Classification of Dengue Haemorrhagic Fever (DHF) using SVM, naive bayes and random forest

R Arafiah*, F Hermin¹, I R Kartika², A Alimuddin³ and I Saraswati³

¹Department of Computer Science, Faculty of Mathematics and Natural Sciences, Jakarta State University, Jl. Rawamangun Muka East Jakarta, Indonesia
²Department of Chemistry, Faculty of Mathematics and Natural Sciences, Jakarta State University, Jl. Rawamangun Muka East Jakarta, Indonesia
³Electrical Engineering Departemen Faculty of Engineering, University of Sultan Ageng Tirtayasa, Jl Jendral Sudirman Km 03 Cilegon, Banten, Indonesia.

*riaarafiah@unj.ac.id

Abstract Handling Dengue Hemorrhagic Fever (DHF) becomes Indonesia's national priority. DHF is an infectious disease whose treatment requires precision and speed of diagnosis. Data mining can be used to build prediction diagnosing DHF disease with supporting database. This paper aims to predict DHF using SVM, Naive Bayes, and Random Forest and then compare it with the accuracy of the result of third method. The accurate DHF prediction system can be used to avoid the error of diagnosing DHF and the treatment of the disorder can be done more quickly and precisely. The input systems are the patient's medical records (i.e. temperature, spotting, rumple leed, and bleeding) and the output system is suffering from DHF or not.

1. Introduction

One of the problems of prevention and eradication of infectious diseases which until now is still a public health problem is Dengue Haemorrhagic Fever (DHF). This dengue fever vector is the widespread of Aedes aegypti and Aedes albopictus mosquitoes throughout Indonesia through repeated bites to vulnerable people. The cause of DHF is dengue virus. This virus is a family of Flaviridae and genus Flavivirus which has 4 serotypes of Den-1, Den-2, Den3 and Den-4 [1]. The course of dengue disease is difficult to predict. Clinical manifestations ranging from asymptomatic, symptomatic (dengue fever, dengue), DHF can be without shock or shock shock (SSD). Patients who, at the time of entry are in a good situation, at any time can fall into a state of shock (SSD), therefore the speed of diagnosis, monitoring, and close supervision are the keys to the success of DBD [2].

DHF prediction based on data can be used to solve this problem. The huge of data that is utilized in the system, it can improve predicted results. As in other areas, data growth is increasing every day. Every month, data growth is estimated to increase 20 times from the previous month [3]. This can lead to data stacking and potentially important data loss. Data mining or Knowledge discovery in Database (KDD) has the ability to search and locate the missing important data

In general, data mining has six functions [4] that is: prediction, description, classification, association, estimation, clustering. The function of data mining as a classification is a process to find a model or function to describe the class or concept of a data. This process is used to describe important data and can predict future data trends. To predict DHF required data mining functions as a
classification. Data mining algorithms commonly used in the classification include: Naïve Bayes, SVM, Random Forest. SVM is used for the detection of heart disease [5] with accuracy 85.10%. The SVM classifier is commonly used and shows a good performance [6]. Using SVN data mining algorithm for prediction of dengue disease with accuracy 90.42% [7]. Use of Naïve Bayes for prediction Dengue fever [8], detection heart disease [5] accuracy 86.42% [9]. While the use of random forest for prediction Dengue fever, with accuracy 0.76 [10]. This paper aims to predict DHF using SVM, Naïve Bayes, and Random Forest and then compare it with the accuracy of the result of third algorithm. All three algorithms are tested against the same data for DHF prediction systems using the Orange tool. The input systems are the patient's medical records (i.e. temperature, spotting, rumple leed, and bleeding) [11] and the output system is suffering from DHF or not. Based on the accuracy of the predicted result selected the best algorithm, the accurate DHF prediction system can be used to avoid the error of diagnosing DHF and the treatment of the disorder can be done more quickly and precisely. Furthermore, this paper will describe the review literature on data mining, SVM, Naïve Bayes; methods, results, conclusions and suggestions.

2. Literature review

2.1. Data mining

Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, machine learning to extract and identify useful information and related knowledge from large databases [11]. Stages performed on the data mining process begins from the selection of data from source data to target data, preprocessing stage to improve data quality, transformation, data mining as well as the stage of interpretation and evaluation resulting in output of new knowledge that is expected to give a better contribution.

Basic steps in preparing data mining [3]: (1) preparing data sets: selecting data, or focusing on sample data, we will look for patterns or relationships. (2) Cleaning data and processing data (3) Reducing data, in this process we will find useful features to represent data tailored to the goal. 4. Adjusting the purpose of data mining with existing data mining methods, such as clustering, regression, classification, fuzzy (5) Exploration analysis, models and hypotheses, the method to be used in finding patterns from a data set (6) Data mining process, looking for patterns and relations from data sets (7) Interpretation of patterns in the process, in this process the process can be iterative from 1 to 7. This process can also be equipped with visualization of the pattern obtained. (8) Using results from patterns and relationships found. This process involves examination and repair.

2.2. Support Vector Machine (SVM)

Classification technique SVM, has advantages [9]: (1) Better Accuracy as compared to other classifier (2) Easily handle complex nonlinear data points (3) Over fitting problem is not as much as other methods. SVM has disadvantages: (1) Computationally expensive (2) The main problem is the selection of right kernel function. For every dataset different kernel function shows different results. (3) As compared to other methods, training process takes more time. (4) SVM was designed to solve the problem of binary class. It solves the problem of multi class by breaking it into pair of two classes such as one-against-one and one-against-all.

2.3. Naïve Bayes

Naïve Bayes is the best classification compared to REP Tree, Random tree, J48 and SMO [10]. This algorithm has advantages when it is applied to large size data, but it is weak in attribute selection [12]. In some case Bayesian classifier is requires small training data set for classification, easier for implementation, fast to classify and more efficient but it is non sensitive to irrelevant features [12]. The main feature of this Naïve Bayes Classifier is a very strong assumption (naïf) of the independence of each condition / event, works very well compared to other classifier models, requires a small amount of training data to determine the required parameters estimates in the process of classification. The
assumption of independence between attributes reduces accuracy (since there is usually a linkage) [13]
the stages of the naive bayes algorithm process are: (1) computing the number of classes / labels (2)
counting the number of cases per class (3) multiplying all class variables 4) comparing the results per
class

2.4. Random Forest (RF)
Random Forest algorithm maintains best accuracy as compared to others classification system [13]. The
steps of the construction of the random forest algorithm are as follows:

2.4.1. Step1. Sampling k training subsets. In the first step, k training datasets are experimented from the
original training dataset S in a bootstrap selection manner. Namely, N records are selected from S by
a random sampling and replacement method in each sampling time. After the current step, k training
subsets are constructed as a collection of training subsets S Train: S Train = {S1, S2......Sk}. At the
same time, the records that are not to be selected in each sampling period are composed as an Out-Of-
Bag (OOB) dataset. In this way, k OOB sets are constructed as a collection of SOOB:
SOOB = {OOB1, OOB2... OOBk},
Where k ≪ N, Si ∩ OOBi = φ and Si ∪ OOBi = S.
To obtain the classification accuracy of each tree model, these OOB sets are used as testing sets after
the training process. [14]

2.4.2. Step 2. Constructing each decision tree model. In an RF model, each Meta decision tree is created
by a C4.5 or CART algorithm from each training subset Si. In the growth process of each tree, m feature
variables of dataset Si are randomly selected from M variables. In each tree node’s dividing process is
done, then gain ratio of each feature variable is computed, and the best one or most priority node is
chosen as the splitting node. This splitting process is repeated until a leaf node is generated. Finally, k
decision trees are trained from k training subsets in the same way. [14]

2.4.3. Step 3. Collecting k trees into an RF model. The k trained trees are collected into an RF model,
which is defined:

\[(X, \Theta j) = \sum_{i=1}^{k} h_i(x, \Theta j), (j = 1,2, ..., m)\],

where \(h_i(x, \Theta j)\) is a meta decision tree classifier, X is the input feature vector of the training dataset, and \(\Theta j\) is
an independently and identically distributed random vector that determines the growth process of the
tree [14].

3. Research method
The research data is secondary data from research of M. Syafii in 2006. The sample of research is taken
from medical record (medical record) hospital Dr. Sardjito Yogyakarta on 13-16 December 2005. The
sample consists of data of patients suffering from dengue fever (DD) as many as 213 patients. According
to international classification of diseases tenth revision (ICD 10) DHF disease is coded A.91 and DD
disease with code A.90. Samples were taken from all hospitalized patients from January to November
2005.

Based on clinical manifestations of DHF, the severity of DHF are divided into 4 grade [2]: grade I:
fever and haemorrhagic manifestations (positive tourniquet test); grade II: as grade I plus spontaneous
bleeding; grade III: as grade II plus circulatory failure, degree IV: as grade III plus profound shock with
undetectable BP and pulse. Then the independent variable which is the input system is the patient's
medical record i.e temperature, spotting, rumple leed (results tornikuet test), and bleeding. The
dependent variable which is the output of the system suffering from DHF or not.

Dengue Fever (DF) is was used as a comparison in this study due to consideration of 4 clinical criteria
and 1 laboratory criteria. Processing and data analysis using Orange 3.4.5 software. Orange is an open
source software for Data Analytics / Data Mining processing. Compared to other Data Mining software, orange excels in visualization or what we call visual programming. Orange provides many widgets that we put on the canvas / drawing board then we connect with other widget. This canvas media will make it easier for users to play with data and perform the process of analytics data intuitively. In addition to widgets, orange also provides some add-ons / modules for problems in certain domains such as Text Mining / Text Analytics, Bioinformatics, Network Data / Social Network, Maps Models, Prototypes Process, and more. Orange is a free data mining tool. The orange app can be downloaded at http://orange.biolab.si/download/.

4. Research results
By using patient medical data input d, which are; fever (temperature), presence of spotting, presence of bleeding, and tornikuet test and output data: diagnose the patient whether or not dengue fever. Data processing using Naïve bayes method with tools Orange, obtained in figure 1:

**Figure 1.** Widget data processing with orange for DHF prediction.

The widget does two things. First, it shows a table with different classifier performance measures, such as classification accuracy and area under the curve. Second, it outputs evaluation results, which can be used by other widgets for analyzing the performance of classifiers, such as ROC Analysis or Confusion Matrix.

**Figure 2.** Prediction test results using Naïve Bayes.

Classification Accuracy (CA) is the accurate percentage of data records classified correctly after testing on the classification results [15]. While recision or confidence is the proportion of positive predicted cases that are also positively correct on the actual data. Recall or sensitivity is the proportion of true positive cases that are correctly predicted positively [16]. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and the worst score at 0. The relative contribution of precision and recall to the F1 score are equal. Based on figure 2 the best prediction is Random Forest with Classification accuracy 0.796, precision 0.841 and Recall 0.822. The
area under curve (AUC) is calculated to measure the performance difference of the method used. AUC RF is 0.898 or has an accuracy rate between 0.80-0.90, meaning good classification [11]. The proportion between the predicted and actual class shows in confusion matrix. From figure 2 we can see the accuracy of the predicted result using Naïve Bayes, SVM, and Random Forest method respectively 77.5%, 56%, and 84.1%. In other words, the use of Naïve Bayes, SVM, and Random Forest algorithms has the wrong predicted respectively 22.5%, 44%, and 15.9%.

Figure 3. Confusion matrix.

Figure 3 shows Naïve Bayes model is able to predict false of 9.403 and true equal to 15.790. In other words accuracy prediction is equal to 69%. The performance of the classification Naïve Bayes algorithm in graphs, the accuracy of Naïve Bayes, SVM, and Random Forest method classification can be seen in the graph of Receiver Operating Characteristic (ROC) in figure 4.

Figure 4. Accuracy of model classification based on Receiver Operating Characteristic (ROC).

From figure 4, performance classification of Naïve Bayes, SVM, and Random Forest algorithm, respectively can be seen from the shape of the green, yellow, and blue curve. The best performance of the algorithm is Random forest, if the resulting curve approaches the baseline line or the line across from the 0.0 point (dotted line), and is good, if the curve approaches the point 0.1.

5. Conclusion and recommendation

By using input data are the patient's medical records (i.e. temperature, spotting, rumple leed, and bleeding) and output data is suffering from DHF or DD built DHF prevention system that uses 3 algorithms. Based on result systems, the best algorithm is Random Forest with classification accuracy 0.796, precision 0.841 and recall 0.822. The area under curve (AUC) is calculated to measure the
performance difference of the method used. AUC RF is 0.898 or has an accuracy rate between 0.80-0.90, meaning good classification. The Naive Bayes and SVM algorithms are usually good in classification and prediction, but in this case, exceptions occur. This may be due to lack of preprocessing of data before being used in the system. Usually, the use of Naive Bayes and SVM algorithms require preprocessing of data, whereas the Random Forest algorithm does not require.

Correlation test between variables can be done to choose the relationship between independent variables and bound. Adding new variables such as weather (i.e temperature, humidity, and rainfall) and other medical features (such as diarrhea) can be done to improve the accuracy of the model.

Acknowledgments
The researcher would like to thank the Directorate of research and community service of the ministry of research technology and higher education Directorate General of Higher Education for the foundation this research and supporting of the Institutes of research and community service, Jakarta State University, Indonesia.

References
[1] I Kurane 2007 Dengue hemorrhagic fever with special emphasis on immunopathogenesis Comp. Immunol. Microbiol. Infect. Dis. 30(5–6) pp. 329–340
[2] WHO 2011 Comprehensive guidelines for prevention and control of dengue and dengue haemorrhagic fever
[3] U Fayyad, G Piatetsky Shapiro and P Smyth 1996 From Data Mining to Knowledge Discovery in Databases AI Mag. 17(3) p. 37
[4] D T Larose and C D Larose 2014 Discovering Knowledge in Data
[5] A F Otoom, E E Abdallah, Y Kilani, A Kefaye and M Ashour 2015 Effective diagnosis and monitoring of heart disease Int. J. Softw. Eng. its Appl. 9(1) pp. 143–156
[6] P T Noi and M Kappas 2018 Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery Sensors (Switzerland) 18(1)
[7] A S Fathima and D Manimeglai 2012 Predictive Analysis for the Arbovirus-Dengue using SVM Classification Int. J. Eng. Technol. 2(3) pp. 521–527
[8] N Iqbal and M Islam 2017 Machine learning for dengue outbreak prediction: An outlook Int. J. Adv. Res. Comput. Sci. 8(1) pp. 93–102
[9] H Leopard, W Kipruto Cheruiyot and S Kimani 2016 A Survey and Analysis on Classification and Regression Data Mining Techniques for Diseases Outbreak Prediction in Datasets Int. J. Eng. Sci. pp. 2319–1813
[10] K Shaukat Dar and S M Ulya Azmeen 2015 Dengue Fever Prediction: A Data Mining Problem J. Data Mining Genomics Proteomics 06(03)
[11] R Arafiah and F Hermin 2018 Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method in Journal of Physics: Conference Series 948(1)
[12] A Sheshasaayee and G Thailambal 2017 Comparison of classification algorithms in text mining Int. J. Pure Appl. Math. 116(22) pp. 425–432
[13] Tate A, Gavhane U, Pawar J, Rajpurohit B and Deshmukh G B 2017 Prediction of Dengue, Diabetes and Swine Flu Using Random Forest Classification Algorithm Int. Res. J. Eng. Technol. 46 pp. 685–690
[14] Chen J, Li K, Tang Z, Bilal K, Yu S, Weng C and Li K 2017 A parallel random forest algorithm for big data in a spark cloud computing environment IEEE Transactions on Parallel & Distributed Systems (1) 1-1.
[15] Han J, Pei J, and Kamber M 2011 Data mining: concepts and techniques Elsevier
[16] B S Bloom, M D Englehard, E J Furst, W H Hill and D R Krathwohl 1956 Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I Cognitive Domain Longmans Green Co LTD 16 p. 207