JOIN (Jurnal Online Informatika)
Volume 6 No.1 | June 2021: 73-78
DOI: 10.15575/join.v6i1.719
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Comparison of Machine Learning Classification Methods in Hepatitis C Virus

Lailis Syafa'ah[1], Zulfatman Zulfatman[2], Ilham Pakaya[3], Merinda Lestandy[4]
[1,2,3]Department of Electrical Engineering, Universitas Muhammadiyah Malang, Indonesia
[4]Department of D3 Electronics Technology, Universitas Muhammadiyah Malang, Indonesia

Article Info

Article history:

Received April 04, 2021
Revised May 03, 2021
Accepted May 05, 2021
Published June 15, 2021

Keywords:

Classification
HCV
KNN
Machine learning
Naïve Bayes
Neural network
Random forest

ABSTRACT

The hepatitis C virus (HCV) is considered a problem to the health of societies are the main. There are around 120-130 million or 3% of the world's total population infected with HCV. Without treatment, most major infectious acute evolve into chronic, followed by diseases liver, such as cirrhosis and cancer liver. The data parameters used in this study included albumin (ALB), bilirubin (BIL), choline esterase (CHE), γ-glutamyl-transferase (GGT), aspartate amino-transferase (AST), alanine amino-transferase (ALT), cholesterol (CHOL), creatinine (CREA), protein (PROT), and Alkaline phosphatase (ALP). This research proposes a methodology based on machine learning classification methods including k-nearest neighbors, naïve Bayes, neural network, and random forest. The aim of this study is to assess and evaluate the level of accuracy using the algorithm classification machine learning to detect the disease HCV. The result show that the accuracy of the method NN has a value of accuracy are high, namely at 95.12% compared to the method KNN, naïve Bayes and RF in a row amounted to 89.43%, 90.24%, and 94.31%.

Corresponding Author:

Merinda Lestandy,
Department of D3 Electronics Technology,
Universitas Muhammadiyah Malang,
Jl. Raya Tlogomas No. 246 Malang – Jawa Timur, Indonesia
Email: merindalestandy@umm.ac.id

1. INTRODUCTION

Hepatitis C virus (HCV) is the main pathogen that is carried through the blood to humans. There are about 120-130 million or 3% of the world's total population infected with HCV (Figure 1). According to the World Health Organization (WHO), every year there are about 3-4 million new cases of infection[1]. HCV has considered a problem of the health of societies are the main. This is because the virus in hepatitis is an etiological factor of chronic hepatitis which often develops into cirrhosis and hepatocellular carcinoma (HCC). In the country forward, the path of transmission of HCV that is most important is the abuse of drugs intravenously, whereas in resource-poor countries invasive procedures or injection-based therapy with contaminated instruments are a major source of new infections[2].

Without treatment, most major infectious acute evolve into chronic, followed by diseases liver, such as cirrhosis and cancer liver. Abuse of alcohol and syndrome Metabolic is a factor of prime which affect the development of diseases liver up and HCC[3]. Each year, approximately one-third of the transplanted liver is done on patients with complications were associated with infection with HCV, with cirrhosis decompensation or HCC[4]. In the decade following, the increase in the burden of hepatitis C is expected due to the aging population which is infected when it [5], [6]. During the period mentioned, the number of cases of cirrhosis related to hepatitis C is expected to increase by 31% and cancer liver about 50% [5] with the effect of an additional result the syndrome, metabolic[7], [8]. Therefore, infection with HCV is a problem of health public main that must be handled with the intervention policy of the strong to effectively identify and treat patients who are infected with HCV.

e-ISSN 2528-1682
p-ISSN 2527-9165

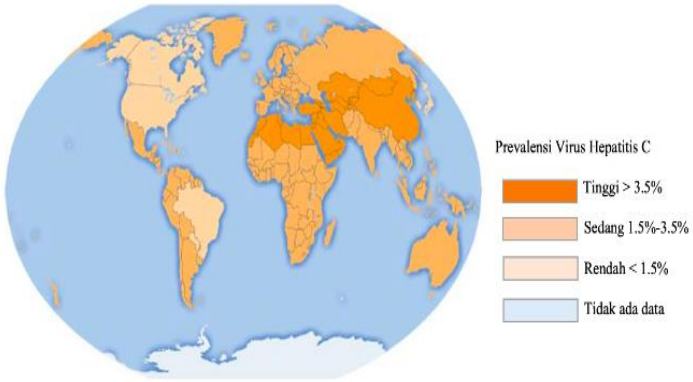


Figure 1. Spread of Hepatitis C Virus Infection in the world [1]

Research on the prediction and classification of HCV has been carried out in[9]–[12]. Various methods have been used to classify data including neural networks[9], decision trees[10], PSO[11], GA [11], logistic regression[12], and Support Vector Machine (SVM)[12]. The accuracy of which is derived from methods quite effective in generating the classification of HCV that is reaching 92% for the neural network[9], decision trees 75.3%[10], PSO 66.4%[11], GA 69.6%[11], Logistic Regression 79.4% [12], and SVM of 80%[12].

The aim of this study is to assess and evaluate the level of accuracy using the algorithm classification machine learning to detect the disease HCV. The method of classification machine learning was used, namely k-nearest neighbors (KNN), naïve Bayes, neural network (NN), and random forest (RF). The data in the study is that is derived from UCI Machine Learning. There are 10 data parameters used in this study including albumin (ALB), bilirubin (BIL), choline esterase (CHE), γ-glutamyl-transferase (GGT), aspartate amino-transferase (AST), alanine amino-transferase (ALT), cholesterol (CHOL), creatinine (CREA), protein (PROT), and Alkaline phosphatase (ALP).

2. METHODS

Data mining is the process of extracting new information based on data that can significantly improve the quality of clinical decisions and provide an important role for intelligent medical systems. Data mining is widely used in engineering including classification, clustering, regression, association analysis, and so on. The machine learning classification method used in this study can be seen in Figure 2.

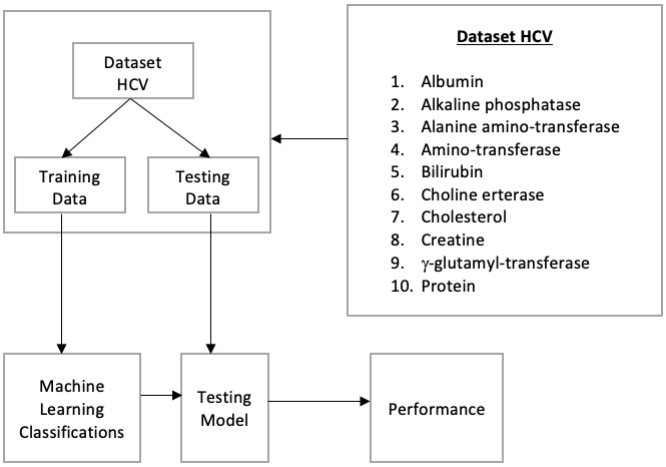


Figure 2. Research Method

2.1 Dataset

HCV Data derived from UCI machine learning with a total of as many as 73 patients (52 male, 21 female), aged 19 to 75 years with a diagnosis of serology and histopathology hepatitis C. As are shown in Table 1. exist 10 parameters of data used in the These studies include Albumin (ALB), bilirubin (BIL), choline esterase (CHE), γ-glutamyl-transferase (GGT), aspartate amino-transferase (AST), alanine amino-transferase (ALT), cholesterol (CHOL), creatinine (CREA), protein (PROT) and Alkaline phosphatase (ALP).

2.2 Classification Method

There are four methods of classification machine learning were used in the study is that the KNN, naïve Bayes, NN, and RF.

2.2.1 k-Nearest Neighbors (KNN)

Larose [13] stated that the KNN is a learning-based algorithm in which the training dataset is stored. KNN is one of the techniques of data mining the most much used in problem classification KNN plain called k-Memory Based Classification because the data training must be in memory at the time of the run-time[14]. Besides being used for classification, the KNN algorithm is also used for estimation and prediction. Calculation of distance Euclidean object to the data training that is given is expressed in Equation 1.

$$d(x,y) = \| x - y \| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

2.2.2 Naïve Bayes

Naïve Bayes is widely used for classification in machine learning. Naïve Bayes also be used for a lot of problems of classification for a simpler and provide accuracy that is better than the methods of machine learning more[15]. Naïve Bayes is a simple probabilistic class that learns from training data and then predicts test data based on the highest probability[16]. The calculation of the hypothetical probability on Naïve Bayes is stated in Equation 2.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

2.2.3 Neural Network

Neural Network (NN) or plain called neural network clone modeled by a network of nerve biologic are contained in the brain or the arrangement of the nerve center. The algorithm is used in machine learning algorithms and can be used for classification / supervised learning. Neurons and synapses are interconnected with each other which allows the passage of messages within them. The three main parts of a neural network are the input layer, hidden layer, and output layer. In NN, neurons are represented by nodes that contain a value that has a weight-specific and prepared plated using many layers of hidden so it can perform the classification based on the data that has been trained plus the use of the function activation[17]. The NN algorithm model can be seen in Figure 3.

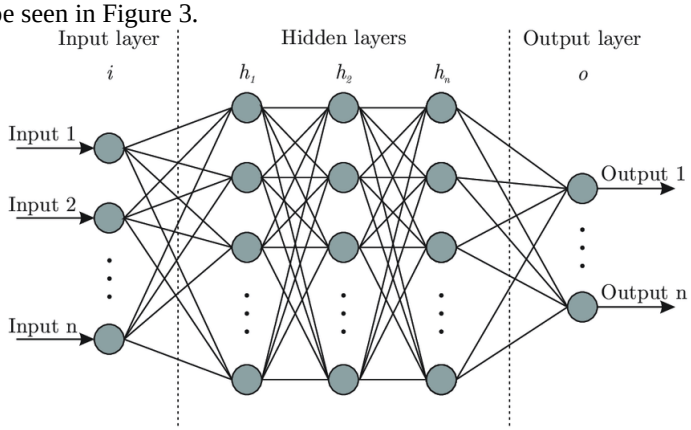
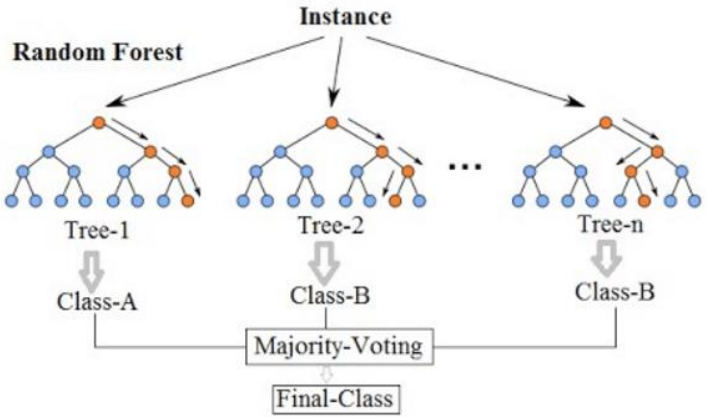


Figure 3. NN Models

2.2.4 Random Forest (RF)

RF comprises a collection of several decision trees like that shown in Figure 4. The higher the number of trees, the better the accuracy results will be obtained. RF uses C4.5 or J48 as classifier. In the year 2001, the RF introduced by Breiman, which combines Bagging by selecting features randomly to the tree decisions. RF is a supervised learning classification [18].



Picture 4. RF Models

2.3 Performance Evaluations

Measurement of the performance of the algorithm classification in research is that by using a confusion matrix. The confusion matrix shows the results of identification between the amount of data predictions are correct and the amount of data predictions are wrong compared with the fact that generated[19]. The confusion matrix table is shown in Table 2.

The parameters used for classification performance are accuracy, precision, and recall. Accuracy is the ratio of performance observation predicted by the right of the total observation. Accuracy can be calculated using Equation 3.

Table 1. Confusion matrix

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Where :
TP (true positive): correctly predicted hepatitis C positive data.
TN (true negative): correctly predicted hepatitis C negative data.
FN (false negative): positive data for hepatitis C which predicted hepatitis C negative data.
FP (false positive): negative data for hepatitis C which predicted hepatitis C positive data.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

Precision is the ratio of observed positive that predicted by the right of the total observation positives were predictable. To look for the value of precision used Equation 4.

$$Precision = \frac{TP}{TP+FN} \quad (4)$$

While the recall could be called sensitivity is the ratio of observation positive that predicted by the right for all the observations in the class that actually. Value recall can be calculated using Equation 5.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

3. RESULTS AND DISCUSSION

Algorithm's classification machine learning was examined in the study is that the KNN, naïve Bayes, NN, and RF using a dataset of HCV. Testing is done by dividing the composition of the training data by 80% and the test data by 20% from the dataset. The results of the classification of the four methods are compared to obtain a method that is appropriate to classify the disease HCV.

The comparison of the Confusion Matrix machine learning classification algorithm is shown in Table 2. The results of the classification algorithm are evaluated based on the results of accuracy, precision, and recall. KNN can classify with an accuracy value of 0.89, a precision value of 0.78, and a recall value of 0.64. The naïve Bayes classification method produces an accuracy value of 0.9, a precision value of 0.69, and a recall value of 0.72. Value of accuracy, the value of precision, and value recall method NN in a row at 0.95, 0.88, and 0.82. While the RF method gets an accuracy value of 0.94, a precision value of 0.83, and a recall value of 0.79.

Table 2. Comparison of confusion matrix between classification methods

Algorithm's	Accuracy	Precision	Recall
KNN	0,89	0,78	0,64
Naïve Bayes	0,9	0,69	0,72
NN	0,95	0,88	0,82
RF	0,94	0,83	0,79

A comparison of accuracy between classification methods is shown in Table 3. The decision trees method in [10] yielded an accuracy of 75.3%, the PSO method 66.4%[11], GA 69.4%[11], logistic regression 79.4%[12], and SVM 80%[12]. Methods of classification are used in the research is showing the results of the accuracy of which is relatively not much different. From the results of the simulation, implementation methods of KNN, naïve Bayes, NN, and RF for the prediction of disease Hepatitis C can perform repairs are indicated by the value of the accuracy of which is high. Rated accuracy of the method KNN, naïve Bayes, and RF in a row amounted to 89.43%, 90.24%, and 94.31%. When compared with the method KNN, naïve Bayes, RF, and[10]–[12], the method NN has a value of accuracies are high, namely at 95.12%.

Table 3. Comparison of accuracy between methods

Algorithm's	Accuracy
KNN	89,43%
Naïve Bayes	90,24%
NN	95,12%
RF	94,31%
Decision Trees[10]	75,3%
PSO[11]	66,4%
GA[11]	69,6%
Regresi Logistik[12]	79,4%
SVM[12]	80%

4. CONCLUSIONS

This research was successfully conducted by comparing the performance of several classification methods. The classification of HCV disease in this study was obtained from the UCI dataset which can be solved using the KNN, naïve Bayes, NN, and RF methods. Methods NN shows the results of the accuracy of the most well which amounted to 95.12% compared to KNN, naïve Bayes, and NN. Wherein each accuracy is at 89.43%, 90.24%, and 94.31%.

There are several suggestions that can be made for the development of further research, namely that it is necessary to increase the amount of training data so that the results of the evaluation of the model are more satisfying and the process of turning back can be carried out. model parameters to improve the accuracy of the classification methods.

ACKNOWLEDGEMENTS

The research team would like to express their gratitude and appreciation to the Faculty of Engineering, the University of Muhammadiyah Malang for their support for the implementation of this work through the Engineering Study and Engineering Center scheme, *Pusat Kajian dan Rekayasa Teknik* (PUSKAREKATEK 2020).

5. REFERENCES

[1] K. Mohd Hanafiah, J. Groeger, A. D. Flaxman, and S. T. Wiersma, "Global epidemiology of hepatitis C virus infection: New estimates of age-specific antibody to HCV seroprevalence," *Hepatology*, vol. 57, no. 4, pp. 1333–1342, 2013, doi: 10.1002/hep.26141.

[2] A. M. Hauri, G. L. Armstrong, and Y. J. F. Hutin, "The global burden of disease attributable to contaminated injections given in health care settings," *Int. J. STD AIDS*, vol. 15, no. 1, pp. 7–16, 2004, doi: 10.1258/095646204322637182.

[3] A. Alberti, "What are the comorbidities influencing the management of patients and the response to therapy in chronic hepatitis C?," *Liver Int.*, vol. 29, no. SUPPL. 1, pp. 15–18, 2009, doi: 10.1111/j.1478-3231.2008.01945.x.

[4] F. R. Ponziani, A. Gasbarrini, M. Pompili, P. Burra, and S. Fagiuoli, "Management of hepatitis C virus infection recurrence after liver transplantation: An overview," *Transplant. Proc.*, vol. 43, no. 1, pp. 291–295, 2011, doi: 10.1016/j.transproceed.2010.09.102.

[5] G. L. Davis, M. J. Alter, H. El-Serag, T. Poynard, and L. W. Jennings, "Aging of Hepatitis C Virus (HCV)-Infected Persons in the United States: A Multiple Cohort Model of HCV Prevalence and Disease Progression,"

Gastroenterology, vol. 138, no. 2, pp. 513-521.e6, 2010, doi: 10.1053/j.gastro.2009.09.067.

[6] H. Razavi *et al.*, "The present and future disease burden of hepatitis C virus (HCV) infection with today's treatment paradigm," *J. Viral Hepat.*, vol. 21, pp. 34–59, 2014, doi: 10.1111/jvh.12248.

[7] F. Kanwal *et al.*, "Increasing prevalence of HCC and cirrhosis in patients with chronic hepatitis C virus infection," *Gastroenterology*, vol. 140, no. 4, pp. 1182-1188.e1, 2011, doi: 10.1053/j.gastro.2010.12.032.

[8] Y. Arase *et al.*, "Sustained virological response reduces incidence of onset of type 2 diabetes in chronic hepatitis C," *Hepatology*, vol. 49, no. 3, pp. 739–744, 2009, doi: 10.1002/hep.22703.

[9] S. Ansari, I. Shafi, A. Ansari, J. Ahmad, and S. I. Shah, "Diagnosis of liver disease induced by hepatitis virus using artificial neural networks," *Proc. 14th IEEE Int. Multitopic Conf. 2011, INMIC 2011*, pp. 8–12, 2011, doi: 10.1109/INMIC.2011.6151515.

[10] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, "Using machine learning techniques to generate laboratory diagnostic pathways—a case study," *J. Lab. Precis. Med.*, vol. 3, pp. 58–58, 2018, doi: 10.21037/jlpm.2018.06.01.

[11] S. Hashem *et al.*, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 3, pp. 861–868, 2018, doi: 10.1109/TCBB.2017.2690848.

[12] G. Suwardika, "Pengelompokan Dan Klasifikasi Pada Data Hepatitis Dengan Menggunakan Support Vector Machine (SVM), Classification And Regression Tree (Cart) Dan Regresi Logistik Biner," *J. Educ. Res. Eval.*, vol. 1, no. 3, p. 183, 2017, doi: 10.23887/jere.v1i3.12016.

[13] S.-H. Wu, "Machine Learning Notation," *IEEE Softw.*, vol. 33, pp. 1–2, 2009, doi: 10.1109/MS.2016.114.

[14] E. Alpaydin, "Voting over Multiple Condensed Nearest Neighbors," *Artif. Intell. Rev.*, vol. 11, no. 1–5, pp. 115–132, 1997, doi: 10.1007/978-94-017-2053-3_4.

[15] R. Kurniawan, N. Yanti, M. Z. Ahmad Nazri, and Zulvandri, "Expert systems for self-diagnosing of eye diseases using Naïve Bayes," *Proc. - 2014 Int. Conf. Adv. Informatics Concept, Theory Appl. ICAICTA 2014*, pp. 113–116, 2015, doi: 10.1109/ICAICTA.2014.7005925.

[16] S. Tschiatschek, K. Paul, and F. Pernkopf, "Integer Bayesian network classifiers," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8726 LNAI, no. PART 3, pp. 209–224, 2014, doi: 10.1007/978-3-662-44845-8_14.

[17] M. M. S. Mishra, "A View of Artificial Neural Network," *IEEE Int. Conf. Adv. Eng. Technol. Res. (ICAETR - 2014), August 01-02, 2014, Dr. Virendra Swarup Gr. Institutions, Unnao, India*, no. c, pp. 5414–5420, 2014, [Online]. Available: https://ieeexplore.ieee.org/document/7012785.

[18] S. Kabiraj *et al.*, "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, pp. 1–4, 2020, doi: 10.1109/ICCCNT49239.2020.9225451.

[19] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digit. Signal Process. A Rev. J.*, vol. 17, no. 4, pp. 694–701, 2007, doi: 10.1016/j.dsp.2006.10.008.