Hepatitis classification using support vector machines and random forest

Jane Eva Aurelia, Zuherman Rustam, Ilsya Wirasati, Sri Hartini, Glori Stephani Saragih
Department of Mathematics, University of Indonesia, Indonesia

Article Info

Article history:

Received Mar 13, 2020
Revised Mar 20, 2021
Accepted Apr 17, 2021

Keywords:

Classification
Hepatitis
Machine learning
Random forest
Support vector machines

ABSTRACT

Hepatitis is a medical condition defined by inflammation of the liver. It can be caused by infection of the liver by hepatitis viruses or is of unknown aetiology. There are 5 main hepatitis viruses, such as virus types A, B, C, D and E. The infection may occur with limited or no symptoms, but also may include some symptoms like abdominal pain, dark urine, extreme fatigue, jaundice, nausea or vomiting. Because Indonesia is a large archipelago, the prevalence of viral infections varies greatly by region of acute hepatitis patients. This research uses data of hepatitis examination result with amount of 113 data and 5 features. The method that used is support vector machines (SVM) and random forest method. SVM is the classification method that uses discriminant hyper-plane, dividing to classes. meanwhile, random forest is a tree-based ensemble depending on a collection of random variables. SVM and random forest (RF) are applied to predict hepatitis data, and then the results will be compared.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Jane Eva Aurelia
Department of Mathematics
University of Indonesia
Jl. Prof. DR. Sudjono D. Pusponegoro, Pondok Cina, Depok, Jawa Barat 16424, Indonesia
Email: janeevaaurelia@gmail.com

1. INTRODUCTION

Hepatitis or known as an inflammation of the liver is a condition that can change to cirrhosis, fibrosis and liver cancer also can be self-limiting [1]. The highest common cause of hepatitis globally is hepatitis viruses but others can also cause hepatitis like autoimmune diseases and toxic substances [1].

There are 5 main types of hepatitis virus, such as virus types A, virus types B, virus types C, virus types D and also virus types E [1], [2]. Because of the encumbrance of ailment and death, these types are the greatest concern, also the possibility for outbreaks and visitation spread [1], [2]. Specially, in hundreds of millions of people, types B and C guide to chronic disease and also the most prevalent cause of cancer and liver cirrhosis [1], [2].

Hepatitis A and hepatitis E usually caused by ingestion of water and food contamination, where Hepatitis B, Hepatitis C and Hepatitis D is caused by infected body fluids which result of parenteral contact [3]. Blood contamination (products), equipment contaminated for medical procedures and transmittal from parent to child at nativity (or family members to kids), as also genital contexture are the prevalent modes of transmission for these viruses [3]. The infection may occur with limited or no symptoms, but also may include some symptoms like abdominal pain, dark urine, extreme fatigue, jaundice, nausea or vomiting [3].

Because Indonesia as a great archipelago, the predominance of viral infections varies exceptionally by territory of patients acute hepatitis [3] 43% to 68% infected by the virus of hepatitis A, 6% to 26% infected by the virus of hepatitis B, and 15% to 37% infected by the virus of non-hepatitis virus A nor B.

In 36% to 100% from a child of 5 year old, non-hepatitis virus A or called as Anti-HAV antibodies were detected [4]. In the general population, the prevalence of HBs-Ag has been estimated at 2.4% to 9.1% and as high as 17% whereas outside Java Island rated [4]. Patients in consort with carcinoma hepatocellular and liver cirrhosis were positive HBs-Ag at 37% to 52% [4].

Hepatitis C virus or called as HCV antibody was come across in 0.5% to 3.4% of blood donors, 10% to 16% of acute hepatitis, 21% to 41% of hepatocellular carcinoma and 31% to 74% of liver cirrhosis patients [4], [5]. In Indonesia, the two most important causes of chronic liver disease are HBV and HCV, although 25% to 29% of hepatocellular carcinoma and 14% to 25% of liver cirrhosis sufferer had no serologic substantiation for HBV or HCV [4], [5].

2. RESEARCH METHOD

2.1. Support vector machines

Support vector machines or known as SVM is a supervised machine learning model for two group classification problems that uses classification algorithms [6]. We'll able to categorize new examples after giving the model sets of label training data for either of two categories [7]. Let $\{x_i, x_j\}_i^N$ is the dataset where, $x_i \in R^D$ is feature of vector, y_i is class label for x_i and N is the number of samples [7]-[11]. This is main formula of support vector machines to find the best hyperplane:

$$f(x) = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{1}$$

To the hyperplane determining its orientation, that formula contains w (weight) as the orthogonal vector, b (bias) as the distance from the origin to the hyperplan, and x indicates the training sample [12]. The aim is to maximize the margin [13]. Moreover, SVM goal is construct the two planes, where the plane for the positive class is $\boldsymbol{w}^T\boldsymbol{x}_i + b \geq +1$, the plane for the negative class is $\boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1$. Figure 1 is an illustration of support vector machines [14].

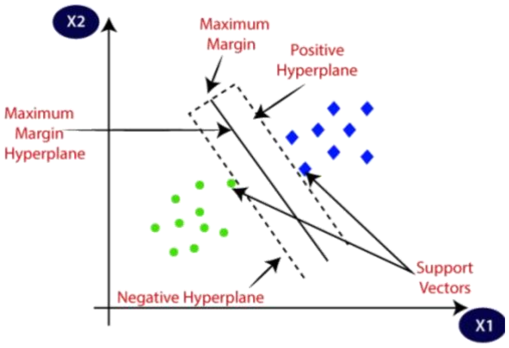

Figure 1. Illustration of support vector machines

The optimization problem of SVM can be summarized as:

Minimize,

$$\frac{1}{2}\|w\|^2 \tag{2}$$

$$\text{subject to, } y_i(\boldsymbol{w}^T \cdot x_i + b) \geq 1, \forall i = 1, \dots, N \tag{3}$$

By solving the problem above, formula of $\boldsymbol{w}$ and b are obtained in (4) and (5):

$$\boldsymbol{w} = \sum_{i=1}^{N} a_i y_i \boldsymbol{x_i} \tag{4}$$

$$b = \frac{1}{N_s}\sum_{i \in S}(y_i - \sum_{m \in S} a_m y_m \boldsymbol{x_m}) \tag{5}$$

Then, decision formulas of SVM can be written as:

$$f(x) = sign(\boldsymbol{w} \cdot \boldsymbol{x} + b) \quad (6)$$

In this study, kernel functions are used in support vector machines [15]. The kernel function resolves problems that are linear in order to be applied to non linear problems [16]. Especially, for algorithms express in inner product between two vectors [16]. There are several kernel functions with the parameters in Table 1 [15], [16].

Table 1. The several kernel function

No.	Name	Kernel function
1.	Linier	$K(x_i,x_j) = [\![x_i]\!]^T x_j.$
2.	Polynomial	$K(x_i,x_j) = [\![(t + [\![x_i]\!]^T x_j)]\!]^d.$
3.	Gaussian Radial Basis Function (RBF)	$K(x_i,x_j) = \exp(-\lVert x_i - x_j \rVert^2/\sigma^2).$

2.2. Random forest

Random forest or RF is a flexible and easy to use machine learning algorithm that produces [17]. A great result will produce most of the time, even without a hyper parameter tuning [17], [18]. Because RF's simplicity and diversity, random forest also one of the most used algorithms [18].

The random forest is a tree-based ensemble which is a combination of each decision tree depending on a collection of random variables [19]. The decision tree is a flowchart shaped like vector [20]. For a n-dimensional, the random vector $x = (x_1, x_2, \dots, x_p)^T$ represents the predictor variables and a random variable y represents the real-valued response [20]. Figure 2 is an illustration of random forest [20].

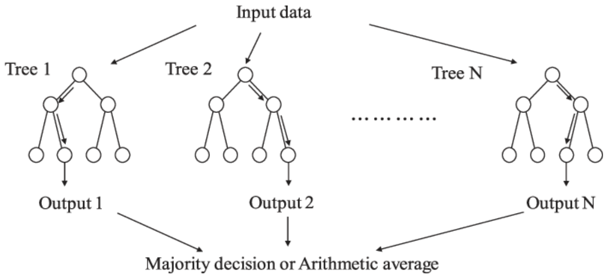

Figure 2. Illustration of random forest

Generated by random forests, most of the options depend on two data objects [21]. With replacement by sampling, for the current tree when the training set is drawn, one till third of the cases are left out of the sample which used to get a running unbiased of this data namely out of bag or OOB that estimate the error of the classification and also the variable importance [22], [23]. After being built, for each pair of cases, all of the data are run down the tree and proximities are computed [23]. The same terminal node occupied by two cases as their proximity by one is increased [23]. At the end of the run, the normalization of proximities is by the number of trees divided [24]. In locating outliers proximities are used, also the missing data replacement and illuminated producing low dimensional views of the data [24].

2.3. Confusion matrix

To calculate the accuracy, confusion matrix is used. The formula (7) for accuracy is [25]:

$$accuracy = \frac{T_p+T_N}{T_P+T_N+F_P+F_N} \quad (7)$$

T_P: Number of samples having hepatitis classified correctly.
F_P: Number of healthy people that were incorrectly classified to hepatitis.
F_N: Number of samples with hepatitis that were incorrectly classified as healthy.
T_N: Number of healthy individuals correctly spotted.

3. RESULTS AND ANALYSIS

3.1. Data

The data used in this study are from hepatitis patients who have inflammation in their liver. This data amounted to 113 data with 5 features proportioned as 90% training data and 10% testing data from the original data, with actual amounts of 84 major data and 29 minor data. Minor data represent data classes that indicate the presence of inflammation with label ‘1’ that used for the dataset, while the major data represent data classes that do not indicate inflammation with label '2' that used for the dataset. Table 2 explains the inflammation data features that will be examined.

Table 2. The several kernel function

No.	Feature	Definition of feature
1.	Gender	Sex (Male or Female)
2.	SGOT	Serum Glutamic Oxaloacetic Transaminase
3.	SGPT	Serum Glutamic Pyruvic Transaminase
4.	Anti-HCV	Non-Hepatitis C Virus
5.	Diagnosis	The Identification of the Nature of an Illness

3.2. Result

This research used training data diverse from 10% to 90%. Table 3 shows the results of the performa of accuracy of the entire method used and Table 4 shows the results of the comparison of running time of each method. As listed in Table 3, the best accuracy obtained was 99.55%, which resulted from the SVM model with a gaussian RBF kernel. Followed by, linear kernel SVM (99.13% accuracy) and random forest with 98.43% accuracy. Meanwhile, the lowest level of accuracy resulted from the SVM model with a polynomial kernel that was equal to 96.64%.

Gaussian RBF has the best accuracy 100% with 10%-40%, 60% and 80% training data. For linear kernel has the best accuracy 100% with 10%-20%, and 40%-50% training data, along with random forest at 10%-60% training data. On the other side, polynomial kernel has the best accuracy of 100% if the model uses 10%, 30% and 60% training data.

In Table 4, gaussian radial basis kernel gives the best performance with an average running time of 2.3158. Followed by, polynomial kernel SVM with an average running time of 2.3542 and linear kernel SVM with an average running time of 2.4578. Lastly is random forest with an average running time of 7.31.

Table 3. The performance of each method

Training data	Accuracy			
	SVM linear	SVM polynomial	SVM gaussian RBF	Random forest
10%	1.0	1.0	1.0	1.0
20%	1.0	0.9565	1.0	1.0
30%	0.9705	1.0	1.0	1.0
40%	1.0	0.9782	1.0	1.0
50%	1.0	0.9824	0.9824	1.0
60%	0.9852	1.0	1.0	1.0
70%	0.9875	0.9875	0.9875	0.9875
80%	0.9890	0.9890	1.0	0.9890
90%	0.9901	0.8039	0.9901	0.8823
Average	0.9913	0.9664	0.9955	0.9843

Table 4. The comparison of running time of each method

Training data	Running time(s)			
	SVM linear	SVM polynomial	SVM gaussian RBF	Random forest
10%	2.3331	2.3688	2.3570	7.31
20%	2.7447	2.3429	2.8153	7.31
30%	2.6945	2.8721	2.2034	7.31
40%	2.2015	2.2014	2.1982	7.31
50%	2.5598	2.3076	2.3521	7.31
60%	2.3193	2.1845	2.2880	7.31
70%	2.4280	2.1740	2.2993	7.31
80%	2.6149	2.3953	2.1403	7.31
90%	2.2248	2.3413	2.1886	7.31
Average	2.4578	2.3542	2.3158	7.31

All of the methods are good for classification of the presence of inflammation in the liver leading to hepatitis. The highest accuracy resulted with a value of 99% are from the linear kernel and the gaussian RBF. However, based on the accuracy and running time, the best method to classify hepatitis is gaussian RBF kernel SVM.

4. CONCLUSION

Predicting the presence of inflammation in the liver of a patient in diagnosing with machine learning can help medical staff to classify hepatitis disease. An early detection can make patients get the right treatment that helps them increase their life and reduce the risk. In this study, there are four method used in: SVM with linear, polynomial, gaussian RBF kernel, and random forest. The experimental results show that the performance of SVM classifiers and Random Forest method are properly and correctly predict the data. However, based on our results, if we see both of the performa and running time, support vector machine with gaussian RBF is the best one to classify Hepatitis data as we can see in Tables 3 and 4. Hopefully, in the future research, this method can be use with a larger dataset so can develop to give more better accuracy for predicting or classifying the other diseases.

ACKNOWLEDGEMENTS

This research full supported financially by University of Indonesia, with a PUTI PROSIDING 2020 research grant scheme (ID Nomor: NKB-976/UN2.RST/HKP.05.00/2020).

REFERENCES

[1] “What is hepatitis?” WHO, https://www.who.int/features/qa/76/en/ (accessed 2020).

[2] “Hepatitis-latest research and news” Nature, https://www.nature.com/subjects/hepatitis (accessed 2020).

[3] “Hepatitis in Indonesia” Springer, https://link.springer.com/chapter/10.1007/978-4-431-68255-4_99 (accessed 2020).

[4] A. D. Dearborn and J. Marcotrigiano, “Hepatitis C Virus Structure: Defined by What It Is Not,” in *Cold Spring Harbor Laboratory Press*, 2020 Jan 2;10(1):a036822., doi: 10.1101/cshperspect.a036822.

[5] D. Vergani, I. R. Mackay and G. M. Vergani, “Hepatitis,” in *Autoimmune Diseases (Sixth Edition)*, Academic Press, 2020, pp. 1117-1147, doi: 10.1016/B978-0-12-812102-3.00057-9.

[6] V. A. Afentoulis and K. I. Lioufi “SVM Classification with Linear and RBF kernels,” 2015, doi: 10.13140/RG.2.1.3351.4083

[7] Z. Rustam, D. A. Utami, R. Hidayat, J. Pandelaki and W. A. Nugroho, “Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infarction Datasets” in *International Journal on Advanced Science Engineering Information Technology*, vol. 9, no. 2, pp. 685-691, 2019, doi: 10.18517/ijaseit.9.2.8615

[8] C. Aroef, R. P. Yuda, Z. Rustam and J. Pandelaki, “Multinominal Logistic Regression and Support Vector Machine for Osteoarthritis Classification,” in *Journal of Physics: Conference Series*, vol. 1417, At. no. 012012, 2019, doi: 10.1088/1742-6596/1417/1/012012.

[9] Z. Rustam and N. Angie, “Prostate Cancer Classification Using Random Forest and Support Vector Machines,” *Journal of Physics: Conference Series*, vol. 1752, 2021, Art. no. 012043, doi: 10.1088/1742-6596/1752/1/012043.

[10] H. Tasman, Z. Rustam, N. A. Darmawan and R. E. Yunus, “Ischemic Stroke Classification Using Local Binary Pattern and Support Vector Machines,” in *International Conference on Recent Advances in Applied Mathematics*, 2020.

[11] L. Sun, B. Zou, S. Fu, J. Chen and F. Wang, “Speech Emotion Recognition Based on DNN-Decision Tree SVM Model,” in *Journal of Speech Communication*, vol. 115, pp. 29-37, 2019, doi: 10.1016/j.specom.2019.10.004.

[12] J. Abukhait and M. Obeidat, “Classification based on Gaussian-kernel Support Vector Machine with Adaptive Fuzzy Inference System,” in *Przegląd Elektrotechniczny*, vol. 1, no. 5, pp. 16-24, doi: 10.15199/48.2018.05.03.

[13] H. Subhani and S. Badugu, “A Study of Liver Disease Classification Using Data Mining and Machine Learning Algorithms,” in *Springer Nature Switzerland AG*, pp. 630-640, 2020, doi: 10.1007/978-3-030-24318-0_72.

[14] A. Tharwat, “Parameter investigation of support vector machine classifier with kernel functions,” 2019 in *Knowledge and Information Systems*, vol. 61, pp. 1269-1302, 2019, doi: 10.1007/s10115-019-01335-4.

[15] B. Bai, Z. Li, J. Zhang and W. Zhang, “Application of support vector machines-based classification extremum method in the flexible mechanisms,” in *ASME*, pp. 1-29, 2020, doi: 10.1115/1.4046210.

[16] Z. Rustam, S. Hartini, T. Siswantining, D. A. Utami and N. K. Putri, “Comparison Between Fuzzy Kernel C-Means, Fuzzy Kernel Possibilistic C-Means and Support Vector Machines in Soft Tissue Tumor Classification,” in *International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD)*, vol. 1103, pp. 92-105, 2019, doi: 10.1007/978-3-030-36664-3_11.

[17] “Random forests–classification description,” https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. (accessed 2020).

[18] “The Random Forest Algorithm: A Complete Guide,” https://builtin.com/data-science/random-forest-algorithm (accessed 2020).

[19] “Machine Learning Basics–Random Forest,” https://shirinsplayground.netlify.com/2018/10/ml_basics_rf/ (accessed 2020).

[20] Z. Rustam, N. A. Dermawan, J. Pandelaki and R. E. Yunus, “Acute sinusitis classification using support and fuzzy support vector machines,” *Journal of Physics: Conference Series*, vol. 1490, 2019, Art. no. 012029, doi: 10.1088/1742-6596/1490/1/012029.

[21] R. Katuwal, P. N. Suganthan and L. Zhang, “Heterogeneous Oblique Random Forest,” in *Journal of Pattern Recognition,* vol. 99, 2019, Art. no. 107078, doi: 10.1016/j.patcog.2019.107078.

[22] A. Arfiani and Z. Rustam, “Ovarian Cancer Data Classification Using Bagging and Random Forest,” in *AIP Conference Proceedings*, 2019, https://doi.org/10.1063/1.5132473.

[23] U. Aprilliani and Z. Rustam, “Osteoarthritis Disease Prediction Based on Random Forest,” in *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, doi: 10.1109/ICACSIS.2018.8618166.

[24] A. Subudhi, M. Dash and S. Sabut, “Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier,” In *Biocybernetics and Biomedical Engineering*, vol. 40, pp. 277-289, 2020, doi: 10.1016/j.bbe.2019.04.004.

[25] M. S. Ozigis, J. D. Kaduk, C. H. Jarvis, P. da C. Bispo and H. Balzter, “Detection of oil pollution impacts on vegetation using multifrequency SAR, multispectral images with fuzzy forest and random forest methods,” in *Environmental Pollution*, vol. 256, 2020, Art. no. 113360, doi: 10.1016/j.envpol.2019.113360.

BIOGRAPHIES OF AUTHORS

Jane Eva Aurelia was born in Jakarta, 19 June 1998. She is a final year student in the Department of Mathematics, University of Indonesia. She is currently working on her thesis, which is firmly about applied mathematics using machine learning. Also, Ms. Jane’s specialties in research are mostly about machine learning, mathematical modeling, and data mining.

Zuherman Rustam is an Associate Professor and a lecturer of the intelligence computation at the Department of Mathematics, University of Indonesia. He obtained his Master of Science in 1989 in informatics, Paris Diderot University, French, and completed his Ph.D. in 2006 from computer science, University of Indonesia. Assoc. Prof. Dr. Rustam is a member of IEEE who is actively researching machine learning, pattern recognition, neural network, artificial intelligence.

Ilsya Wirasati is a final year student in the Department of Mathematics, University of Indonesia, who is currently working on her thesis. Her research is firmly about applied mathematics using machine learning in medical field. Ms. Ilsya’s specialties in research are mostly about machine learning, mathematical modeling, and data mining.

Sri Hartini is a Bachelor of Science from the Department of Mathematics, University of Indonesia, who is also completing the Master of Science at the University of Indonesia and is currently pursuing a Ph.D. in intelligence computation. Ms. Hartini is passionately researching machine learning, computer vision, neural networks and deep learning in various fields.

Glori Stephani Saragih was born in Medan, 17 January 1997. She is a Bachelor of Science from Department of Mathematics, Universitas Indonesia, who is completing the Master of Science at Universitas Indonesia and is currently pursuing a Ph.D. in intelligence computation. Ms. Glori is currently a Process Improvement Manager in PT. Aplikasi Karya Anak Bangsa (Gojek). Her current research is machine on machine learning and neural network in various fields, especially medical and finance.