Dengue fever prediction modelling using data mining techniques

Wipawan Buathong and
Pita Jarupunphol*

Department of Digital Technology,
Phuket Rajabhat University,
Phuket, Thailand
Email: w.buathong@pkru.ac.th
Email: p.jarupunphol@pkru.ac.th
*Corresponding author

Abstract: This research experiments on several combinations of feature selection and classifier to obtain the most efficient classification model for predicting dengue fever. The features of relationship patterns for predicting dengue fever were investigated. In order to obtain the most effective classification model, several feature selection techniques were ranked and experimented with well-recognised classifiers. The measurement results of different models were illustrated and compared. The most efficient model is the neural network with three layers. Each layer contains 100 nodes with ReLu activation function. Five features were classified using information gain with 64.9% accuracy, 71.8% F-measure, 65.7% precision, and 79.0% recall. Other competitive machine learning models with slightly similar efficiency are: (1) the combined Naive Bayes and information gain; (2) the combined neural network and ReliefF; (3) the combined Naive Bayes and FCBF. SVM, on the other hand, is considered as the least efficient model when experimented with selected feature selection techniques.

Keywords: dengue fever; data mining; classification; feature selection; ranking.

Reference to this paper should be made as follows: Buathong, W. and Jarupunphol, P. (2021) 'Dengue fever prediction modelling using data mining techniques', *Int. J. Data Mining and Bioinformatics*, Vol. 25, Nos. 1/2, pp.103–127.

Biographical notes: Wipawan Buathong received her PhD in Information Technology from King Mongkut's University of Technology North Bangkok, Thailand. She is an Assistant Professor at the Department of Digital Technology, Phuket Rajabhat University. Her research interests include data mining, dimensionality reduction, and data classification.

Pita Jarupunphol received his PhD in Computer Science from the University of Auckland, New Zealand. He is an Assistant Professor at the Department of Digital Technology, Phuket Rajabhat University. His research interests cover information security, human-computer interaction, and data science.

1 Introduction

The dengue incidence has increased significantly in recent decades. About half of the world's population is now at risk of being infected by dengue fever (World Health Organisation, 2020). In several cases, severe dengue fever can be associated with different symptoms, e.g., bleeding, organ impairment and/or plasma leakage. When the dengue becomes severe, there is a higher risk of death if it cannot be dealt with appropriately. Currently, there is no specific treatment for dengue fever. Dengue fever must be coped with by medical professionals in hospitals. Severe dengue has become a leading cause of hospitalisation and death among children and adults in Thailand (Ministry of Public Health, 2020). Dengue prevention and control depend upon effective vector control measures and sustained community involvement. An early prediction of disease progression is associated with severe dengue. An access to proper medical care is capable of reducing fatality rates of severe dengue to below 1%. According to the dengue fever report conducted by the Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health (Thailand) (2020), Thai citizens are infected with dengue fever each year. Especially in 2013, there are 62 people died from dengue fever out of a total of 66,299 people infected with dengue fever. As such, rapid preparation and response to dengue fever must be taken into account. Table 1 represents cases of dengue fever between 2012 and 2016 from the Bureau of Epidemiology report (Thailand).

Table 1 Dengue fever cases between 2012 and 2016

Details/Year	*2012*	*2013*	*2014*	*2015*	*2016*
Patients (income)	20,660	66,299	12,918	30,457	20,395
Deaths (persons)	18	62	12	17	17
Rate per patient (hundred thousand)	32.15	103.47	19.89	46.77	31.17
Rate (%)	0.09	0.09	0.09	0.06	0.08

Source: Ministry of Public Health (2020)

Nowadays, data mining has played a significant role in providing essential data perspectives as it involves several data processes from searching to analysing the data. The analysed data can be converted into meaningful information for the public and private sectors to identify data relationships, predict and make a more accurate decision from a large volume of data. In particular, dimensionality reduction or feature reduction is the process of reducing the number of random variables under consideration by obtaining a set of key variables (Agarwal, 2014; Witten et al., 2016). In other words, dimensionality reduction is a data dimension downsizing technique that reduces the data size by eliminating unnecessary data features to increase the accuracy of classification. Feature selection and feature extraction are dimensionality reduction techniques that have been applied in data mining. As with feature selection, irrelevant features can be filtered out from the dataset. In contrast, feature extraction creates a new set of features but stills contain useful information.

Recently, several data mining techniques have been proposed to classify data features in various kinds of diseases (e.g., leukaemia, breast cancer, and hepatitis). The most weighted data features can be utilised as representative features to predict such diseases. In addition to non-communicable diseases, dengue fever is a virus infectious disease caused by dengue illness with female *Aedes aegypti* mosquitoes being a primary carrier (Dasgupta et al., 2019). The bites of an infected female Aedes mosquito, which gets the virus while feeding on the infected person's blood, transmit the virus to others. Dengue transmission is effective in tropical climates with higher vapour pressure and rainfall rates. The severe dengue fever categories comprise dengue fever (DF), dengue haemorrhagic fever (DHF), and dengue shock syndrome (DSS). The severity of these dengue fevers can increase in the event of a loss of life if diagnosis and treatment cannot be made on time.

While either feature selection or classifier can be used to classify data features and reduce data dimensions, different combination techniques have been proposed by research scholars, who search for the most efficient data classification model for obtaining the best features (Arabshahi and Fazlollahtabar, 2018; Manek et al., 2017; Priya and Ranjith Kumar, 2015; Renuka Devi et al., 2016). Likewise, an efficient model is also required in this research as there are many data features that can be used to identify dengue fever. Therefore, it is questionable that:

1 how many features are sufficient to predict dengue fever with high accuracy

2 what combination techniques can be considered as the most efficient dengue fever prediction model.

In this case, the objective of this research is to measure the efficiency of models based on a combination of feature selection techniques and classifiers for facilitating dengue fever prediction. The article starts from identifying data features that can be utilised to predict dengue fever. The parameters required for the linear kernel will also be determined.

In this research, several well-recognised feature selection and classification techniques in the literature were chosen for the experiment. Different classifier algorithms will be applied with feature selection techniques to enhance the classification efficiencies and obtain the most effective model. In this research, feature selection techniques, including GINI, gain ratio, info gain, ReliefF, chi-square and FCBF, will be experimented with decision tree, Naive Bayes, neural network, and K-nearest neighbours. Four criteria, including accuracy, precision, recall, and f-measure are utilised to measure and rank the effectiveness of each model. The experimental results will reveal and rank the efficiency of each model for dengue fever prediction when data features are downsized to different dimensions.

2 Literature reviews

This section commences from reviewing the literature related to the research, including feature selection techniques and classifiers. After that, applications of data mining in dengue fever will also be discussed.

2.1 Feature selection

Feature selection is generally used to define the tools and techniques for reducing inputs to a controllable size for processing and analysis. Feature selection is also used for machine learning and other data mining applications (Colaco, 2016). As feature selection is one of data dimensionality reduction techniques, obtaining the most useful feature selection has become a challenging topic in machine learning. There are several efforts on dimensionality reduction using feature selection techniques for a large amount of data attributes (Colaco, 2016). As mentioned previously, feature selection is frequently used to remove inappropriate and noisy features to retrieve relevant features for data classification. In several cases, however, the classification process can be time-consuming (Renuka Devi et al., 2015). The data with many features will affect the classification performance since the methods used in learning to create classifiers do not assure the data with a large number of features. Feature selection and data classification are critical functions of data mining specifying data attributes in target classes. According to Colaco (2016), the number of data generated worldwide is increasing in multiple fields such as social media, bioinformatics and healthcare. These data contain redundant, irrelevant or noisy data which causes high dimensionality. Feature selection techniques experimented in this research are discussed below.

2.1.1 Gini ratio

Gini ratio is widely recognised as Gini index or Gini coefficient, which is a statistical distribution measurement introduced by Corrado Gini (Giorgi, 2011). This technique is often applied for measuring income distribution among a population due to its simplicity. The measurement value is based on the degree between 0 and 1. While 0 implies perfect equality, perfect inequality is indicated by 1. In this case, a higher Gini index represents greater inequality and vice versa. Nevertheless, the Gini index popularity has nowadays declined due to its bias and limitations. The applications of Gini ratio have been evidenced in different research works (Arabshahi and Fazlollahtabar, 2018; Manek et al., 2017; Pandiyarajan and Thangairulappan, 2019). The Gini ratio equation is calculated in (1) where $p(t_i)$ is the proportion of data contained in class i with respect to all data.

$$Gini(t_i) = 1 - \sum_{i=1}^{n} [p(t_i)]^2 \tag{1}$$

2.1.2 Information gain

Information gain (IG) is another well-respected feature selection technique used by decision tree algorithms (Furnkranz, 2010). This technique measures how much ‘information’ a feature provides about the class. An attribute with highest information gain will be tested and separated at an initial stage. The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is based on finding the attribute that returns the highest information gain. Consequently,

information gain is considered as biased towards choosing attributes with a large number of values as root nodes (Azhagusundari and Thanamani, 2013; Muller and Muller, 2012; Raileanu and Stoffel, 2004). Information gain related formulas are illustrated in (2) and (3); where

1 $-\sum_t$ is the sum of the probabilities of j occurring in the t class

2 $Entropy(p)$ is the entropy of the root

3 $\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$ is the entropy of each subnode.

$$Entropy(t) = -\sum_t p(j \mid t) \log_2 p(j \mid t) \tag{2}$$

$$Gain = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right) \tag{3}$$

2.1.3 Gain ratio

Gain ratio is a revised version of the information gain, which was proposed to address bias and limitations associated with the information gain (Buathong, 2012; Harris, 2002; Chandra and Saxen, 2004). Gain ration is different from the information gain when considering instance distribution of entropy. Gain ratio does not ignore intrinsic information of a split when selecting an attribute, i.e., number and size of branches. A generic overview of Gain ratio can be illustrated in (4); where

1 T represents the set of the training set

2 x represents the attribute chosen as the classifier

3 $info(T)$ is a function that specifies the required amount of data so that the desired characteristics can be recognised.

$$Gain(x) = info(T) - info_x(T) \tag{4}$$

2.1.4 Chi-square (χ^2)

A chi-square (χ^2) algorithm is widely used in statistics for testing the independence of two categorical variables. In data mining, this algorithm measures how expected features deviate each other (Agarwal, 2014; Witten et al., 2016). The chi-square can be used to determine if a relationship between two features in a sample (test) is capable of reflecting of a real association between these features in the population from the probability (p-value). Besides, the algorithm can also determine if there is a difference between the two features for comparison purposes. How χ^2 is calculated is represented in (5), where

1 A_{ij} is the actual frequency of a sample with i and class j

2 E_{ij} is the expected frequency of a sample with i and class j

3 m is the number of characteristic values

4 n is the number of classes.

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{5}$$

2.1.5 FCBF

A fast correlation-based filter (FCBF) is based on the notion of 'predominant correlation' in which features with high correlation with the target variable but with minor correlation with other variables are chosen (Balakrishnan and Narayanaswamy, 2009; Kavitha et al., 2017; Senliol et al., 2008; Zeng et al., 2010). Given that there exists a predominant feature 'X', no other feature is more correlated to 'X' than 'X'. Then, the predominant correlation feature arises after the features become more correlated with 'X' than with the class have been tested. The principle of characterisation and redundancy analysis of the FCBF is calculated using a coefficient rating 'symmetrical uncertainty' (SU), which is used to determine the feature 'conditional uncertainty' level and the criteria of X over Y. If SU value is 1, one feature can be a representative and predictive feature of other features. On the other hand, the value 0 implies that two features are not correlated. However, the coefficient must be greater than the specified criterion to determine the number. Redundant features will be eliminated, leaving only the dominant feature. The FCBF equation is represented in (6).

$$SU(X,Y) = 2\left[\frac{H(X) - H(X \mid Y)}{H(X) + H(Y)}\right] \tag{6}$$

2.1.6 ReliefF

The ReliefF algorithm (Kira and Rendell, 1992), which was developed from Relief, is another feature selection algorithm that calculates the weight from data with consideration similar to the random data. ReliefF is a method of selecting a feature that can be done with more than two data classes. In this case, it is available for all data types and resistant to incorrect and incomplete data (Robnik-Ŝikonja and Kononenko, 2003). Nevertheless, there is a disadvantage of ReliefF as it can be performed with only nominal and numeric data. The ReliefF algorithm estimates the difference in characteristics with nearby samples (K-nearest neighbours or K-NN) in the same or different classes between 0 and 1 of K. To find K values, the algorithm commences from searching within the same class. While the value 0 indicates that the characteristics are similar, the value 1 implies significant differences of the characteristics. There have been several data mining applications based on this algorithm (Huang et al., 2007; Symeonidis et al., 2006). The ReliefF equation is illustrated in (7), where

1 $W[A]$ represents all feature weights

2 $diff(A,R,H)$ is the difference of all features (A), a random instance (R), and nearest hits (H)

3 $P(C)$ is the class probability

4 M is nearest miss for each class.

$$W[A] := W[A] - \frac{diff(A,R,H)}{m} + \sum_{\neq class(R)} \left(P(C) * \frac{diff(A,R,M_c)}{m} \right) \tag{7}$$

2.2 Data mining

Besides feature selection techniques, there are prominent classifiers for data mining appropriate for particular data types and features.

2.2.1 Support vector machine

Support vector machine (SVM) is a widely recognised technique generally used to create machine learning models for data classification due to its high accuracy. SVM was discovered to be more efficient than other data classification techniques. The efficiency of SVM is also supported by other research scholars (Buathong, 2012; Chen and Lin, 2006; Lu and Wang, 2004; Weston et al., 2001; Zou and Jin, 2018), who applied different techniques and parameters to optimise the data classification performance of SVM. SVM utilises the maximum margin of the decision hyperplane to determine the decision plane and use that plane to classify the data. Since the decision plane may contain many planes, selecting only the best decision plane must be taken into account. Scattered points on the plane and the highest distance or margin of each data group must be considered. The decision plane used for data classification can be either linear or nonlinear. In many cases, if the sample data used to create a machine learning model with SVM are incapable of identifying the right decision plan, methods of converting data to higher dimensions are preferable for many researchers. Although SVM works effectively with datasets, the implementation of all of the data in this SVM may be delayed, and the dimensions should be initially reduced to make data mining more efficient.

2.2.2 Decision tree

Decision tree is a widely used classifier due to its characteristics represented in the form of an organisation chart, which is understandable for many researchers and practitioners. Decision tree establishes classification or regression models in the form of a tree structure. This classifier breaks down a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The decision tree comprises node, branch, and leaf. While each node represents the attribute, each branch shows the testing result and the leaf node represents the defined class. The root node, which is the

topmost decision node in a tree, corresponds to the best predictor (Arabshahi and Fazlollahtabar, 2018; Chandra and Saxen, 2004). There are advantages and disadvantages of using a decision tree. For instance, decision trees can handle both categorical and numerical data. All potential decisions are taken into account and can be traced to a conclusion. Incomplete decision nodes that must be further analysed can also be identified from the decision tree. However, the complete decision tree structure can be affected by a small change in the data causing inaccuracy in predictive performance of the decision tree. Besides, the classification performance declines as the number of attributes increases. Kumar and Suman (2011) classified data types and symbols into two classes using the decision tree algorithm.

2.2.3 Naive Bayes

Naive Bayes is a classifier used to create a probability principle that classifies data using Bayesian rules. Naive Bayes is based on the assumption that all features are independent of each other with defined conditions Bayesian learning is an easy way to classify data. The experimental results are comparable to those from more sophisticated algorithms such as C4.5 (Dougherty et al., 1995), which is an algorithm of decision tree that classifies data into branches and establish rules for the data.

2.2.4 Neural network

The neural network or artificial neural networks (ANN) is a classifier based on an artificial intelligence (AI) technology for calculating functions from data groups (Majumder, 2015; Ozyildirim and Avci, 2013). It is a method that provides machine learning from a prototype. The machine is trained to understand how to address generic problems. The neural network is based on the transmission of input-output nodes and processes distributed in different layers, including input, output, and hidden layers.

2.2.5 K-nearest neighbours

K-NN is a neighbour's data classification technique equal to K, which is another competitive technique for data classification methods. K-NN algorithm is considered as unsupervised learning without using training data to simulate the data model but uses the experimental data as a model. To implement the K-NN algorithm, a positive integer number must be specified for K, which will indicate the number of target classes that must be utilised to predict new cases, e.g., 1-NN, 2-NN, 3-NN,…, K-NN. To illustrate this, 3-NN implies the algorithm will search for three cases and determine cases similar to the new cases. This data classification technique is straightforward to implement. The data of interest will be compared with others if there is much resemblance. If the target data is closest to the criteria, the system will consider the nearest data as the answer.

2.3 Data mining applications in dengue fever

This section discusses some research works on data mining applications related to dengue fever classification.

2.3.1 Delen et al.

Delen et al. (2005) detailed survival data for breast cancer patients from the experimental data conducted by the American Cancer Society. The data contains 202,932 patients and 17 variables classified into survivors and non-survivors. Three data classification techniques were applied, including neural network, decision tree, and degradation logic. Furthermore, there were two performance measurement approaches for data classifications. While the first approach measures accuracy in terms of overall, sensitivity, and specificity respectively, the results indicated that the artificial neural network algorithm gave the values at 91.2% 94.37% and 87.48%. Besides, the decision tree algorithm could provide the values of 93.62% 96.02% and 90.66%. For the degradation logic, the performance reached 89.20%, 90.71%, and 87.86%. However, the second measurement approach is based on the logical degeneration measurement algorithm using the k-fold cross-validation. It was found that the decision tree algorithm provided the highest accuracy at 93.62%.

2.3.2 Thitiprayoonwongse et al.

Thitiprayoonwongse et al. (2011) classified dengue data into four classes using decision trees, including dengue fever (DF), dengue haemorrhagic fever (DHF) (I), DHF (II) and DHF (III). The data was taken from Songklanagarind Hospital in Songkhla and Srinagarind Hospital in Khon Kaen. Accuracy was used as the primary evaluation criterion for efficiency measurement. Two experiments were conducted to ensure the result validity. The experimental results revealed high accuracy for the average at 96.50% and the second experiment with classification rules at 96.50%.

2.3.3 Tanner et al.

Tanner et al. (2008) investigated data mining for dengue fever to predict potential outcomes of the disease. Initially, the data was divided into dengue fever and non-dengue fever. After that, the dengue fever data was further classified into two classes, including dengue fever and severe dengue fever. Patient information was taken from Dong Thap Hospital in Singapore. The Decision Tree algorithm and validation methods were based on the 10-fold cross-validation. The data classification performance results are at 71.20%, 90.10%, and 84.70% for sensitivity, specificity, and F-measure, respectively.

2.3.4 Husam et al.

Husam et al. (2017) investigated three feature selections, including particle swarm optimisation (PSO), genetic algorithm (GA) and rank search (RS) to identify features with a higher accurate prediction of dengue outbreaks. Based on the selected features, three predictive modelling techniques (J48, DTNB and Naive Bayes) were applied for

dengue outbreak detection. The trial dataset was obtained from the Public Health Department in Seremban, Malaysia. The outcomes affirmed that the predictive accuracy was ventured forward by utilising feature selection process before the predictive modelling process. The examination likewise indicated a set of features to represent dengue outbreak detection for Malaysian health agencies. This work encourages capturing the relations and patterns inside the data and enhancing the predictive accuracy of dengue outbreaks. Among the three selected feature selection algorithms, PSO reached the best accuracy uncovering a new set of features representing dengue data, i.e., year, a cumulative week from 2003 to 2010, number of dengue fever for the current week, minimum temperature value, average temperature value, rainfall and race. The authors argued that this work determines the most related attributes of dengue data before modelling the issue by applying feature selection process, unlike most previous works.

2.3.5 Dasgupta et al.

Dasgupta et al. (2019) experimented on the classification algorithms to distinguish the key features liable for spreading the dengue. To find the features identified with the spread of dengue disease, the authors applied three mainstream machine learning algorithms, including random forest classifier (RFC), decision tree classifier (DTC) and linear support vector machine (LSVM). In addition, predictive mean matching and percentage split are also applied for data preprocessing and resampling, respectively. In this work, forward selection (FS) and backward elimination (BE) are two feature selections involved to help the algorithm run more effectively and improve the classification accuracy.

2.3.6 Pandiyarajan and Thangairulappa

Pandiyarajan and Thangairulappa (2018) stated that different diagnosing methods like ELISA, Platelia, haemocytometer, RT-PCR, decision tree algorithms and support vector machine algorithms are used to diagnose the dengue infection using the detection of antibodies IgG and IgM, but the recognition of the IgM is not possible between thirty and ninety days of dengue infection. These methods were incapable of identifying the correct result and require a certain volume of the blood inappropriate to be applied in children. In this case, Pandiyarajan and Thangairulappa (2019) proposed a classification method of dengue infection based on informative and most significant genes in the gene expression of dengue patients. The proposed method was based on a combination of Gini-index and information gain for feature selection and rule extraction from the neural network for classifying dengue serotypes. Several symptoms, climate risk factors, patients' records and gene sequence of the patients were used to diagnose the dengue in later stages. The authors asserted that serotypes based on the function of the protein can be easily classified for the biologist if the structure of the protein is recognised. The proposed method extracted rules and identified the cause of amino acids for the dengue. The experimental results revealed that the proposed method accuracy is 96% for classifying the dengue serotypes.

2.3.7 Renuka et al.

According to Renuka Devi et al. (2015), the genetic algorithm (GA) applications in data mining techniques are currently evolving to fulfil the efficiency of traditional classification techniques. Renuka Devi et al. (2015) proposed alternative classification techniques based on GA to obtain the most optimal and relevant features. The same authors, Renuka Devi et al. (2016) also further experimented on the same dataset using particle swarm optimisation (PSO), which is an evolutionary algorithm similar to GA, to model a novel classification system. PSO utilises a population of individuals prevailing within a multi-dimensional space. These proposed methods were applied to the dengue dataset obtained from the Gene Expression Omnibus (GEO) repository of the National Center for Biotechnology Information (NCBI). The authors claimed that optimal features could be obtained and the results illustrated the accuracy and validity of the proposed methods. The dataset contains 18 attributes for 1,275 patients.

2.3.8 Shaukat et al.

Shaukat et al. (2015) experimented on five data mining techniques for classifying and predicting dengue fever from District Headquarter (DHQ) Hospital Jhelum in Pakistan, including Naive Bayes (NB), regression tree representative (REPtree), random tree (RT), J48 and sequential minimal optimisation (SMO). Weka, developed by the University of Waikato, New Zealand, was utilised as a data mining tool for the experiment. Selected attributes for the experiment include fever, bleeding, myalgia, flu, fatigue and results. TP rate, ROC rate, error rate, and accuracy are evaluation criteria for measuring the classifiers. Based on the experimental results, Naive Bayes is summarised by the authors as the most effective classifier in comparison with others in terms of both TP rate and accuracy at 92%.

3 Methodology

As has already been discussed, various feature selection and classification approaches have been applied in dengue fever prediction area. While some researchers focus on measuring the performance of data mining classifiers, some concentrate on a combination of feature selections and classifiers. As long as dengue fever is still considered as one of the most serious health illnesses, an effective prediction model will always be required. In this research, experimental models will be based on a combination of well-recognised feature selection and classification techniques and compared according to evaluation criteria. The methodology in this research comprises several steps to achieve the objectives, including data collection, data preparation, data integration, data cleansing, data transformation, data visualisation, and data mining or dimensionality reduction.

3.1 Data collection

The data was collected in collaboration with Vachira Hospital in Phuket in order to obtain data attributes for dengue fever classification. The main data sources were medical records and social medicine offices. The data was filled out by medical records officers

via the electronic system designed and prepared by the researcher. The hospital staff were also informed that the data will be used for research only and no sensitive data will be used and revealed to the public due to the Thailand National Data Protection Act enforcement. There are 1,000 patients used as sampling units and three data classes derived from the dengue patient data, including dengue fever (DF), dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS). The dataset contains 19 attributes, including six number attributes, 12 quality attributes, and one class attribute, for dengue fever identification. For example, *Platelet1*, *Platelet2*, and *Platelet3* record the number of patient's platelets daily measured at the first, second, and third time respectively. Attributes *White_blood_cell_min* and *White_blood_cell_max* store the patient's daily minimum and maximum white blood cells. There are different types of antibody attribute. For instance, *Dengue IgM* contains the result of an antibody that the body produces within 3–4 days since the first symptom of a high fever. *Dengue IgG* is an attribute that records an antibody made up of dengue virus infection. This type of immunity is a long lasting homotypic immunity that may be produced approximately 2 weeks within the first stage of a high fever or 1–2 days since the first symptom of a high fever. Besides, *Dengue NS1Ag* contains the protein NS1A that can stimulate the immune system to produce antibody, which can be found in the plasma of dengue virus infected patients.

Other dengue fever symptoms, including headache, retro-orbital pain, joint pain, nausea, pulse, rash, leakage, shock, bleeding, and temperature, are also used to predict dengue fever. In particular, dengue fever diagnostic results (Dx) is a class attribute that categorises dengue fever into:

1 dengue fever (DF)

2 dengue haemorrhagic fever (DHF)

3 dengue shock syndrome (DSS).

Table 2 represents data features for dengue fever prediction.

Table 2 Primary 19 features for predicting dengue fever

No.	*Attributes*	*Descriptions*
1	Platelet1	The number of platelets at the first time
2	Platelet2	The number of platelets at a second time
3	Platelet3	The number of platelets at a third time
4	White_blood_cell_min	The daily minimum white blood cell
5	White_blood_cell_max	The daily maximum white blood cell
6	Dengue IgM	Positive/negative
7	Dengue IgG	Positive/negative
8	Dengue NS1Ag	Positive/negative
9	Headache	Yes/No
10	Retro-orbit pain	Yes/No
11	Myalgia arthralgia/joint pain	Yes/No
12	Nausea/vomiting	Yes/No

Table 2 Primary 19 features for predicting dengue fever (continued)

No.	*Attributes*	*Descriptions*
13	Pulse	Yes/No
14	Temperature	The patient's temperature (Celsius)
15	Rash	Yes/No
16	Leakage	Yes/No
17	Shock	Yes/No
18	Bleeding	Yes/No
19	Dx (dengue fever diagnostic results)	DF/DHF/DSS

3.2 Data preparation

After the data has been collected in accordance with the national acts, data preparation is a crucial step involving several data processes, i.e., data integration, data cleansing and data transformation, data visualisation, and dimensionality reduction. There are certain tools and applications used in data preparation. For example, Microsoft Excel was used in data cleansing process to handle the errors and organise the data in the same format. While the LIBSVM program was used to identify data features, MATLAB and Weka were primarily used for data mining.

The proportion of data for training and testing are based on the mean of ten iterative trials for the reliability of the model performance. The ten folds cross-validation method was used by dividing the dataset into the K portion evenly. The K-1 data portion was utilised to construct a machine learning model while the other data portion was reserved for testing the validity. This process was repeated until all data portions had been applied to test the machine learning model. The data classification accuracy of each cycle was then combined and averaged to reflect the learning efficiency following four measurement criteria, including accuracy, precision, recall, and F-measure.

3.2.1 Data integration

Data integration is an essential process of merging data from various sources into a unified view after the data has been prepared. The data integration criteria were considered to find the data that is more or less unusually high (outlier) and address missing values. In case the attribute value is quantitative that can be represented by the attribute average value or in the event that the important attributes in determining the data classification are missing. The patient's data will be excluded from the dataset.

3.2.2 Data cleansing and transformation

After the collected data had been merged, data cleansing was considered as it could help eliminate purposeless data caused by recording errors (noise). This step is required before the data selection covering the cooperation of doctors and staff of medical records and social medicine. When selected attributes are irrelevant to the research, reducing the number of factors must be conducted to increase the reliability of the data mining results.

Moreover, data transformation is another essential step to find the representative data after the data have been cleaned. As the data have been stored for a period of time, the patient data records might be different and duplicated. For instance, there might be more than one record for one patient. The data also comprise the range of independent variables or features that must be scaled or normalised. In this experiment, the data were limited to only a single number or text to represent the data of each patient. The data were normalised using min-max technique to adjust the data in the same range for higher accuracy of data classification. After the patient data have been through the cleansing and transformation, a multi-class classification is performed as features must be classified into one of three or more classes. In this research, there are 800 patients classified into 344 DF, 452 DHF, and 4 DSS.

3.2.3 Data visualisation

After the collected data have been through several data processes, the data were further analysed to identify which data characteristics could be used as data classification criteria. In this case, data visualisation is another required step since the data must be represented in an understandable manner. The data relationships in this research were graphically represented in a form of image. The representative data can be observed if there are confusions in DF and DHF.

3.3 Dimensionality reduction

Dimensionality reduction is the last data preparation process aiming to reduce the original data size but preserve essential data characteristics and the result accuracy. At this stage, all dengue data features were downsized and experimented with chosen classifiers to obtain the efficiency of each combination model. As this research is aimed to experiment on a combination of feature selection techniques and classifiers, feature selection techniques, including Gini ratio, gain ratio, information gain, ReliefF, chi-square and FCBF, were experimented with selected classifier algorithms, i.e., SVM, decision tree, Naive Bayes, neural network, and K-nearest neighbours. The data that have been cleansed and transformed were further experimented to obtain training data. Figure 1 summarises all experimental processes in data preparation in addition to other methodological steps.

4 Results and discussion

The analysis of the sample features derived from the actual patient records reveals that there was a loss of data at 13.3%. The average/most frequent method was applied to solve the lost data and calculate the average of numerical data features. The character data feature was based on the method used to determine the highest frequency. All these processes are to ensure that the data can be used for further analysis. To find connections to the features for predicting dengue fever, the researchers ranked features from the sample data using information gain, gain ratio, Gini index, chi-square (χ^2), ReliefF and FCBF. In this case, the top ten weighted features were ranked in descending order according to the feature selection methods.

Figure 1 Research methodology (see online version for colours)

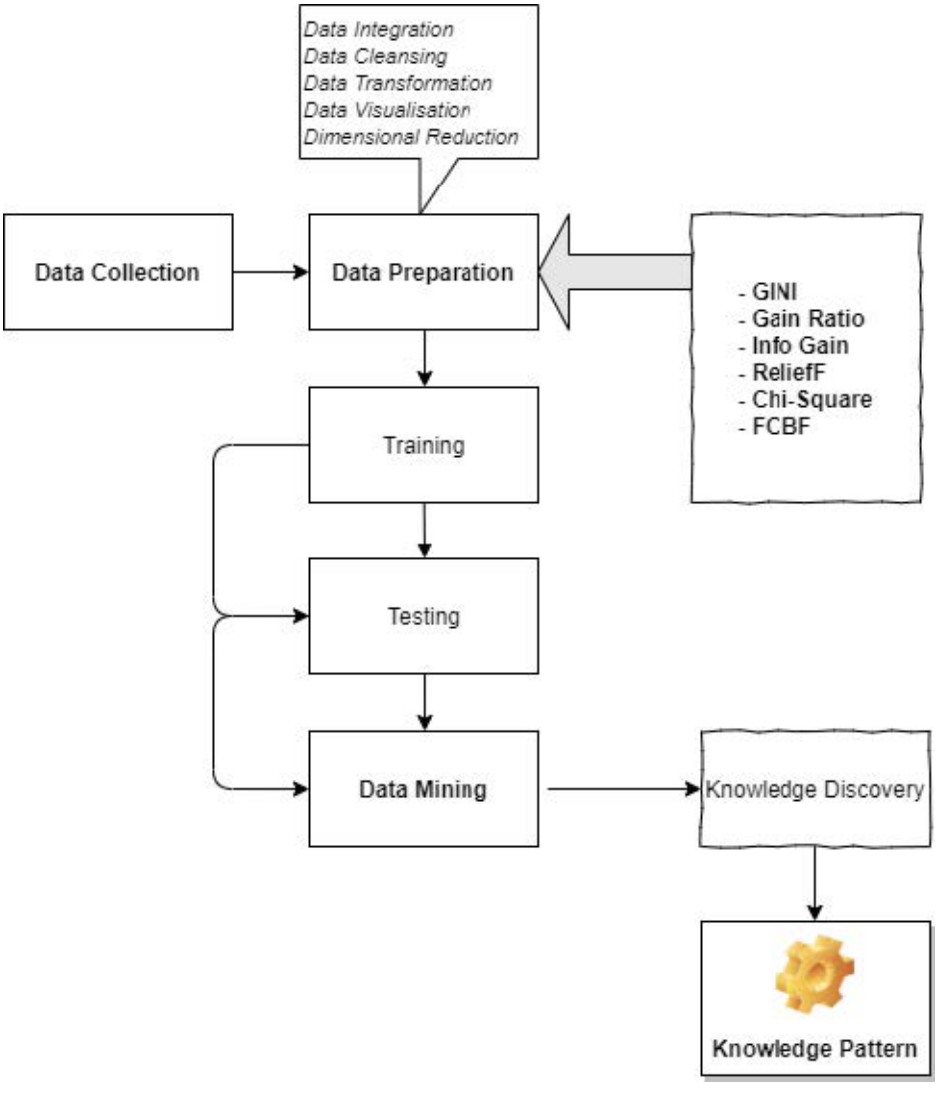


Table 3 prioritises the outstanding features from dengue data using the information gain. It can be seen that Platelet_2 and Platelet_1 are the most weighted characteristics, with scores of 0.035 and 0.034 respectively, while bleeding and leakage scores at 0.026 and 0.025, which is also considered an outstanding feature. However, Platelet_3 has a weight of 0.013, which is less than the weight score of temperature and nausea at 0.019 and 0.017 respectively.

Table 3 Information gain-based ranking

Rating	*Feature*	*Score*
1	Platelet_2	0.035
2	Platelet_1	0.034
3	Bleeding	0.026
4	Leakage	0.025
5	Temperature	0.019
6	Nausea	0.017
7	Platelet_3	0.013
8	Pulse	0.013
9	Shock	0.012
10	Age	0.011

Table 4 ranks outstanding features from dengue data using the Gini ratio. 'Shock' could be observed as the single most prominent feature with a weight score of 0.103, while Dengue_lgM, Dengue_lgG, and Dengue_NS1Ag features are in second of weighting with the same score of 0.074. In addition, bleeding and leakage features received the fifth and sixth scores with 0.046 and 0.044 respectively. Nevertheless, please note that platelet features, which are distinctly weighted by other feature selections, are not significant with this experimentation. Platelet_3 scored 0.022 and Platelet_2 scored the lowest at 0.018, while Platelet_1 did not appear in the table.

Table 4 Gain ratio-based ranking

Rating	*Feature*	*Score*
1	Shock	0.103
2	Dengue_lgM	0.074
3	Dengue_lgG	0.074
4	Dengue_NS1Ag	0.074
5	Bleeding	0.046
6	Leakage	0.044
7	Platelet_3	0.022
8	White_Blood_Cell_3	0.019
9	Nausea	0.018
10	Platelet_2	0.018

Table 5 lists the outstanding features from dengue data using the Gini index method. Platelet_1 and Platelet_2 are the most weighted features, with scores of 0.021 and 0.019 respectively. However, leakage, temperature and bleeding have the same weighting score at 0.011. Please also note that Platelet_3, career and White_Blood_Cell_2 have the smallest weight of 0.006, which is less than that of the age at 0.007.

Table 5 Gini index-based ranking

Rating	*Feature*	*Score*
1	Platelet_1	0.021
2	Platelet_2	0.019
3	Leakage	0.011
4	Temperature	0.011
5	Bleeding	0.011
6	Nausea	0.010
7	Age	0.007
8	Platelet_3	0.006
9	Career	0.006
10	White_Blood_Cell_2	0.006

Table 6 ranks outstanding features from chi-square dengue data. Shock can be seen as the single most prominent feature with a weight score of 64.003. Other features such as Platelet_1, bleeding, leakage and White_Blood_Cell_3 are the second weighted data with scores of 28.642, 28.301, 27.384, and 21.472 respectively. Other remaining ranked data features received scores less than 12.000.

Table 6 χ^2 -based ranking

Rating	*Feature*	*Score*
1	Shock	64.003
2	Platelet_1	28.642
3	Bleeding	28.301
4	Leakage	27.384
5	White_Blood_Cell_3	21.472
6	Pulse	11.834
7	Nausea	11.721
8	Temperature	9.435
9	Platelet_3	7.786
10	White_Blood_Cell_2	3.927

Table 7 prioritises the outstanding features from dengue data using the ReliefF method. Leakage can be seen as the single most prominent feature with a weight score of 0.063, while pulse, bleeding, shock, and Platelet_2 are second weighted data features with scores of 0.039, 0.023, 0.020, and 0.016 respectively. The data features White_Blood_Cell_3 and Rash have the lowest weight of 0.001.

Table 7 ReliefF-based ranking

Rating	*Feature*	*Score*
1	Leakage	0.063
2	Pulse	0.039
3	Bleeding	0.023
4	Shock	0.02
5	Platelet_2	0.016
6	Temperature	0.009
7	Retro_Orbit_Pain	0.009
8	Platelet_1	0.003
9	White_Blood_Cell_3	0.001
10	Rash	0.001

Table 8 ranks the outstanding features from the dengue fever data using the FCBF method. Bleeding and leakage are the most prominent features, with the same weight score of 0.033. Nausea, pulse, and temperature have lower weight scores at 0.018, 0.014, and 0.013 respectively. However, it is interesting to note that the remaining data features have the same weight of 0.000. In other words, all these features are not worth to be considered as predictive features for dengue fever.

Table 8 FCBF-based ranking

Rating	*Feature*	*Score*
1	Bleeding	0.033
2	Leakage	0.033
3	Platelet_2	0.024
4	Nausea	0.018
5	Pulse	0.014
6	Temperature	0.013
7	Platelet_1	0.000
8	Shock	0.000
9	Platelet_3	0.000
10	White_Blood_Cell_3	0.000

4.1 Dengue fever prediction modelling

The features expected to be indicators of different dengue fever types are ranked and further classified using decision tree, Naive Bayes, SVM, neural network, and kNN. Four major assessment criteria, including accuracy, f-measure, precision and recall, were used as a basis to propose dengue fever prediction models.

Figure 2 shows the efficiency of dengue fever prediction model based on a combination of information gain and the classifiers. It can be seen that the most effective classification model for dengue fever data is the neural network that consists of five features with 64.9% accuracy, 71.8% F-measure, 65.7% precision and 79.0% recall. As for the dengue fever data classification model, which is similarly effective is Naive Bayes, consisting of four features, with accuracy = 64.4%, F-measure = 69.4%, precision = 67.4% and recall = 71.5%. However, SVM appears to be the least efficient classification model when compared with others at six features as SVM represents the lowest accuracy (46.2%).

Figure 2 The model accuracy based on information gain and classifiers for predicting dengue fever (see online version for colours)

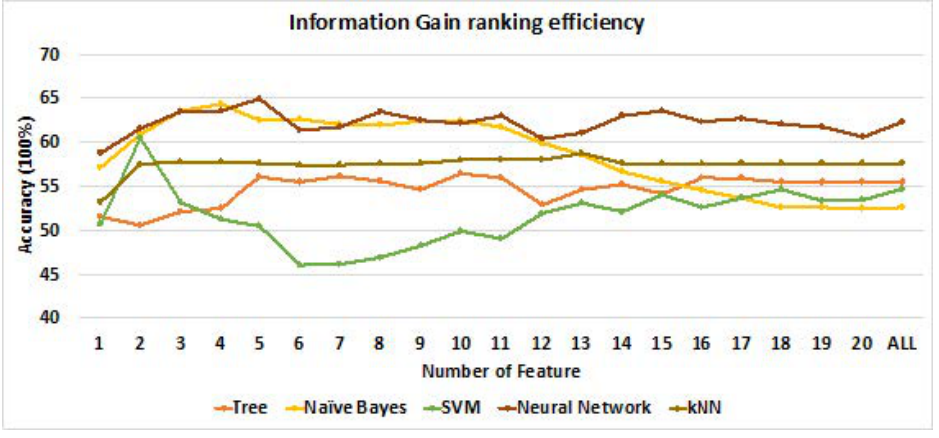


Figure 3 illustrated the efficiency of gain ratio-based dengue fever prediction models experimented with the classifiers. The most efficient model is a combination of gain ration and the neural network consisting of 20 features with 63.7% accuracy, 68.6% F-measure, 67.6% precision and 69.7% recall. Naive Bayes can achieve its highest accuracy 62.1% at 14 features, which is slightly different from that of the neural network at the same features. At 12 features, nevertheless, SVM shows the lowest accuracy (43.8%) representing the most deficient classification in comparison with other classifiers.

Figure 4 shows that the most efficient prediction model for Gini index is the neural network consisting of five features, with 64.2% accuracy, 70.9% F-measure, 65.6% precision and 77.2% recall. Naive Bayes can achieve its highest accuracy 63.0% at six features, which is higher than that of the neural network at the same features. Nevertheless, please note that SVM also shows its lowest accuracy (45.6%) at six features. At this point, SVM can be considered as the most deficient classification model in comparison with other classifiers. Please note that the SVM accuracy rises when the number of features increases.

Figure 3 The model accuracy based on gain ratio and classifiers for predicting dengue fever (see online version for colours)

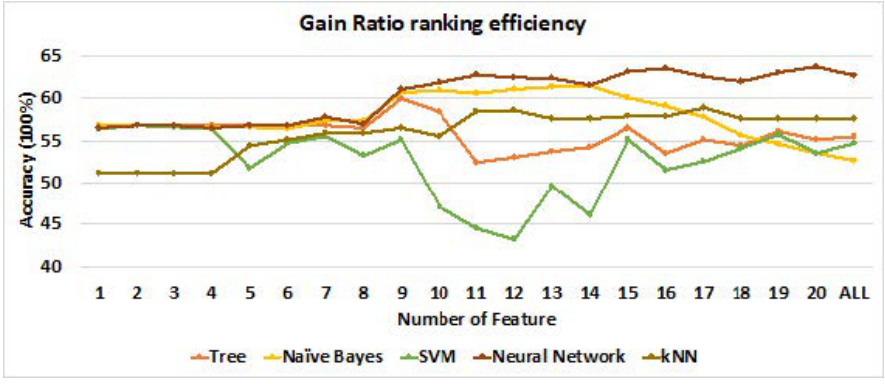
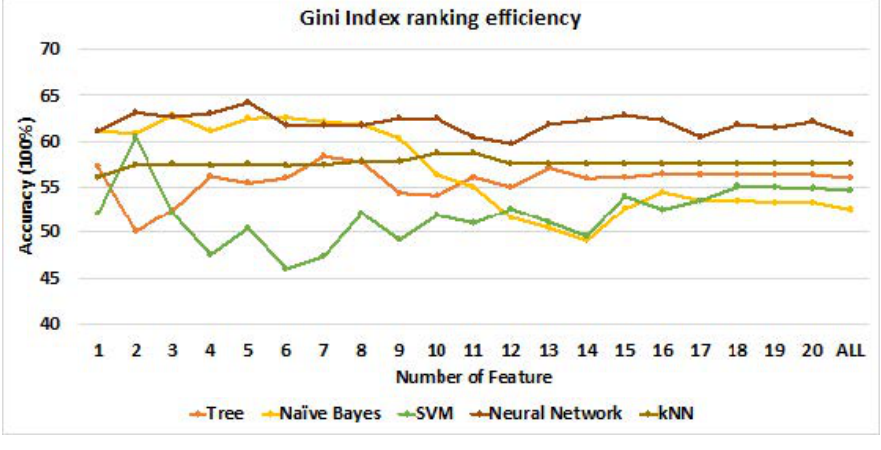
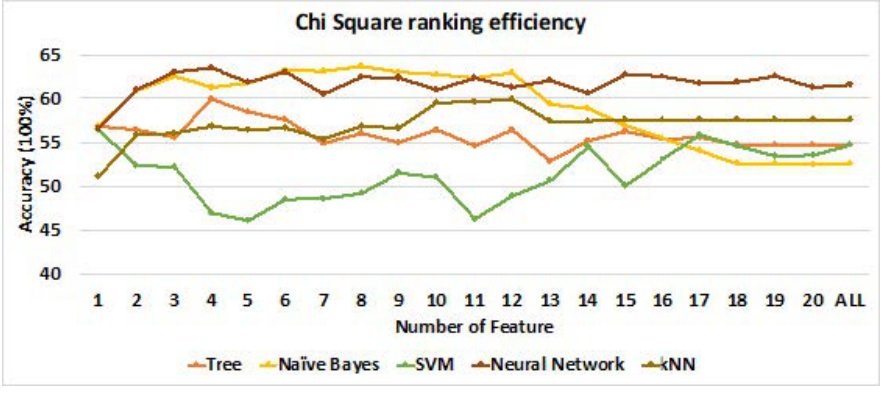


Figure 4 The model accuracy based on Gini index and classifiers for predicting dengue fever (see online version for colours)



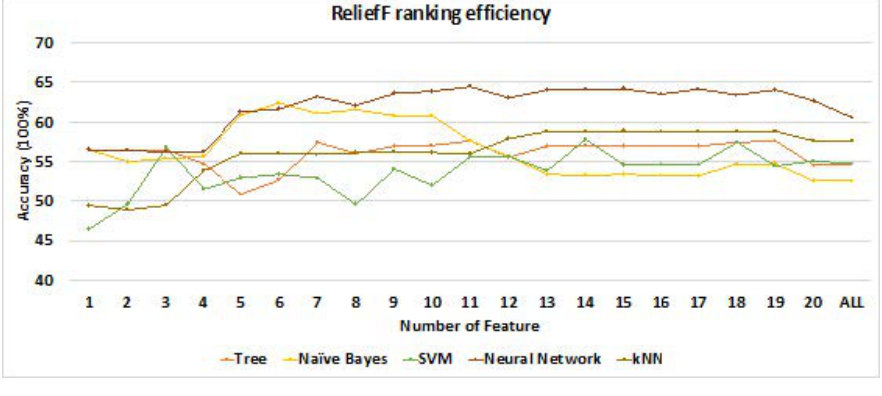
According to Figure 5, the most efficient prediction model for chi-square is a combination with Naive Bayes at eight features with 63.7% accuracy, 69.0% F-measure, 67.2% precision and 70.8% recall. A competitive prediction model is based on the neural network at four features with 63.5% accuracy, 70.6% F-measure, 65.2% precision and 77.0% recall. In the meantime, SVM also shows its lowest accuracy (46.0%) at 5 and 11 features. This may imply an inconsistency between the dataset and the selected classifier.

Figure 5 The model accuracy based on chi-square (χ^2) and classifiers for predicting dengue fever (see online version for colours)



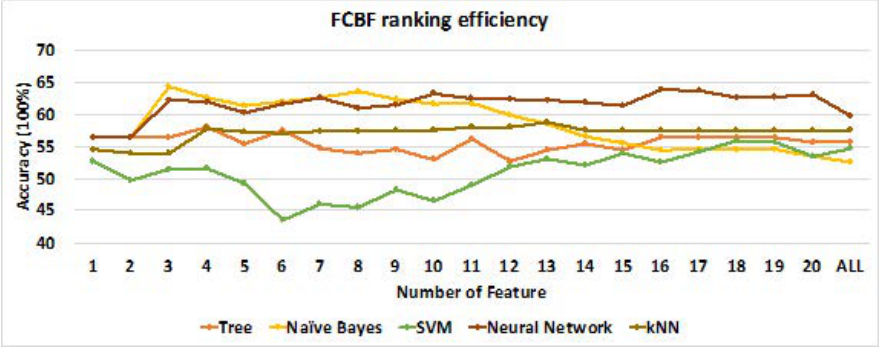
As shown in Figure 6, a combination of ReliefF and the neural network is considered the most efficient dengue prediction model that represents 64.4% accuracy, 69.7% F-measure, 67.1% precision and 72.6% recall at 11 features. Naive Bayes can provide its highest accuracy 62.5% at six features, which is higher than that of the neural network at 61.9% with the same features. However, the lowest accuracy (46.7%) performed by SVM is noticeable at one feature.

Figure 6 The model accuracy based on ReliefF and classifiers for predicting dengue fever (see online version for colours)



From Figure 7, the most effective classification model is Naive Bayes consisting of three features with 64.4% accuracy, 68.5% F-measure, 68.4% precision and 68.6% recall. The neural network can also be considered as an effective classification model as its accuracy is 64.1% at 16 features. Besides, SVM represents its lowest accuracy (43.8%) at six features.

Figure 7 The model accuracy based on FCBF and classifiers for predicting dengue fever (see online version for colours)



5 Conclusions and future work

The overall efficiency of most models is slightly similar. The most efficient model for predicting dengue fever is the machine learning model using a combination of information gain and neural network algorithm at five features. Other models with similar efficiency are also those based on machine learning. For instance, the Naive Bayes algorithm at four features. Through the information gain sequencing model, the machine-based model using the neural network algorithm consists of 11 features. Through the ReliefF sequencing method and the model using machine learning. The Naive Bayes algorithm consists of three features. It is interesting to note that a combination of FCBC and Naive Bayes provides the most efficient results at the lowest dimension. The FCBF ranking method is summarised in Table 9. However, further research must be done as the accuracy of all models is still not satisfactorily high and the results are based on modelling from the dataset only.

Table 9 The effectiveness of the model for the classification of dengue fever data

Algorithm and ranking method	*Accuracy*	*F-measure*	*Precision*	*Recall*
Information gain (neural network)	64.9%	71.8%	65.7%	79.0%
Information gain (Naïve Bayes)	64.4%	69.4%	67.4%	71.5%
ReliefF (neural network)	64.4%	69.7%	67.1%	72.6%
FCBF (Naïve Bayes)	64.4%	68.5%	68.4%	68.6%

The SVM performance, in particular, was not capable of representing desirable accuracy when experimented with the selected feature selection algorithms. As such, further experiment on SVM with its purposely designed feature selection algorithm such as LSVM may be worthwhile. In addition, the model accuracy is still not very high, which might be due to be data quality issues, i.e., the data might be imbalanced. As the efficiency rate is still inadequate for medical health implementations expecting a reliably

predictive model. Further work can be done to obtain improved accuracy. For instance, the data can be further cleaned or analysed thoroughly before mining. Besides, applying ensemble methods (e.g., stacking, bagging, and boosting) may be desirable as they are based on multiple learning algorithms for addressing issues in case the overall quality of the system is higher than the best algorithm. While the model is based on the obtained dataset, further practical research should be conducted to validate whether the model is capable of providing reliable efficiency for assisting medical practitioners to predict dengue fever.

References

Agarwal, S. (2014) 'Data mining: data mining concepts and techniques', *Proceedings – 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*.

Arabshahi, H. and Fazlollahtabar, H. (2018) 'Classifying innovative activities using decision tree and Gini index', *International Journal of Innovation and Technology Management*, Vol. 15, No. 3, pp.1–14.

Azhagusundari, B. and Thanamani, A.S. (2013) 'Feature selection based on information gain', *International Journal of Innovative Technology and Exploring Engineering*, Vol. 2, No. 2, pp.18–21.

Balakrishnan, S. and Narayanaswamy, R. (2009) 'Feature selection using FCBF in type II diabetes databases', *International Conference on IT to Celebrate S. Charmonman's 72nd Birthday*, March, Thailand.

Buathong, W. (2012) 'A comparison of dimensionality reduction techniques using information gain, gain ratio and linear SVM weights ranking methods', *Proceedings of the 5th ACTIS National Conference and 2012 International Conference on Applied Computer Technology and Information Systems*, Songkhla, Thailand, pp.185–189.

Chandra, B. and Saxena, G. (2004) 'A new selection measure for classification using decision trees', *Journal of Information and Knowledge Management*, Vol. 3, No. 1, pp.1–7.

Chen, Y.W. and Lin, C.J. (2006) 'Combining SVMs with various feature selection strategies', *Studies in Fuzziness and Soft Computing*.

Colaco, S., Kumar, S., Tamang, A. and Biju, V.G. (2016) 'A review on feature selection algorithms', *Advances in Intelligent Systems and Computing*, Vol. 906, pp.133–153, Springer Verlag.

Dasgupta, S., Sharma, N., Sinha, S. and Raghavendra, S. (2019) 'Evaluating the performance of machine learning using feature selection methods on dengue dataset', *International Journal of Engineering and Advanced Technology*, June, Vol. 8, No. 5, pp.2679–2685.

Delen, D., Walker, G. and Kadam, A. (2005) 'Predicting breast cancer survivability: a comparison of three data mining methods', *Artificial Intelligence in Medicine*, Vol. 34, No. 2, pp.113–127.

Dougherty, J., Kohavi, R. and Sahami, M. (1995) 'Supervised and unsupervised discretization of continuous features', *Machine Learning Proceedings 1995*.

Furnkranz, J. (2010) *Decision Tree*, pp.263–267, Springer US, Boston, MA.

Giorgi, G. (2011) 'Corrado Gini: the man and the scientist', *Metron – International Journal of Statistics*, April, Vol. 69, No. 1, pp.1–28.

Harris, E. (2002) 'Information gain versus gain ratio: a study of split method biases', *ISAIM*.

Huang, Y., McCullagh, P., Black, N. and Harper, R. (2007) 'Feature selection and classification model construction on type 2 diabetic patients' data', *Artificial Intelligence in Medicine*, Vol. 41, No. 3, pp.251–262.

Husam, I.S., Abuhamad, A.B., Zainudin, S., Sahani, M. and Ali, Z.M. (2017) 'Feature selection algorithms for Malaysian dengue outbreak detection model', *Sains Malaysiana*, February, Vol. 46, No. 2, pp.255–265.

Kavitha, K.R., Gopinath, A. and Gopi, M. (2017) 'Applying improved SVM classifier for leukemia cancer classification using FCBF', *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*.

Kira, K. and Rendell, L.A. (1992) 'A practical approach to feature selection', *Machine Learning Proceedings 1992*.

Kumar, D. and Suman (2011) 'Performance analysis of various data mining algorithms: a review', *International Journal of Computer Applications*, Vol. 32, No. 6, pp.9–15.

Lu, X. and Wang, Z. (2004) 'A comparison among four SVM classification methods: LSVM, NLSVM, SSVM and NSVM', *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, Vol. 7.

Majumder, M. (2015) *Artificial Neural Network*, pp.49–54, Springer.

Manek, S., Shenoy, D., Mohan, C. and Venugopal, R. (2017) 'Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier', *World Wide Web*, Vol. 20, No. 2, pp.135–154.

Ministry of Public Health (Thailand) (2020) *Dengue Fever Prognosis Report 2019*, in Thai, Technical report.

Muller, M.E. and Muller, M.E. (2012) Information gain', *Relational Knowledge Discovery*, pp.92–120, Cambridge University Press, Cambridge.

Ozyildirim, B.M. and Avci, M. (2013) 'Generalized classifier neural network', *Neural Networks*, Vol. 39, No. 1, pp.18–26.

Pandiyarajan, P. and Thangairulappan, K. (2018) 'Classification of dengue gene expression using entropy-based feature selection and pruning on neural network', *Advances in Intelligent Systems and Computing*, Vol. 736, pp.519–529, Springer Verlag.

Pandiyarajan, P. and Thangairulappan, K. (2019) 'Classification of dengue serotypes using gini-index based feature selection and rule extraction from neural network', *Journal of Advanced Research in Dynamical and Control Systems*, Vol. 11, No. 4, Special Issue, pp.1620–1629.

Priya, M. and Ranjith Kumar, P. (2015) 'A novel intelligent approach for predicting atherosclerotic individuals from big data for healthcare', *International Journal of Production Research*, Vol. 53, No. 24, pp.7517–7532.

Raileanu, L.E. and Stoffel, K. Theoretical comparison between the Gini Index and Information Gain criteria', *Annals of Mathematics and Artificial Intelligence*, Vol. 41, No. 1, pp.77–93.

Renuka Devi, B., Nageswara Rao, K. and Pallam Setty, S. (2015) 'Towards better classification using improved genetic algorithm and decision tree for dengue datasets', *International Journal of Applied Engineering Research*, Vol. 10, No. 8, pp.20313–20326.

Renuka Devi, B., Nageswara Rao, K. and Pallam Setty, S. (2016) 'Towards better classification using improved particle swarm optimization algorithm and decision tree for dengue datasets', *International Journal of Soft Computing*, Vol. 11, No. 1, pp.18–25.

Robnik-Šikonja, M. and Kononenko, I. (2003) 'Theoretical and empirical analysis of ReliefF and RReliefF', *Machine Learning*, Vol. 53, No. 1, pp.23–69.

Senliol, B., Gulgezen, G., Yu, L. and Cataltepe, Z. (2008) 'Fast correlation based filter (FCBF) with a different search strategy', *2008 23rd International Symposium on Computer and Information Sciences, ISCIS 2008*.

Shaukat, K., Masood, N., Mehreen, S. and Azmeen, U. (2015) 'Dengue fever prediction: a data mining problem', *Journal of Data Mining in Genomics & Proteomics*, Vol. 6, No. 3, pp.1–5.

Symeonidis, A.L., Nikolaidou, V. and Mitkas, P.A. (2006) 'Exploiting data mining techniques for improving the efficiency of a supply chain management agent', *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, December, pp.23–26

Tanner, L., Schreiber, M., Low, J.G.H., 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. and Ooi, E.E. (2008) 'Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness', *PLoS Neglected Tropical Diseases*, Vol. 2, No. 3, pp.1–9.

Thitiprayoonwongse, D., Suriyaphol, P. and Soonthornphisaj, N. (2011) 'Data mining on dengue virus disease', *Proceeding of the 13th Conference on Enterprise Information Systems (ICEIS2011)*, pp.32–41, Beijing, China.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V. (2001) 'Feature selection for SVMs', *Advances in Neural Information Processing Systems, Papers from Neural Information Processing Systems (NIPS) 2000*, MIT Press, Denver, CO, USA, Vol. 13, pp.668–674.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. (2016) *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Elsevier Science.

World Health Organisation (2020) 'Dengue and severe dengue', *WHO Fact Sheets*, March, Vol. 117.

Zeng, X.Q., Li, G.Z. and Chen, S.F. (2010) 'Gene selection by using an improved fast correlation-based filter', in *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW 2010)*, IEEE Computer Society, Hong Kong, pp.625–630.

Zou, H. and Jin, Z. (2018) 'Comparative study of big data classification algorithm based on SVM', *2018 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference, CSQRWC 2018*.