and Neural Network algorithms. Figure 3 shows the configuration in the Orange tool. The resampling procedure to evaluate the model of this study is 5-fold cross-validation and random sampling with 70% of training set size and 30% of testing set size.

Figure 3: Model Development Configuration in Orange



Performance Evaluation

The performance of the above-mentioned algorithms is calculated using the following metrics (Bhargav et al., 2018):

Confusion Matrix

One of the easiest metrics to find the correctness of a machine learning model is the confusion matrix. It is used to evaluate the performance of a classification model, and the size of the matrix depends on the target classes of the dataset used. The confusion matrix compares the actual values with the values predicted by the machine learning model. Terms that are used in the confusion matrix are shown in Table 2.

Table 2: Confusion Matrix Description

CONFUSION MATRIX		**PREDICTED VALUES**	
		True (Positive)	**False (Negative)**
ACTUAL VALUES	**True (Positive)**	TP	FN
	False (Negative)	FP	TN

Accuracy, Precision, Recall, F1-Score

Other metrics that are used to evaluate the performance of the model are Accuracy, Precision, Recall, and F1-Score (Kaunang & Rotikan, 2018; Powers, 2020). Accuracy is the percentage of correctly classified instances. Precision is the fraction of correctly predicted positive instances to the total predicted positive instances. Recall shows the number of the actual positive values that were correctly predicted by the model. F1-score is the harmonic mean of Precision and Recall. In

other words, the metric used the combination of the other two metrics. Below is the equation of each metric:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} \quad (4)$$

Experimental Results

After all of the above-mentioned steps were done, different results were obtained. Tables 3 and 4 show the classification method results using 5-fold cross-validation and random sampling of 70% of the training set size.

Table 3: Experimental Results of Different Machine Learning Algorithms Using 5-fold Cross-Validation

Evaluation Metrics	Machine Learning Algorithms					
	kNN	**SVM**	**Random Forest**	**Neural Network**	**Naïve Bayes**	**Logistic Regression**
Accuracy	97.1%	97.1%	96.8%	97.8%	95.9%	98.3%
Precision	97.1%	97%	96.7%	97.7%	96.2%	98.3%
Recall	97.1%	97.1%	96.8%	97.8%	95.9%	98.3%
F1-score	96.9%	97%	96.7%	97.7%	96%	98.3%

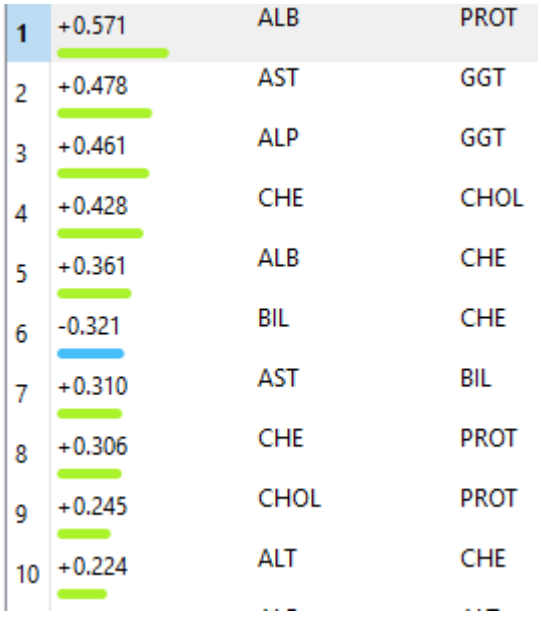
Table 4: Experimental Results of Different Machine Learning Algorithms Using Random Sampling

Evaluation Metrics	Machine Learning Algorithms					
	kNN	**SVM**	**Random Forest**	**Neural Network**	**Naïve Bayes**	**Logistic Regression**
Accuracy	96.5%	96.7%	97.3%	97.1%	96%	97.9%
Precision	96.4%	96.6%	97%	97%	96.3%	97.8%
Recall	96.5%	96.7%	97.1%	97.1%	96%	97.9%

F1-score	96.3%	96.5%	96.1%	96.9%	96.1%	97.8%

The aim of this study is to carry out a comparison of different machine learning algorithms to predict the Hepatitis C data. From the results, it can be seen that all algorithms generate good performance. However, among all the algorithms used in this study, Logistic Regression carried the optimum result of 97.9% accuracy, followed by RandomForest algorithm with 97.3% of accuracy. This shows that Logistic Regression can be used to predict Hepatitis C patients. Moreover, to show the correlations between attributes, the Pearson pairwise correlation was used. The correlations create 55 pairwise cases, and Figure 4 shows that the highest positive correlation is found between ALB and PROT attributes. This means that the number of Albumin in the blood can affect the Total Protein in the blood of a donor or patient.

Figure 4: Correlation between attributes



1	+0.571	ALB	PROT
2	+0.478	AST	GGT
3	+0.461	ALP	GGT
4	+0.428	CHE	CHOL
5	+0.361	ALB	CHE
6	-0.321	BIL	CHE
7	+0.310	AST	BIL
8	+0.306	CHE	PROT
9	+0.245	CHOL	PROT
10	+0.224	ALT	CHE

CONCLUSION

In this study, there are six different machine learning algorithms were chosen to predict the Hepatitis C dataset. Based on the performance of the algorithms, Logistic Regression turned out to be the best algorithm in order to build a prediction model for Hepatitis C disease. However, this conclusion needs further investigation due to an imbalance in the number of data between the two classes. Therefore, in the future study, a much deeper analysis like finding out the most informative attributes for this data as well as finding out a further correlation between the attributes.

Acknowledgement

This work was supported by Universitas Advent Indonesia.

REFERENCES

AbuSharekh, E. K., & Abu-Naser, S. S. (2018). *Diagnosis of hepatitis virus using artificial neural network*.

Agarwal, G. G., Singh, A. K., Venkatesh, V., & Wal, N. (2019). Determination of risk factors for hepatitis C by the method of random forest. *Annal of Infectious Disease and Epidemiology*, *4*(1).

Alter, M. J. (2007). Epidemiology of hepatitis C virus infection. *World Journal of Gastroenterology: WJG*, *13*(17), 2436.

Bailey Jr, D. E., Landerman, L., Barroso, J., Bixby, P., Mishel, M. H., Muir, A. J., Strickland, L., & Clipp, E. (2009). Uncertainty, symptoms, and quality of life in persons with chronic hepatitis C. *Psychosomatics*, *50*(2), 138–146.

Bhargav, K. S., Thota, D., Kumari, T. D., & Vikas, B. (2018). Application of machine learning classification algorithms on hepatitis dataset. *International Journal of Applied Engineering Research*, *13*(16), 12732–12737.

Bréchot, C. (1996). Hepatitis C virus. *Digestive Diseases and Sciences*, *41*(12), 6S-21S.

Clemens, J. M., Taskar, S., Chau, K., Vallari, D., Shih, J. W., Alter, H. J., Schleicher, J. B., & Mimms, L. T. (1992). *IgM antibody response in acute hepatitis C viral infection*.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., & Starič, A. (2013). Orange: data mining toolbox in Python. *The Journal of Machine Learning Research*, *14*(1), 2349–2353.

Firdaus, R., Saha, K., Biswas, A., & Sadhukhan, P. C. (2015). Current molecular methods for the detection of hepatitis C virus in high risk group population: A systematic review. *World Journal of Virology*, *4*(1), 25.

Hashem, S., Esmat, G., Elakel, W., Habashy, S., Raouf, S. A., Elhefnawi, M., Eladawy, M. I., & ElHefnawi, M. (2017). Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(3), 861–868.

Hoffmann, G., Bietenbeck, A., Lichtinghagen, R., & Klawonn, F. (2018). Using machine learning techniques to generate laboratory diagnostic pathways—a case study. *J Lab Precis Med*, *3*, 58. https://archive.ics.uci.edu/ml/datasets/HCV+data

Kao, H.-J., Weng, S.-L., Huang, K.-Y., Kaunang, F. J., Hsu, J. B.-K., Huang, C.-H., & Lee, T.-Y. (2017). MDD-carb: A combinatorial model for the identification of protein carbonylation sites with substrate motifs. *BMC Systems Biology*, *11*. https://doi.org/10.1186/s12918-017-0511-4

Kaunang, F. J., & Rotikan, R. (2018). Students' academic performance prediction using data mining. *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018*. https://doi.org/10.1109/IAC.2018.8780547

Lauer, G. M., & Walker, B. D. (2001). Hepatitis C virus infection. *New England Journal of*

Medicine, *345*(1), 41–52.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

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

Parkes, J., Guha, I. N., Roderick, P., & Rosenberg, W. (2006). Performance of serum marker panels for liver fibrosis in chronic hepatitis C. *Journal of Hepatology*, *44*(3), 462–474.

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*.

Saadeh, S., Cammell, G., Carey, W. D., Younossi, Z., Barnes, D., & Easley, K. (2001). The role of liver biopsy in chronic hepatitis C. *Hepatology*, *33*(1), 196–200.

Sandag, G. A., & Kaunang, F. (2019). Klasifikasi Fungsi Family Protein Transport Menggunakan Radial Basis Neural Network. *CogITo Smart Journal*, *5*(2), 203–214.

Shepard, C. W., Finelli, L., & Alter, M. J. (2005). Global epidemiology of hepatitis C virus infection. *The Lancet Infectious Diseases*, *5*(9), 558–567.

Somi, M. H., Etemadi, J., Ghojazadeh, M., Farhang, S., Faramarzi, M., Foroutan, S., & Soleimanpour, M. (2014). Risk factors of HCV seroconversion in hemodialysis patients in tabriz, iran. *Hepatitis Monthly*, *14*(6).

Uttreshwar, G. S., & Ghatol, A. A. (2009). Hepatitis B diagnosis using logical inference and generalized regression neural networks. *2009 IEEE International Advance Computing Conference*, 1587–1595.

Weng, S.-L., Huang, K.-Y., Kaunang, F. J., Huang, C.-H., Kao, H.-J., Chang, T.-H., Wang, H.-Y., Lu, J.-J., & Lee, T.-Y. (2017). Investigation and identification of protein carbonylation sites based on position-specific amino acid composition and physicochemical features. *BMC Bioinformatics*, *18*. https://doi.org/10.1186/s12859-017-1472-8

Ziol, M., Handra-Luca, A., Kettaneh, A., Christidis, C., Mal, F., Kazemi, F., De Lédinghen, V., Marcellin, P., Dhumeaux, D., & Trinchet, J. (2005). Non-invasive assessment of liver fibrosis by measurement of stiffness in patients with chronic hepatitis C. *Hepatology*, *41*(1), 48–54.