Healthcare Data Mining: Predicting Hospital Length of Stay of Dengue Patients

Iwan Inrawan Wiratmadja, Siti Yaumi Salamah & Rajesri Govindaraju*

Department of Industrial Engineering, Faculty of Industrial Technology
Institut Teknologi Bandung, Jalan Ganesha 10, Bandung 40132, Indonesia
*E-mail: rajesri_g@ti.itb.ac.id

Abstract. Dengue is regarded as the most important mosquito-borne viral disease. Recently dengue has emerged as a public health burden in Southeast Asia and other tropical countries. At times when dengue re-emerges as an epidemic, hospitals are required to be able to handle patient flow fluctuation while maintaining their performance. This research applied a data mining technique to build a model that can predict in-patient hospital length of stay from the time of admission, which can be useful for effective decision-making that may lead to better clinical and resource management in hospitals. Using the C4.5 algorithm and a decision tree classifier, an accuracy of 71.57% and an area under the receiver operating characteristic (ROC) curve value of 0.761 were obtained. The decision tree showed that only 7 out of 21 input attributes affect the hospital length of stay prediction of dengue patients. The attribute with the highest impact was monocytes, followed by diastolic blood pressure, hematocrit, leucocytes, systolic blood pressure, comorbidity score, and lymphocytes. In this research also a prototype of a prediction system using the resulting model was developed.

Keywords: *data mining; decision tree; dengue; hospital; length of stay; prediction.*

1 Introduction

Dengue is one of the most prevalent mosquito-borne viral diseases in humans. Dengue is considered a national public health concern in Indonesia. It is one of the diseases that trigger the highest numbers of in-patients at hospitals. However, the number of the patients fluctuates because of the fluctuation of the number of dengue fever and dengue hemorrhagic fever incidents as well as seasonal increase and decrease in the number of patients. Hence, hospitals are required to be able to handle a dramatic increase in patient influx while maintaining their performance. Due to the limited resources of hospitals, including medical facilities and medical staff, a significant increase in the number of patients hospitalized over some period increases the probability that a patient will undergo a longer hospital stay because of the chance that the medical service quality that the hospital can provide is lower. This situation has been observed in a local hospital in Bandung, West Java, Indonesia. Therefore in this study, a prediction model was developed for this

Received April 5th, 2017, Revised March 4th, 2018, Accepted for publication March 6th, 2018.
 ISSN: 2337-5779, DOI: 10.5614/j.eng.technol.sci.2018.50.1.8

hospital to support the decision-making process and enable better clinical and resource management.

Hospital length of stay correlates positively with the cost of hospitalization. Therefore, the ability to control and identify factors affecting in-patient length of stay could be a huge advantage for many hospitals. A system that can predict in-patient length of stay at the time of admission provides important benefits.

There are huge amounts of healthcare data available in hospitals, including dengue patient data. These data can be used for decision-making if they are mined to obtain useful information. Data mining can be employed to discover patterns in hospital data to accurately predict length of stay. Data mining has been used to identify patterns and gain knowledge from dengue patient datasets for several different purposes. Farooqi and Ali [1] applied data mining to classify dengue patients into dengue fever (DF) and dengue hemorrhaghic fever (DHF) patients. Data mining has also been used to classify patients into five levels of dengue disease, i.e. DF, DHF I, DHF II, DHF III, and DHF IV [2-4]; to detect the day of defervescence [3]; to diagnose dengue disease by classifying patients into dengue and non-dengue patients [5-7]; and to predict the probability of dengue disease severity [7].

Several studies have been conducted to predict hospital length of stay, using various data mining techniques. These researches either used a dataset of in-patients in general [8] or focused on predicting length of stay of patients with a specific diagnosis [9-15]. Azari, *et al.* [8] proposed an approach to predict hospital length of stay by clustering datasets and performing length of stay classification using various classifier models. These models were then statistically compared by combining the values of accuracy, Kappa statistic, precision, recall, and area under receiver operating characteristic (ROC) curve. The present research focused on certain attributes, i.e. specialty of the service provider, days elapsed since the first act of the year, primary diagnosis code, and categorized comorbidity score.

Blais, *et al.* [9] employed a multivariate analysis to predict length of stay for an acute care medical psychiatric in-patient service. The results indicated that the factors for estimating length of stay were available at the time of admission. Combes, *et al.* [10] identified models based on linear regression to predict length of stay in a pediatric emergency department. Even though a simple model was obtained, it suffers from the disadvantage of linearity assumption. Hachesu, *et al.* [11] compared several data mining techniques to predict length of stay of cardiac patients, resulting in various degrees of accuracy.

Table 1 Related work.

References	Objective	Dataset	Methods
Farooqi and Ali [1]	Classify DF and DHF	Dengue patients	Naïve Bayesian, KNN, decision tree, MLP, SVM
Farooqi, *et al.* [2]	Classify DF and DHF I-IV	Dengue patients	Decision tree
Thitiprayoonwongse, *et al.* [3]	Detect the day of defervescene; Classify DF and DHF I-IV	Dengue patients	Decision tree, Fuzzy logic
Thitiprayoonwongse, *et al.* [4]	Classify DF and DHF I-IV	Dengue patients	Decision tree
Kumar [5]	Diagnose dengue	Dengue patients	Alternating decision tree
Shakil, *et al.* [6]	Diagnose dengue	Dengue patients	Decision tree
Tanner, *et al.* [7]	Classify the probability of dengue in early phase; Predict the probability of dengue disease severity	Dengue patients	Decision tree
Azari, *et al.* [8]	Predict LOS		Various data mining models
Blais, *et al.* [9]	Predict LOS	Acute care medical psychiatric patients	Statistical multivariate analysis
Combes, *et al.* [10]	Predict LOS in Emergency Department	Pediatric Emergency Department patients	Regression models
Hachesu, *et al.* [11]	Predict LOS	Coronary artery disease (CAD) patients	Decision tree, SVM, ANN
Lella, *et al.* [12]	Predict LOS		Novel method (growing neural gas), SOM, OneR, ZeroR, C4.5
Liu, *et al.* [13]	Compare different DM techniques to predict LOS	Clinics, stroke patients	Naïve Bayesian, Decision tree
Tanuja, *et al.* [14]	Compare different DM techniques to predict LOS		ANN, NB, KNN, C4.5
Yang, *et al.* [15]	Predicting LOS	Burn patients	SVM, regression model

Lella, *et al.* [12] presented a novel unsupervised length of stay prediction model, which turned out to have higher accuracy compared to several other standard data mining methods. This study also considered only attributes provided at admission. A sampling procedure was conducted to speed up the training process. Liu, *et al.*

[13] conducted a comparative analysis of data mining algorithms to predict in-patient length of stay using two different datasets: a clinical dataset and a stroke dataset. The result demonstrated that the naïve Bayesian and decision tree algorithm performed much better than a neural network approach. Tanuja, *et al.* [14] compared different data mining techniques to predict length of stay using patient physiological measurements, demographic details, and lab test results as input. In terms of accuracy, multilayer back propagation performed better than naïve Bayes, K-NN, and C4.5. Yang, *et al.* [15] studied length of stay prediction of burn patients and demonstrated that artificial intelligence-based prediction techniques are more effective than regression techniques. An overview of the previous researches reviewed in this study is shown in Table 1.

Numerous studies have been conducted to predict length of stay, as mentioned above. Also, numerous researches applied data mining on dengue patient datasets. However, limited research is available on predicting length of stay of dengue patients. Although studies have been done that considered length of stay prediction for diseases in general, these studies cannot directly show the factors that affect length of stay for dengue patients in particular. Thus, this research was aimed at discovering the most important factors that influence length of stay of patients diagnosed with DF and DHF using data mining.

A comparative analysis of some data mining algorithms demonstrated the applicability and suitability of the decision tree algorithm in predicting in-patient length of stay [13]. Studies also showed that decision trees are reliable and have better accuracy in clinical decision-making [11]. By comparing the advantages and disadvantages of various classification techniques [16], decision tree turned out to be the most advantageous technique in view of interpretability, which is very important to enable physicians to understand and clarify the result of the length of stay prediction. It can also easily process high-dimension data and handles both numerical and categorical data [16].

A decision tree can be constructed using several algorithms, including CART (classification and regression tree) and C4.5, among others. These algorithms differ in the way they select splits and stop the tree from splitting [17]. Lim, *et al.* [18] showed that the C4.5 algorithm produces good classification accuracy and is considered the fastest algorithm in processing among similar data mining or machine learning algorithms. C4.5 also has the ability to split into more than two branches of categorical data, while CART can only process binary output [17]. Therefore, considering the suitability of C4.5, this algorithm was employed to predict length of stay of dengue patients in this research.

The main objective of this research was to develop a decision tree model to predict length of stay for hospitalized dengue fever and dengue hemorrhagic fever patients

using admission data from a hospital in Bandung. This model can be useful for effective decision-making and enable better clinical and resource management in the hospital. This research further developed a prototype of a prediction system that applies the proposed model so that it can be used directly by the hospital. The remainder of this paper is organized as follows: Section 2 describes the literature study on factors that may affect length of stay of dengue patients, Section 3 presents the methodologies adopted in this research, Section 4 discusses the results, while Section 5 summarizes the conclusions and discusses the limitations of this research as well as suggestions for future work.

2 Factors for Dengue Patients Length of Stay Prediction

A number of attributes, i.e. a mixture of demographic attributes, illness attributes, and treatment attributes obtained at the time of admission, can be assessed to estimate length of stay and further incorporated into clinical decision-making [9]. Several studies related to dengue were reviewed to select initial attributes/factors that may affect length of stay of dengue patients. A summary of our literature review on these factors is presented in Table 2.

3 Methods

Thorough data pre-processing was conducted to prepare the input data before building the predictive model. After pre-processing the data, the dataset was randomly split into two subsets for training and testing respectively. One subset of the data was trained to classify patient LOS into three groups, i.e. short, medium, and long, while another subset was used to test the accuracy of the model. A flowchart illustrating the approach used in this research is shown in Figure 1.

3.1 Dataset

This research was conducted using demographic and illness- or health-related data of 370 dengue fever (DF) and dengue haemorrhagic fever (DHF) patients of Dr. M. Salamun Hospital, a local hospital located in Bandung, Indonesia. This dataset was collected from the hospital database system and medical records.

Based on the literature study, factors that appear frequently in studies related to dengue were chosen as initial input attributes. These attributes were validated by medical personnel to ensure that no factors unrelated to length of stay were used as predictors. Two attributes were ruled out by the hospital's physician as they were considered not related to dengue patient length of stay in this hospital: day of admission and time of admission. A total of 27 attributes available at the time of admission, including patient personal data, clinical symptoms, and laboratory features, were selected as input attributes, whereas the attribute 'Length of stay'

was identified as the target attribute. The selected attributes are presented in Table 3.

Table 2 Factors for dengue patient length of stay prediction.

Attributes/ Factors	References
Age	Carrasco, *et al.* [19], Tanuja, *et al.* [14], Yang, *et al.* [15], Xiao, *et al.* [20], Lee, *et al.* [21]
Gender	Carrasco, *et al.* [19], Yang, *et al.* [15], Lee, *et al.* [21]
Blood Pressure	Tanuja, *et al.* [14], Farooqi & Ali [1], Lee, *et al.* [21]
Pulse	Thein, *et al.* [22], Tanuja, *et al.* [14], Farooqi, *et al.* [2], Kumar [5], Farooqi & Ali [1], Lee, *et al.* [21]
Temperature	Tanuja, *et al.* [14], Tanner, *et al.* [7], Shakil, *et al.* [6], Farooqi & Ali [1], Lee, *et al.* [21]
Haemoglobin	Aroor, *et al.* [23], Kumar [5], Shakil, *et al.* [6], Lee, *et al.* [21]
Leucocytes	Aroor, *et al.* [23], Ho, *et al.* [24], Carrasco, *et al.* [19], Thein, *et al.* [22], Kalayanarooj, *et al.* [25], Thitiprayoonwongse, *et al.* [3], Farooqi, *et al.* [2], Kumar [5], Tanner, *et al.* [7], Shakil, *et al.* [6], Farooqi & Ali [1], Lee, *et al.* [21]
Haematocrit	Carrasco, *et al.* [19], Tanuja, *et al.* [14], Thitiprayoonwongse, *et al.* [3], Thitiprayoonwongse, *et al.* [4], Farooqi, *et al.* [2], Tanner, *et al.* [7], Shakil, *et al.* [6], Farooqi & Ali [1], Lee, *et al.* [21]
Platelet	Aroor, *et al.* [23], Ho, *et al.* [24], Thein, *et al.* [22], Tanuja, *et al.* [14], Thitiprayoonwongse, *et al.* [3], Thitiprayoonwongse, *et al.* [4], Farooqi, *et al.* [2], Tanner, *et al.* [7], Shakil, *et al.* [6], Farooqi & Ali [1], Lee, *et al.* [21]
Neutrophil	Kalayanarooj *et al.* [25], Tanner, *et al.* [7]
Lymphocytes	Thitiprayoonwongse, *et al.* [3], Tanner, *et al.* [7], Lee, *et al.* [21]
Monocytes	Kalayanarooj, *et al.* [25]
Dengue IgM	Farooqi & Ali [1]
Dengue IgG	Tanner, *et al.* [7], Farooqi & Ali [1]
SGOT (AST)	Aroor, *et al.* [23], Ho, *et al.* [24], Kalayanarooj, *et al.* [25], Thitiprayoonwongse, *et al.* [3], Thitiprayoonwongse, *et al.* [4], Lee, *et al.* [21]
SGPT (ALT)	Aroor, *et al.* [23], Ho, *et al.* [24], Kalayanarooj, *et al.* [25], Thitiprayoonwongse, *et al.* [3], Lee, *et al.* [21]
Creatinine	Thein, *et al.* [22], Tanuja, *et al.* [14], Lee, *et al.* [21]
Bleeding Manifestation	Aroor, *et al.* [23], Thitiprayoonwongse, *et al.* [3], Thitiprayoonwongse, *et al.* [4], Farooqi, *et al.* [2], Farooqi & Ali [1], Lee, *et al.* [21]
Nausea/vomiting	Aroor, *et al.* [23], Ho, *et al.* [24], Carrasco, *et al.* [19], Kalayanarooj, *et al.* [25], Premaratna, *et al.* [26], Shakil, *et al.* [6], Lee, *et al.* [21]
Diarrhea	Aroor, *et al.* [23], Ho, *et al.* [24], Shakil, *et al.* [6]
Abdominal pain	Aroor, *et al.* [23], Carrasco, *et al.* [19], Thein, *et al.* [22], Thitiprayoonwongse, *et al.* [3], Thitiprayoonwongse, *et al.* [4], Shakil, *et al.* [6], Lee, *et al.* [21]
Joint/muscle pain	Premaratna, *et al.* [26], Shakil, *et al.* [6], Farooqi & Ali [1]
Headache	Ho, *et al.*[24], Premaratna, *et al.* [26], Kumar [5], Shakil, *et al.* [6], Farooqi & Ali [1]
Rash	Aroor, *et al.* [23], Ho, *et al.* [24], Premaratna, *et al.* [26], Thitiprayoonwongse, *et al.* [4], Farooqi & Ali [1], Lee, *et al.* [21]
Diagnosis	Azari, *et al.* [8]
Comorbidity score	Thein, *et al.* [22], Kalayanarooj, *et al.* [25], Azari, *et al.* [8], Yang, *et al.* [15], Xiao, *et al.* [20], Thitiprayoonwongse, *et al.* [4], Lee, *et al.* [21]
Febrile days before admission	Ho, *et al.* [24], Carrasco, *et al.* [19], Azari, *et al.* [8]
Day of admission	Xiao, *et al.* [20]
Time of admission	Xiao, *et al.* [20]

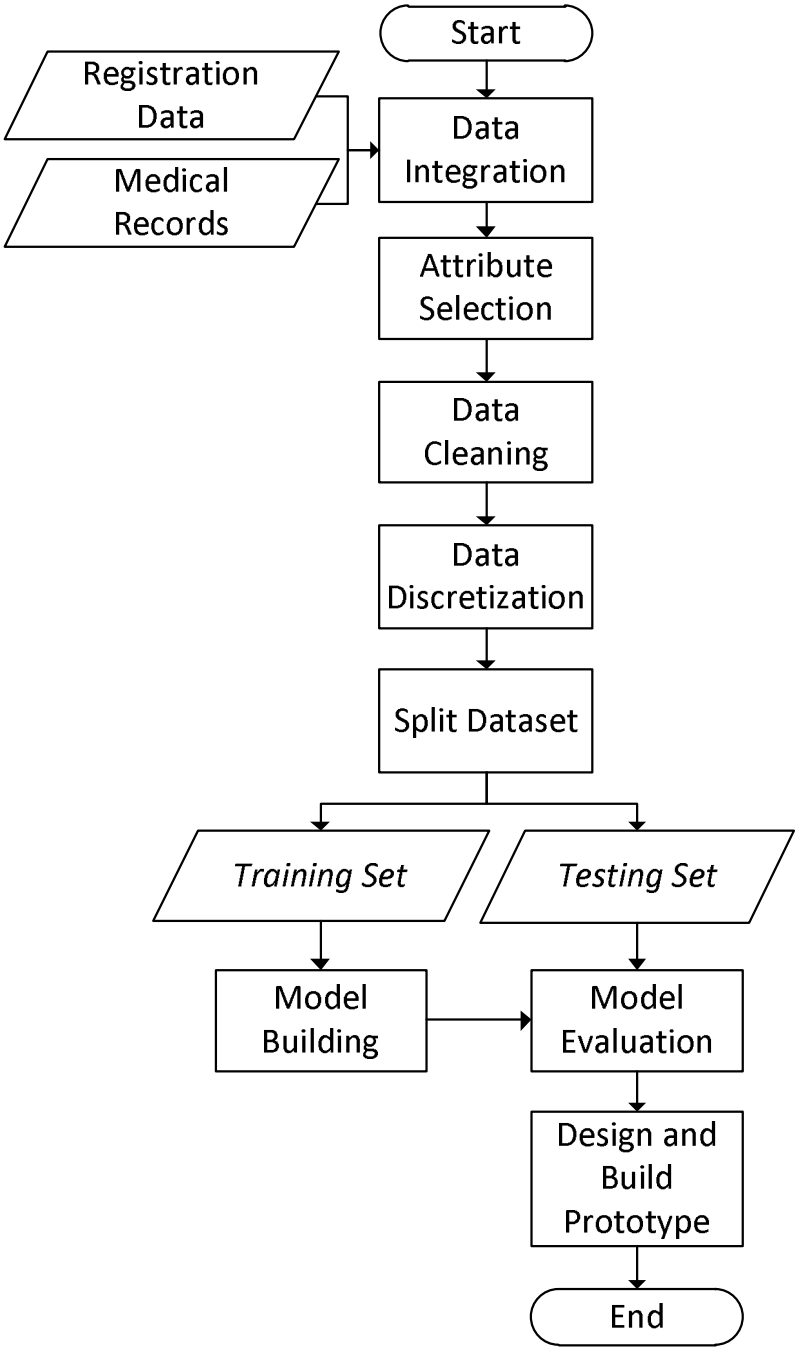


Figure 1 Methodology.

Table 3 Attributes characteristics of the length of stay prediction dataset.

No.	Attributes	Unit	Type	Value	Missing
		Input Attributes			
1	Age	Year	Numeric	0-85	0%
2	Gender		Nominal	L; P	0%
3	Systolic blood /	mmHg	Numeric	50-170	13%
	Diastolic blood pressure	mmHg	Numeric	42-110	13%
4	Pulse	bpm	Numeric	20-152	13%
5	Temperature	C	Numeric	33.1-39.6	4%
6	Haemoglobin	g/dL	Numeric	9.2-21.5	7%
7	Leucocytes	/mm3	Numeric	1100-18400	7%
8	Haematocrit	%	Numeric	27-61	1%
9	Platelet	/mm3	Numeric	16000-282000	1%
10	Neutrophil	%	Numeric	21-89	33%
11	Lymphocytes	%	Numeric	4-73	32%
12	Monocytes	%	Numeric	2-25	33%
13	Dengue IgM		Numeric	Positive; negative	48%
14	Dengue IgG		Numeric	Positive; negative	48%
15	SGOT (AST)	U/L	Numeric	12-1281	71%
16	SPGT (ALT)	U/L	Numeric	10-621	71%
17	Creatinine	mg/dL	Numeric	0.44-1.9	93%
18	Bleeding manifestation		Nominal	Yes; no	33%
19	Nausea/vomiting		Nominal	Yes; no	14%
20	Diarrhea		Nominal	Yes; no	85%
21	Abdominal pain		Nominal	Yes; no	43%
22	Joint/Muscle pain		Nominal	Yes; no	68%
23	Headache		Nominal	Yes; no	63%
24	Rash		Nominal	Yes; no	57%
25	Diagnosis		Nominal	DF; DHF	0%
26	Comorbidity score		Numeric	0-7	0%
27	Febrile day(s) before admission	Day	Numeric	2-15	0%
		Target Attribute			
28	Length of stay	Day	Numeric	1-11	0%

3.2 Data Pre-processing

3.2.1 Data Cleaning

One of the characteristics of hospital data is their incompleteness. There are many attributes with missing values that have to be processed to obtain an accurate model. The handling of missing values in this research adopted the techniques used

in Azari, *et al.* [8] and Hachesu, *et al.*[11]. Attributes and records with more than 50% missing values were removed from the dataset, while imputation was used for attributes with missing values below 50%. Imputation was done by replacing the missing data with the mean value of its length of stay class for numeric attributes and the mode for nominal attributes. The new constant value 'Unknown' was used to replace missing values of categorical attributes that were excluded from some of the data intentionally. These techniques were applied based on the characteristics of each attribute.

3.2.2 Data Discretization

This research employed a classification tree model, where the target attribute is of a nominal or categorical type. The raw length of stay attribute is a numeric attribute. Therefore, a discretization procedure was done to change the attribute type from numeric to nominal. Tanuja, *et al.* [14] used expert judgment to classify length of stay considering ease of analysis. Azari, *et al.* [8] applied the equal frequency technique, which resulted in unacceptable intervals of classes. Consequently, expert judgment was employed to categorize patient length of stay into three classes considering the level of intervention needed during hospitalization. This research used the equal frequency technique to discretize and validate the result with an expert to guarantee the suitability and applicability of the classes for implementation afterwards.

After pre-processing the data, the final dataset was randomly split into two subsets, a training set and a testing set. 70% of the records went to the training set and the remainder 30% was used as the testing set.

3.2.3 Model Building

A data mining algorithm, decision tree C4.5, was used to discover hidden information in the data. By using this algorithm, a length of stay prediction model of dengue patients was built in the form of a decision tree. This algorithm was executed using Weka, an open-source software application for data mining and machine learning [27].

The first step before running the C4.5 algorithm was setting the values of the parameters, known as parameter tuning. In this research, two parameters were considered for constructing the model using Weka: pruning confidence and minimum number of instances, denoted by C and M respectively. The values of the parameters were decided by searching for the combination of parameters that resulted in the best accuracy for the classifier model. Weka's CVParameterSelection feature was employed to perform a grid search on the parameters and generate the best combination.

3.2.4 Model Evaluation

Evaluation of the performance of the model was done by comparing the prediction result to the actual target attribute class. This was done using the testing dataset, a separate subset of the data, which had not been seen and was used to build the predictive model. A confusion or classification matrix was used to show the classification result in detail for each actual and predicted length of stay class combinations. Performance measures analysed in this research were as follows:

1. Accuracy

   Accuracy is the rate of correctly classified records.

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

2. Area under receiver operating characteristic (ROC) curve

   Area under ROC curve is calculated to measure the trade-offs between true the positive rate and the false positive rate shown in the curve. The true positive rate and the false positive rate are defined by:

   $$True\ positive\ rate = \frac{TP}{TP+FN} \quad (2)$$

   $$False\ positive\ rate = \frac{FP}{FP+TN} \quad (3)$$

   TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of true negatives.

After validating the model's effectiveness, a prototype of the prediction model was built to facilitate the user for easier decision-making in the hospital.

4 Results and Discussion

4.1 Length of Stay

The equal frequency technique was used to change the type of the 'Length of stay' attribute from numeric to nominal. The data were classified into 3 bins, which represent short, medium, and long hospital stays for DF/DHF patients. Table 4 shows the result of the equal frequency calculation.

Table 4 Length of stay discretization result.

No	Class	Number of records
1	0-2 days	82
2	3-4 days	182
3	> 4 days	94

The class labels obtained after discretization of the attribute 'Length of stay' were 0-2 days, 3-4 days, and more than 4 days. These classes were meant to indicate the amount of intervention required to reduce or prevent prolonged LOS, where short stays require minimum intervention, medium stays require moderate intervention, and long stays require aggressive intervention [8]. From a medical perspective, these classes are logically acceptable to represent the amount of intervention needed for dengue patients. By discretizing the target attribute this way the classification accuracy will be higher while still resulting in clinically meaningful predictions.

4.2 Classifier Model

The result showed that the combination of parameters that generates the best accuracy for the classifier model is:

1. Pruning confidence factor (C): 35%
2. Minimum number of instances (M): 6

A pruning confidence of 35% means that the algorithm removing branches to avoid an overly large and complex tree is 35% or more confident in doing so. This will prevent over-fitting, which happens when there are branches that are too specific and not representative for generalization [5]. The other parameter represents the minimum number of instances required in each leaf or node for decision tree growing. It means that the tree avoids dividing into nodes with too few supporting cases.

The resulting tree has a size of 25, which indicates low complexity. The resulting model was presented in the form of a decision tree that can be read as a set of if-then rules that can be easily interpreted by nontechnical users. A visualization of the tree is shown in Figure 2.

Four rules were generated for '0-2 days' classification, 5 rules for '3-4 days' classification, and 4 rules for 'more than 4 days' classification. Furthermore, 7 significant attributes out of 21 input attributes that contribute to predicting length of stay of dengue patients were found. These attributes were: systolic and diastolic blood pressure, hematocrit, leucocytes, lymphocytes, monocytes, and comorbidity score.

For each classification rule from the decision tree a different level of support and confidence was obtained when training the data, as shown in Table 5. Support is the percentage of the training data that applies to the condition of the rule (if-statement). Meanwhile, confidence measures the accuracy of the rule. These measures can provide additional consideration for physicians using the model to predict the length of stay of a patient.

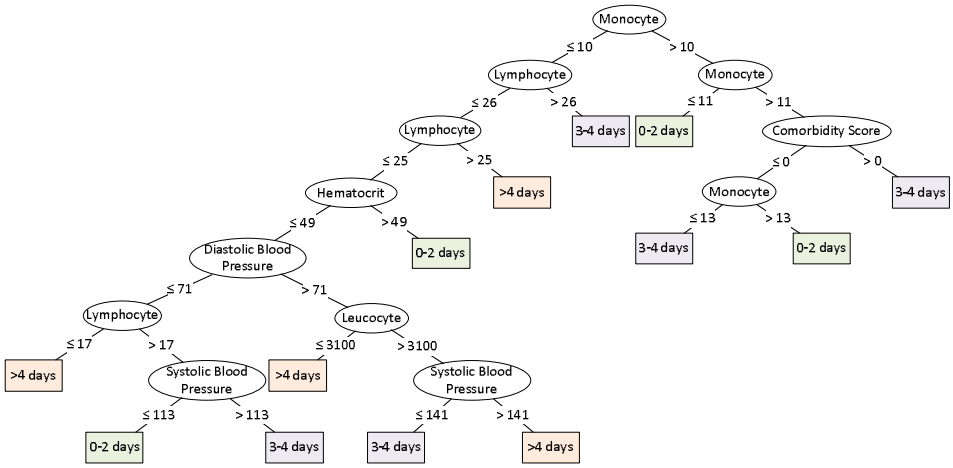


Figure 2 Visualization of the decision tree.

Table 5 Support and confidence of rules.

Rule	Class	Applied If-Then*	Applied If**	Support	Confidence
1	> 4 days	7	2	4%	78%
2	0-2 days	7	1	3%	88%
3	3-4 days	6	2	3%	75%
4	> 4 days	7	0	3%	100%
5	3-4 days	19	4	9%	83%
6	> 4 days	6	1	3%	86%
7	0-2 days	6	2	3%	75%
8	> 4 days	30	2	13%	94%
9	3-4 days	111	21	52%	84%
10	0-2 days	20	4	9%	83%
11	3-4 days	16	6	9%	73%
12	0-2 days	15	5	8%	75%
13	3-4 days	6	2	3%	75%

*Instances that satisfy both hypothesis (if-statement) and conclusion (then-statement) of the rule.
**Instances that failed to satisfy the conclusion of the rule.

From the data in Table 5 it can be seen that the classification rule with the most supporting instances was rule number 9, which is the rule to classify patients in the '3-4 days' class. It shows that the majority of patients in Dr. M. Salamun Hospital, 52% in fact, satisfied the characteristics and had the attribute values of rule number

9, i.e. monocytes ≤ 10 and lymphocytes > 26. Apart from the support value, each rule has considerably high confidence values, which is reassuring for the physicians who use the method.

Information discovered from the model was then evaluated using the testing data set. The frequency of correct and incorrect predictions of instances from the testing dataset is displayed in the confusion matrix shown in Table 6.

Table 6 Confusion matrix.

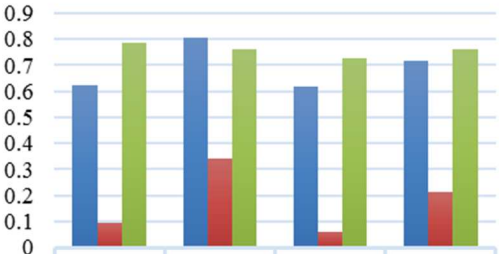
		Prediction		
		0-2 days	**3-4 days**	**> 4 days**
Actual	**0-2 days**	18	9	2
	3-4 days	7	42	3
	> 4 days	0	8	13

It was found that the model correctly classified 71.57% instances, while the remainder 28.4% fell into a wrong class. For comparison with the traditional accuracy measure, area under receiver operating characteristic (ROC) curve was calculated, which gives the trade-off value between the true positive rate and the false positive rate. It shows the probability that a randomly chosen positive instance in the test data is ranked above a randomly chosen negative instance [27]. In this case, a 0.761 ROC area was obtained, as shown in Figure 3. This number implies that the ranking produced by the classifier is not essentially random and is relatively close to the best outcome, i.e. an ROC area of 1.

4.3 Model Implementation

In predicting the length of hospital stay, when a patient is classified into a shorter stay class, a lower priority treatment is expected to be given compared to a patient in a longer stay class. This could be a drawback for the patient and for the hospital as well. On the other hand, when a patient is predicted to have a longer stay than he or she actually needs, the hospital may plan to use resources that turn out to be unneeded, which could lead to waste. Therefore, length of stay predictions require a considerably high accuracy.

Having the right input attributes is critical to reach high classification accuracy [28]. This research focused mostly on the clinical attributes as large amounts of data containing these attributes are available to mine. There may be many influencing factors that could affect length of stay and the selected clinical attributes may only represent a small percentage. Dealing with this issue, a model that can predict length of stay at the time of admission with accuracy higher than 70% is considered to be effective.



	0-2 days	3-4 days	> 4 days	Weighted Average
True Positive Rate	0.621	0.808	0.619	0.716
False Positive Rate	0.096	0.340	0.062	0.213
Area ROC	0.785	0.761	0.727	0.761

Figure 3 Detailed accuracy by class.

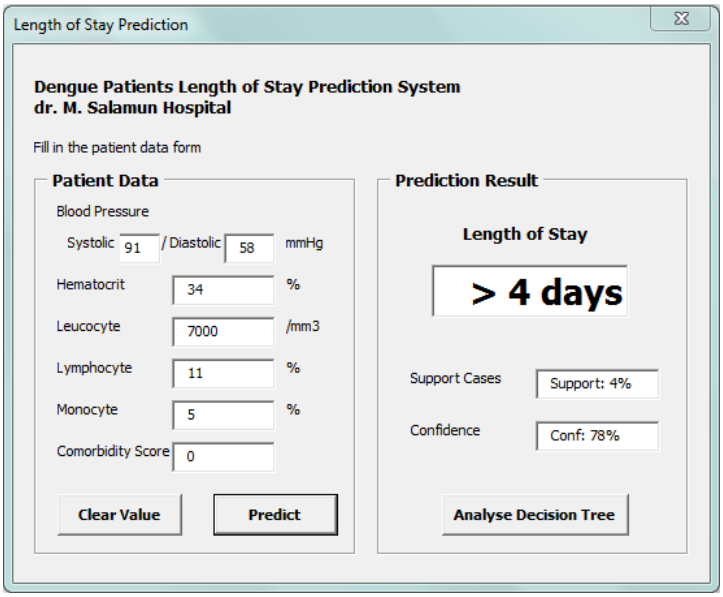


Figure 4 User interface.

Based on the evaluation results, the resulting predictive model was implemented in a system that can be used directly by the hospital. Figure 4 shows the user interface design of the dengue patient length of stay prediction system. This prediction system has been verified and validated according to its objectives and functional and non-functional requirements.

5 Conclusion

This research provides insight into the factors influencing dengue in-patient length of stay in the form of a decision tree that can easily be interpreted by physicians. Attributes that were identified as predictors were systolic and diastolic blood pressure, hematocrit, leucocytes, lymphocytes, monocytes, and comorbidity score, with monocytes as the most significant predictor. A decision tree was built using the C4.5 algorithm, which produced 71.57% accurate results.

An acceptable prototype of the dengue patient length of stay prediction system was developed using the resulting decision tree classification rules. By entering the following laboratory test data of a dengue patient: monocytes, diastolic blood pressure, hematocrit, leucocytes, systolic blood pressure, comorbidity score, and lymphocytes, the developed classification model stored in the system can be used to predict the length of stay of the patient. This system can assist the hospital in making decisions related to clinical and resource management, and therefore can become a part of the hospital's information system. As the data used to construct the model were collected from one local hospital only, the generalization of the results of this study to other hospitals requires further research involving data from a large number of hospitals.

References

[1] Farooqi, W. & Ali, S., *A Critical Study of Selected Classification Algorithms for Dengue Fever and Dengue Haemraghic Fever*, in 11th International Conference on Frontiers of Information Technology, IEEE, pp. 140-145, 2013.

[2] Farooqi, W., Ali, S. & Wahab, A., *Classification of Dengue Fever Using Decision Tree*, VAWKUM Transactions on Computer Sciences, **3**, pp. 15-22, 2014.

[3] Thitiprayoonwongse, D., Suriyaphol, P. & Soonthornphisaj, N., *A Data Mining Framework for Building Dengue Infection Disease Model*, in The 26th Annual Conference of The Japanese Society for Artificial Intelligence, pp. 1-8, 2012.

[4] Thitiprayoonwongse, D., Suriyaphol, P. & Soonthornphisaj, N., *Data Mining of Dengue Infection Using Decision Tree*, in Latest Advances in Information Science and Applications, pp. 154-159, 2012.

[5] Kumar, M.N., *Alternating Decision Trees for Early Diagnosis*, Cornell University, http://arxiv.org/abs/1305.7331, 2013 (Retrieved 23 April 2016).

[6] Shakil, K.A., Anis, S. & Alam, M., *Dengue Disease Prediction Using WEKA Data Mining Tool,* Cornell University, http://arxiv.org/abs/1502.05167, 2015 (Retrieved 22 April 2016).

[7] Tanner, L., Schreiber, M., Low, J.G., Ong, A., Tolfvenstam, T., Lai, Y. L., Ng, L.C., Leo, Y.S., Puong, L.T., Vasudevan, S.G., Simmons, C.P., Hibberd,

M.L. & Ooi, E.E., *Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in Early Phase of Illness*, PLOS Neglected Tropical Diseases, **2**, pp. 1-9, 2008.

[8] Azari, A., Janeja, V.P. & Mohseni, A., *Healthcare Data Mining: Predicting Hospital Length of Stay (PHLOS)*, International Journal of Knowledge Discovery in Bioinformatics, **3**(3), pp. 44-66, 2012.

[9] Blais, M.A., Matthews, J., Lipkis-Orlando, R., Lechner, E., Jacobo, M., Lincoln, R., Gulliver, C., Herman, J.B. & Goodman, A.F., *Predicting Length of Stay on an Acute Care Medical Psychiatric In-patient Service*, Administration and Policy in Mental Health, **31**, pp. 15-29, 2003.

[10] Combes, C., Kadri, F. & Chaabane, S., *Predicting Hospital Length of Stay Using Regression Models: Application to Emergency Department*, in 10ème Conférence Francophone de Modélisation, Optimisation et Simulation, https://hal.archives-ouvertes.fr/hal-01081557, 2014 (Retrieved 18 February 2016).

[11] Hachesu, P.R., Ahmadi, M., Alizadeh, S. & Sadoughi, F., *Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients*, Healthcare Informatics Research, **19**, pp. 121-129, 2013.

[12] Lella, L., di Giorgio, A. & Dragoni, A.F., *Length of Stay Prediction and Analysis through a Growing Neural Gas Model*, in 4th International Workshop on Artificial Intelligence and Assistive Medicine, pp. 11-21, 2015.

[13] Liu, P., El-Darzi, E., Vasilakis, C., Chountas, P. & Huang, W., *Comparative Analysis of Data Mining Algorithms for Predicting In-patient Length of Stay*, in Pacific Asia Conference on Information Systems, pp. 1087-1097, 2004.

[14] Tanuja, S., Acharya, U.D. & Shailesh, K., *Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay*, Journal of Pharmaceutical and Biomedical Sciences, **7**, pp. 1-4, 2011.

[15] Yang, C.S., Wei, C.P., Yuan, C.C. & Schoung, J.Y., *Predicting The Length of Hospital Stay of Burn Patients: Comparison of Prediction Accuracy among Different Clinical Stages*, Decision Support Systems, **50**, pp. 325-335, 2010.

[16] Tomar, D. & Agarwal, S., *A Survey on Data Mining Approaches for Healthcare,* International Journal of Bio-Science and Bio-Technology, **5**, pp. 241-266, 2013.

[17] Larose, D.T., *Discovering Knowledge in Data: An Introduction to Data Mining*, New Jersey: John Wiley & Sons, Inc., 2005.

[18] Lim, T.S., Loh, W.Y. & Shih, Y.S., *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms*, Machine Learning, **40**, pp. 203-229, 2000.

[19] Carrasco, L.R., Leo, Y.S., Cook, A.R., Lee, V.J., Thein, T.L., Go, C.J. & Lye, D.C., *Predictive Tools for Severe Dengue Conforming to World Health*

Organization 2009 Criteria, PLOS Neglected Tropical Diseases, **8**, pp. 1-9, 2014.

[20] Xiao, J., Douglas, D., Lee, A. H. & Vemuri, S. R., *A Delphi Evaluation of The Factors Influencing Length of Stay in Australian Hospitals*, International Journal of Health Planning and Management, **12**, pp. 207-218, 1997.

[21] Lee, V.J., Lye, D.C., Sun, Y., Fernandez, G., Ong, A. & Leo, Y.S., *Predictive Value of Simple Clinical and Laboratory Variables for Dengue Hemorrhagic Fever in Adults*, Journal of Clinical Virology, **42**, pp. 34-39, 2008.

[22] Thein, T.L., Leo, Y.S., Fisher, D.A., Low, J.G., Oh, H.M., Gan, V.C., Wong, J.G.X. & Lye, D.C., *Risk Factors for Fatality among Confirmed Adult Dengue In-patients in Singapore: A Matched Case-Control Study*, PLOS ONE, **8**, pp. 1-6, 2013.

[23] Aroor, A.R., Saya, R.P., Sharma, A., Venkatesh, A. & Alva, R., *Clinical Manifestations and Predictors of Thrombocytopenia in Hospitalized Adults with Dengue Fever*, North American Journal of Medical Sciences, **7**, pp. 547-552, 2015.

[24] Ho, T.S., Wang, S.M., Lin, Y.S. & Liu, C.C., *Clinical and Laboratory Predictive Markers for Acute Dengue Infection*, Journal of Biomedical Science, **20**, pp. 1-8, 2013.

[25] Kalayanarooj, S., Vaughn, D., Nimmannitya, S., Green, S., Suntayakorn, S., Kunentrasai, N., Viramitrachai, W., Ratanachu-eke, S., Kiatpolpoj, S., Innis, B.L., Rothman, A.L., Nisalak, A. & Ennis, F.A., *Early Clinical and Laboratory Indicators of Acute Dengue Illness*, The Journal of Infectious Diseases, **176**, pp. 313-321, 1997.

[26] Premaratna, R., Pathmeswaran, A., Amarasekara, N., Motha, M., Perera, K. & Silva, H.D., *A Clinical Guide for Early Detection of Dengue Fever and Timing of Investigations to Detect Patients Likely to Develop Complications*, Transactions of the Royal Society of Tropical Medicine and Hygiene, **103**, pp. 127-131, 2009.

[27] Witten, I.H., Frank, E. & Hall, M.A., *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Burlington, MA: Morgan Kaufmann, 2011.

[28] Han, J., Kamber, M. & Pei, J., *Data Mining Concepts and Techniques*, 3rd ed., Waltham, MA: Morgan Kaufmann, 2012.