Data mining in medical diagnostic support system

Khoa Nguyen
The health and education are always a vital issue for any countries in the world. In recent years, Vietnamese government has especially invested in these two main spearhead sectors adopting policies and capitals for infrastructure equipment and scientific research. In the field of scientific research, there have been more and more medical scientific works. However, there are not many information technology applications used to solve medical problems. Due to geographical characteristics of a tropical country, there are many diseases related to ultra-viral fever, such as petechial fever which is a very dangerous disease. Dengue vaccine is contemporarily not so popular in Vietnam. Therefore, this project focuses mainly on laws of dengue diagnose in Vietnam applying data mining technique. Based on clinical and subclinical symptoms, we can classify diseases of patients to help doctors diagnose and treat them better.

This research is conducted in four main steps: (1) Finding the medical specialist skills for petechial fever. Next, (2) collecting and pre-processing the data. Then, (3) Learning data mining classification algorithms to choose which would be suitable for the inquiries and collected data.

In addition, the thesis also proposes a method of coordination among information technology and medical experts to build diagnostic support system for different types of diseases. This brings out a practical meaning for rural and remote areas because of lacking in medical equipment for the first aids.
Table of contents

1 Overview of the thesis topic ................................................................. 1
   1.1 Introduction to data mining ............................................................... 1
   1.2 The background of the project ........................................................... 1
   1.3 Research results inside and outside the country ..................................... 2
       1.3.1 The world’s results ................................................................. 2
       1.3.2 The research results inside Vietnam .......................................... 2
   1.4 The aim of the project ........................................................................ 2
   1.5 Objectives and research methods ....................................................... 2
   1.6 Meaning behind the project ............................................................... 3
       1.6.1 Scientific meaning .................................................................... 3
       1.6.2 Practical meaning ..................................................................... 3
   1.7 Outline of the thesis ........................................................................... 3
2 Theoretical framework ......................................................................... 4
   2.1 Overview of data mining techniques .................................................. 4
       2.1.1 Data mining concept ................................................................. 4
       2.1.2 Stages of data mining process (Cuong, 2012) ............................. 4
   2.2 Overview of decision support system ................................................. 5
2.3 Classification in data mining ............................................................... 6
   2.3.1 Definition of classification ............................................................. 6
   2.3.2 Data classification process ............................................................ 6
   2.3.3 Data classification with decision tree .......................................... 9
   2.3.4 Evaluation the effectiveness of classification .................................. 12
   2.3.5 C4.5 algorithm to build the decision tree .................................... 12
2.4 Medical database ............................................................................... 15
   2.4.1 Overview of dengue disease .......................................................... 15
   2.4.2 Clinical progress of dengue diseases .......................................... 16
   2.4.3 Diagnosis ................................................................................... 17
3 Building medical diagnostic support system ....................................... 18
   3.1 Database model ............................................................................... 18
       3.1.1 Medical records in data warehouse ........................................... 18
       3.1.2 Preprocessing data ................................................................... 20
       3.1.3 Analysis of digital medical records ......................................... 23
       3.1.4 Diagnostic rules ...................................................................... 25
       3.1.5 Sample medical records ............................................................ 25
       3.1.6 Diagnose ................................................................................ 25
   3.2 Building the application .................................................................. 25
       3.2.1 Introducing to the program ...................................................... 25
List of Figures

Figure 1. Result of classification process .......................................................... 7
Figure 2. Classification model building ............................................................... 8
Figure 3. Classification step ............................................................................... 8
Figure 4. Model of decision tree in classifying weather data ................................ 10
Figure 5. Solution for supporting diagnosis model .......................................... 18
Figure 6. Collected data .................................................................................. 19
Figure 7. Training data set is entered into the system ....................................... 23
Figure 8. NS1 is selected due to the best measurement .................................... 24
Figure 9. Starting screen .................................................................................. 26
Figure 10. Select attributes screen .................................................................... 27
Figure 11. Create decision tree screen .............................................................. 29
Figure 12. 10 popular rules screen .................................................................... 30
Figure 13. Testing result model ........................................................................ 31
Figure 14. Data test result ................................................................................. 32
Figure 15. Cross validation screen .................................................................... 33
Figure 16. Result of cross validation screen ...................................................... 33
Figure 17. Diagnosis screen .............................................................................. 34
Figure 18. Result from created decision tree ................................................... 35
Figure 19. Image of dengue with warning sign .................................................. 38
Figure 20. Image of dengue .............................................................................. 39
Figure 21. Image of severe dengue ................................................................... 40
List of Tables
Table 1. Table of weather data ........................................................................................................9
Table 2. Example of weather within 3 days ......................................................................................10
Table 3. Data classification result for table 2 (Jiawei, Micheline & Jian, 2011) ...................11
Table 4. Training with buys computer classification attribute .....................................................14
Table 5. Table of physical body condition based on BMI (Binh, 2007) .................................19
Table 6. Discrete HCT data .............................................................................................................20
Table 7. Discrete PLT data ...............................................................................................................21
Table 8. Discrete WBC data .............................................................................................................21
Table 9. Data table of attributes .....................................................................................................21
Table 10. Table of classification statistics and diagnostic time support (the results of conducting the application). .................................................................................................................36
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBP</td>
<td>Systolic blood pressure</td>
</tr>
<tr>
<td>DBP</td>
<td>Diastolic blood pressure</td>
</tr>
<tr>
<td>BMI</td>
<td>Body mass index</td>
</tr>
<tr>
<td>BP</td>
<td>Blood pressure</td>
</tr>
<tr>
<td>RBC</td>
<td>Red blood cells</td>
</tr>
<tr>
<td>WBC</td>
<td>White blood cells</td>
</tr>
<tr>
<td>IgM</td>
<td>Immunoglobulin M</td>
</tr>
<tr>
<td>IgG</td>
<td>Immunoglobulin G</td>
</tr>
<tr>
<td>NS1</td>
<td>Non-structural protein 1</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and communications technology</td>
</tr>
<tr>
<td>IT</td>
<td>Information technology</td>
</tr>
<tr>
<td>BI-RADS</td>
<td>Breast Imaging Reporting and Data System</td>
</tr>
<tr>
<td>HTC</td>
<td>Hematocrit</td>
</tr>
<tr>
<td>PLT</td>
<td>Platelets</td>
</tr>
<tr>
<td>ID3</td>
<td>Iterative dichotomiser 3</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and regression tree</td>
</tr>
<tr>
<td>SLIQ</td>
<td>Supervised learning in quest</td>
</tr>
<tr>
<td>SPRINT</td>
<td>Scalable PaRallelization INduction of decision trees</td>
</tr>
<tr>
<td>ECG</td>
<td>Electrocardiographic</td>
</tr>
<tr>
<td>Num</td>
<td>Number</td>
</tr>
<tr>
<td>ID</td>
<td>Identity</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
1 Overview of the thesis topic

1.1 Introduction to data mining

Nowadays, most industrial and business fields are applying information technology to the storage and processing data, this has created a very large amount of data that stored and increased constantly. It is a good chance for mining the data in warehouse to provide useful knowledge with query tools, query tables, and data mining. Data mining is a technology based on many theories such as probability, statistics, machine learning to find potential knowledge in large sized data warehouses where users have difficulty to make use by conventional techniques. The source of medical data is quite great. If applying data mining in this field it will bring a lot of meaning to the health sector. It will provide valuable information to assist in early examination and treatment to help patients get off many serious sicknesses.

In the field of current Vietnamese medicine; especially; the situation of communes, deep-lying and remote areas, they still lack medical personnel with professional qualifications and lack necessary analysis devices. Therefore, it is important to build a diagnostic support system that is essential for the ongoing health sector in Vietnam. The support system will work with medical staff to help early detection of some dangerous diseases which can reduce economic burden for patient’s families and society. In order to demonstrate the benefits which is offered by the analytic support system, the project selects dengue data for testing and evaluating.

Application of data classification techniques in data mining to build a diagnostic support system is one of the main scrutiny directions of the thesis. After analysing a few algorithms as well as the characteristics of collected data on dengue fever, the thesis proposes to apply a decision tree classification model using C4.5 method to find the unknown rules in the data.

1.2 The background of the project

According to the announcement of the World Health Organization, there are 3.9 billion people in the world living in the circulating dengue region including 1.8 billion people in Asia Pacific region (Vietnam General Department of Preventive Medicine, 2015). Vietnam is a country with a high prevalence of dengue disease, dengue is always one of the leading high-risk infectious diseases every year. There are over 100,000 dengue patients and nearly 100 people die because of this infection and Vietnam ministry of health always pays attention to the key tasks of national anti-dengue prevention program (Vietnam General Department of Preventative Medicine, 2015). Thereupon, building a medical examination support system is to contribute to rapid diagnosis and early detection of risks and it is
a considerable issue of several families and society. The topic applies ICT to build the system using data collected from dengue.

1.3 Research results inside and outside the country

1.3.1 The world’s results

The world has launched many applications from support system for quick examination and for better treatment of sickness, for instance Harry Pope’s Caduceus medical diagnosis system (Siegel, 2012); Diagnosis Pro Medical Expert System (Aronson, 1997); MYCIN system supports to diagnosis blood infection (Buchanan & Shortliffe, 1984); PUFF is used to analyze test results of lung function (Aikins, Kunz, Shortliffe, & Fallat, 1983); PSG-Expert diagnoses insomnia (Fred, Filipe, Partinen, & Paiva, 2000); BI-RADS breast cancer diagnosis (Ngah & Aziz, 2007); Naser developed a diagnostic system for skin diseases in 2008 (Naser & Akkila, 2008); Compete manages hypertensive patients, diabetes and chronic diseases and so on.

1.3.2 The research results inside Vietnam

In Vietnam, the circumstance of IT application in health is still relatively rare. At the end of 1980, there were also studies that supported doctors to inspect internal illnesses, acupuncture and diagnosis of oriental medicine (Thuy, 1996), support decisions making system in clinical diagnosis (Thanh, 2000) etc. Nevertheless, researches on medical diagnostics aim to build decision support system is extremely limited.

1.4 The aim of the project

Acknowledge about the importance of datamining technology in real life, the main task of this thesis project will first focus on discussing about data mining classification and then deeply illustrate its applicability in medical diagnosis field. Thereby, filtering the algorithms as a premise for research and building specific applications. Additionally, gathering data of a specific disease is also noticeable. In other words, petechial fever information used in the thesis is taken from Pham Ngoc Thach hospital in Vietnam. After that, analysing the characteristics of collected data and selecting the appropriate algorithms for the data. Next, building and evaluating the quality and effectiveness of the diagnostic support system is another goal of the research.

1.5 Objectives and research methods

The project addresses to study classification techniques in data mining (namely, investigate algorithm C4.5) to apply to the analysis of medical databases. The dissertation thesis
collects data on dengue diseases of all patients (regardless to age and gender) who examined and cured at Pham Ngoc Thach hospital. Making use retrospective cohort study method with professional support of specialist doctors. The research conducts on the basis of data classification algorithms in data mining.

1.6 Meaning behind the project

1.6.1 Scientific meaning

Thanks to the aid of computers, the project contributes a procedure to support medical staff in daily working routine. The results and experiences gained from implementing this project will help health personnel detect diseases earlier for patients, and contemporarily expect those who working in the field of medicine and computer science to sit back together to find better solutions for health care.

1.6.2 Practical meaning

Both disease diagnosis and disease detection are a long process that requires not only strong professional skills but also sufficient and adequate medical equipment so as to treat patients precisely. Incorrect symptoms of a disease examination drives to undetachable disease and wrong therapeutics. This leads to huge mental and physical losses for patients and their families. The possibility of treatment failure will decrease if it is advance inspected or patients’ families can make appropriate medical attention decisions. Accordingly, by applying the technology, it can help medical team identify symptom more accurately and strategically while it can monitor, warn and counsel patients to avoid dangerous complications and to reduce the economic burden for their families as well as for society.

1.7 Outline of the thesis

The thesis includes the following sections:
Section 1: Overview of the thesis topic
Introduction to issues related to data classification in data mining, background of the project, objective, scope of research, meaning behind the research and thesis layout.
Section 2: Theoretical background
This section speaks up the approach and solution of the thesis problems. Present the mathematical basis and apply the theory to those problems. Basic knowledge about dengue is also explained.
Section 3: Building the diagnostic support system
In this chapter, the characteristics of the data are mentioned, the steps of pre-processing data are found out before being applied to the system. Build and develop diagnostic applications based on dengue data.

Section 4: Experimental setup and evaluation
Launch the program with the training data set. Draw the decision tree diagram and its rule set. Test and evaluate the program with the testing data set.

Section 5: Conclusion
Practical significance of the community, limitations and development direction of the project.

2 Theoretical framework

2.1 Overview of data mining techniques

2.1.1 Data mining concept

Data mining has been attracting the attention of ICT and social industry in recent years. Along with the development of information technology, data stored every day is becoming an extremely huge database. Relying on this volume of data, people use data mining techniques to extract them into useful information or to draw new knowledge from collected data (Jiawei, Micheline & Jian, 2011). On the other hand, professor Tom M. Mitchell defines data mining as following: “using historical data to discover regularities and improve future decisions”.

Data mining can be used in public health, market analysis, construction... and can be viewed as a goal of tremendous IT evolution. Especially, data mining in health care will contribute to create diagnostic systems that support doctors in diseases diagnosis.

2.1.2 Stages of data mining process (Cuong, 2012)

a. Business and data understanding

In this step, we identify the problem that needs to be solved. Understanding the goal of the project and needs from the business view. Since, we determine the exact data source to collect. At the same time, it is necessary to understand its data structure, meaning and importance so that we can come up with a specific solution for that matter (Tutorials Point, n.d).

b. Data preparation

This step involves using data pre-processing techniques to handle the collected data, so data mining algorithms can be understood. This phase is where data mining researchers spend most of their time. The following tasks include:
• Settle missing or lost data: missing or lost values will be replaced by more appropriate values or deleting wrong domain and resolving the inconsistencies in data.
• Eliminate data duplication: remove duplicated data.
• Reduce data noise: noisy data will be adjusted or removed from the database.
• Transform data: digital data will be disjointed into a form which is suitable for mining, by performing summary or aggregation operations.
• Reduce dimensionality: eliminate attributes with little information, random variables to save time and resources for computer.

c. Data modelling
At this step, using the algorithms of data mining to fine the rules of data. The most important in this period is to find the right algorithm to solve the problem.

d. Post-processing and model evaluations
This stage is concerned in changing from the drawn rules (of the previous period), from the training set to the appropriate form for researching problem. Besides, it will also be the evaluation phase of the consultants based on the testing data set. From the comments and support of experts, the models of previous stages will be promptly adjusted. Satisfactory models will be implemented in professional use.

e. Deploy the model
Approved models will be built into implemen tal application programs to support decision making as users’ requirement.

2.2 Overview of decision support system

Decision support system is an information system which provides or analyses data to assist decision making as references and solutions. Decision support systems can be used either indirectly or directly for individuals or organizations (Mora, Forgionne & Gupta, 2003).

In medicine, knowledge-based decision-making support systems provide diagnostic information to medical officers. This information is filtered to wisely provide a valuable way for more effective diagnosis, monitoring and treatment of illnesses. We can see then, there are some benefits of decision-making support in medical care:

• Improve the quality of diagnosis and treatment.
• Reduce the risks of errors to minimize dangerous circumstances for patients.
• Enhance the effectiveness of applying information technology in health care to reduce unnecessary paperwork as well as redundant procedures and so on.
2.3 Classification in data mining

2.3.1 Definition of classification

Classification is a process of data analysis that describes and distinguishes data classes or concepts. These models are called classifiers or classification models. They are used to predict the class of objects whose class label is categorical (discrete, unordered) for new data objects (Kulwinder Kaur, 2018).

2.3.2 Data classification process

Data classification is a two-step process:

- The first step (training phase) of the classification process is the learning of a mapping or function. The learning process aims to build a classifier model which includes pre-conceptual data class from the input data set. Training phase uses a classification algorithm to classify the tuples of training data while the training data set is structured with corresponding attributes (Jiawei, Micheline & Jian, 2011).

- In the second step (Figure 1), the models found in the first step will be used for sorting new data. We use a test set, made up of test tuples and their associated class labels, to compare the output of the classifier. These test tuples have not been used to construct classifier in step 1 (Jiawei, Micheline & Jian, 2011). The results of the classification model are shown:
IF $a = y$ and $b = y$ then class $x$
IF $a = n$ and $c = y$ and $d = y$ then class $x$

Illustrated example:

Step 1: Model building

Purpose: Classify patients into 2 classes: “Positive” and “Negative” in the classifier labelled “Diagnostic results”. Each patient has the following attributes for classification: HCT (Hematocrit), PLT (Platelets), NS1 (Non-structural protein 1). After training, we get the following classification model
Step 2: Classification

Evaluating the results of the classification model from step 1, we use the test data set. With a new sample, use the classifier to classify this sample into one of the classes drawn from the model in step 1. In the test data of Figure 3, patient Khai has the values: HCT = 59.3; PLT = 160.1; NS1 = “Positive”, the model will classify this case as “Diagnostic Results” = “Positive” (Figure 3).

Some problems of classifier that need to be addressed:

- **Accuracy**: The reliability is based on the accuracy of classification.
- **Speed**: Under some circumstances, the speed of classification is considered as an important factor.
• Easy to understand: An easy-to-understand classifier will give users more confidence in the system, while also help them avoid misunderstanding the results given by the system.

• Simplicity: The result of the decision tree is related to its size.

• Time to learn: When a system operates in an environment that changes frequently, it requires the system to learn very quickly the rule of classification or adjust the rule that has been learned to better suit the reality.

The techniques of classification:

• Classification model using decision tree classification

• Classification using neural network

• Classification using Bayesian network

• Classification with the K-nearest neighbour classifier

• Statistical analysis

• Generic algorithms

• Rough set approach

2.3.3 Data classification with decision tree

a. Definition:

Decision tree is the result of training a data set process with tuples which have already had attributes. Decision tree is a common tool in data mining and data classification (Shafer, Agrawal & Mehta, 1996).

The characteristic of a decision tree is that the tree has a flowchart structure, in which:

• Root: is the top node of the tree

• Internal node: represents the test on a single attribute (oval shape)

• Branch: represents the results of the test on the node

• Leaf node: represents the classes or the class distribution (square or rectangle shape)

For instance: we have a table with 14 objects and 5 attributes, in which the “Play” attribute is a classification attribute.

Table 1. Table of weather data

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>True</td>
<td>No</td>
</tr>
<tr>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Mild</td>
<td>High</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cool</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cool</td>
<td>Normal</td>
<td>True</td>
<td>No</td>
</tr>
<tr>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>True</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Based on the classification process, we can create a decision tree as follows (Jiawei, Micheline & Jian, 2011):

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>False</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Mild</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>True</td>
<td>Yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>True</td>
<td>Yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Mild</td>
<td>High</td>
<td>True</td>
<td>No</td>
</tr>
</tbody>
</table>

b. Using decision tree in predicting the class for unknown data:

The purpose of decision tree is to predict the class for unknown data objects. Suppose we know that the weather in the 3 upcoming days with known attributes “Outlook”, “Temperature”, “Humidity”, and “Windy”. However, we still do not know how sport players decide (the “Play” classification attributes), whether “Yes” or “No”. Hence, with 4 given attributes, we use the above decision tree (Figure 4) to predict how they will decide.

Table 2. Example of weather within 3 days
We start from the root node of the tree (the “Outlook” attribute in Figure 4). If “Outlook” is Overcast, the player’s decision will be Yes. If “Outlook” is Sunny, the tree decides to continue considering “Humidity” attribute. If “Humidity” is High, the decision tree will be No, and if “Humidity” is Normal, the decision will be Yes. Next, if “Outlook” is Rainy, the tree decides to continue considering “Windy” attribute. If “Windy” is False, the decision will be Yes while “Windy” is True, the decision will be No.

In summary, according to the above decision tree, there are some arising series of the rules:

Rule 1: If outlook is Overcast then Play = Yes
Rule 2: If outlook is Sunny and Humidity is High then Play = No
Rule 3: If outlook is Sunny and Humidity is Normal then Play = Yes
Rule 4: If outlook is Rainy and Windy is False then Play = Yes
Rule 5: If outlook is Rainy and Windy is True then Play = No

Based on these rules, we have the classification results for the data in Table 2 as follows:

Table 3. Data classification result for table 2 (Jiawei, Micheline & Jian, 2011)

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cool</td>
<td>High</td>
<td>True</td>
<td>No</td>
</tr>
</tbody>
</table>

Through the above example, we can see that the classification using decision tree is relatively easy to understand. However, decision tree depends greatly on training data and classification algorithms.

c. Evaluating decision tree in the field of data mining:

Advantages:

- The process of building a decision tree does not require the knowledge about current research data or input parameters.
- The results of the training process are represented in the form of a tree so that they are easy to understand and friendly to users.
- In general, the decision tree algorithms give results of relatively high accuracy.

Disadvantages:

- For the data sets with many attributes, the decision tree will be large, calculations can get very complex, particularly if many values are uncertain, thus, reducing the “easy to understand” factor.
• The ranking of attributes to branch will be based on the previous branching and ignore the interdependence among attributes.

• When using information gain to determine branching attributes, the attributes with multiple values are usually preferred.

\[ \text{d. The algorithms of the decision tree:} \]

• ID3 (Decision tree)
• C4.5
• CART (Classification and regression tree)
• SLIQ (Supervised Learning in Quest)
• SPRINT (Scalable parallelization induction of decision trees)

2.3.4 Evaluation the effectiveness of classification

The classification model, after being created, needs to be evaluated for the effectiveness. In order to evaluate the model, this thesis will mention 2 popular assessment methods: Holdout and K-fold cross-validation. These 2 techniques are both based on the original random data set. Usually, the training data is divided into 2 parts with 70% for training and 30% for testing (Milener, Guyer & Rabeler, 2018). Nonetheless, this ratio can be changed freely.

a. Holdout method:
The original data set is randomly divided into 2 parts (70% for training and 30% for testing) as mentioned above. After the training process, the model is formed and the program will rely on the testing data to check how many percentages of correctness the model has in total number of testing data.

b. K-fold cross-validation method:
This is an important method in evaluating and developing training models. Same as the holdout method, the original data set is also divided into 2 parts but with “k” subset (fold). Then the training data will be randomly changed with “k” times to create “k” different models. Each model formed from the training set will be evaluated with the corresponding testing data set. The model with the best result will be selected.

In addition, it is possible to evaluate the effectiveness of classification from an independent testing data set which configured similar to the training set.

2.3.5 C4.5 algorithm to build the decision tree

a. Overview:
C4.5 (an improved algorithm of ID3) is a classification algorithm to create decision trees. This algorithm was developed by J. Ross Quinlan (Morgan Kaufmann Publishers, 1993).
The decision trees created by the C4.5 algorithms are simple, easy to use and understandable because the rules created at the leaf node of the tree can be represented in the form of if-then command.

b. Pseudocode of C4.5 algorithm:

1. **Function** C45_builder(set_A, set_attribute)
2.  
3.     if (all records in set_A are all in the same class)
4.         
5.             return a leaf node labeled by that class  
6.         
7.     else
8.     
9.         if (set_attribute is empty)
10.            
11.                return leaf node is labeled by all classes in set_A
12.            
13.        else
14.        
15.            Select a P attribute, taking it as the root for the current tree;  
16.        
17.            For each (V value of P)
18.  
19.                Create a branch of the tree labeled V;  
20.                Put in partition V the examples in set_A which have value V at attribute P;  
21.                Call C45_builder (partition V, set_attribute), attach results to branch V;
22.  
23.        
24.        


c. C4.5 algorithm using Gain-entropy:

The smaller size the tree after being trained, the higher the accuracy and the easier to understand. Therefore, C4.5 algorithms will base on measurement to choose the best attribute. Two measurements used in C4.5 are information gain or gain ratio.

S: training set  
S_i: class of the set of classes C_i (i=1,…,m)
\[ I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{m} \frac{S_i}{S} \log_2 \frac{S_i}{S} \]

- Information for classification (Li, 2011):

- Entropy for classification (Li, 2011):

Suppose attribute A is chosen for training, \( A = \{S'_1, S'_2, \ldots, S'_3\} \) then entropy of A will be calculated as follows:

\[ \text{Ent}(A) = \sum_{j=1}^{v} \frac{S'_j}{S} \left( -\sum_{i=1}^{m} \frac{S'_{ij}}{S'_j} \log_2 \frac{S'_{ij}}{S'_j} \right) \]

In which \( S'_{ij} \) is classification cases of \( S' \)

- Information gain for classification (Li, 2011):

Information gain obtained by branching on attribute A is calculated as follows:

\[ \text{Gain}(A) = I(S_1, S_2, \ldots, S_m) - \text{Ent}(A) \]

The attribute with the largest information gain is chosen as the division criteria.

- Gain ratio (Li, 2011):

\[ IV(A) = -\sum_{j=1}^{v} \frac{S'_j}{S} \log_2 \frac{S'_j}{S} \]

\[ \text{Gain_Ratio} = \frac{\text{Gain}(A)}{IV(A)} \]

Describe how to calculate information gain

For example: we have the following data on Buy Computer:

Table 4. Training with buys computer classification attribute

<table>
<thead>
<tr>
<th>Rid</th>
<th>Age</th>
<th>Income</th>
<th>Student</th>
<th>Credit_rating</th>
<th>Class: buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>&lt;30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>30-40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>30-40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>&lt;30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>&lt;30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>&lt;30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>30-40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>30-40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>

In the training data set above:

\( S_1 \): Buys_computer = "yes".

\( S_2 \): Buys_computer = "no".
With the attributes: Age; Income; Student and Credit_rating. We calculate:

- Information:
  \[ I(S) = I(S_1, S_2) = I(9, 5) = -\frac{9}{14} \times \log_2 \frac{9}{14} - \frac{5}{14} \times \log_2 \frac{5}{14} = 0.94 \]

- Entropy: Calculate entropy of all the attributes in table 4 as follows:

  **Age attribute:** The age attribute has been discretized into values \( S'_1, S'_2, S'_3 \) corresponding to age \(< 30\), age \(30-40\), and age \(> 40\) respectively.

  - With age = \( S'_1 = "<30" \) then \( S'_{11} = 2; S'_{21} = 3 \)
    \[ I(S'_1) = I(S'_{11}, S'_{21}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971 \]
  
  - With age = \( S'_2 = "30-40" \) then \( S'_{12} = 4; S'_{22} = 0 \)
    \[ I(S'_2) = I(S'_{12}, S'_{22}) = 0 \]
  
  - With age = \( S'_3 = ">40" \) then \( S'_{13} = 3; S'_{23} = 2 \)
    \[ I(S'_3) = I(S'_{13}, S'_{23}) = 0.971 \]

  In which entropy of age attribute is

  \[ Ent(age) = \sum \frac{|S_i|}{|S|} \times I(S_i) = \frac{5}{14} \times I(S'_{11}) + \frac{4}{14} \times I(S'_{12}) + \frac{5}{14} \times I(S'_{13}) = 0.694 \]

- Information gain: calculate information gain for each attribute

  \[ Gain(age) = I(S) - Ent(age) = 0.246 \]

  Now, similarly to the other attributes, we can calculate entropy and information gain:

  - Gain(income) = 0.029
  - Gain(student) = 0.151
  - Gain(credit_rating) = 0.048

  The age attribute is the one with the largest information gain. Therefore, age is chosen as the development attribute at the node under consideration.

### 2.4 Medical database

#### 2.4.1 Overview of dengue disease

Dengue is a disease caused by dengue virus (this virus has 4 different types), the disease occurs in areas with tropical and subtropical climate. The disease is spread by Aedes aegypti mosquito. Currently, dengue has only one licensed vaccine which is Dengvaxia and two other candidate vaccines are being evaluated in large phase 3 trials (World Health Organization, 2017). Unfortunately, it has no specific anti-viral treatment. This disease can happen to any individual or outbreak in a residential area. The symptom can be easily confused with some other diseases, so if not being diagnosed and healed promptly, it will be very dangerous and can lead to death.
2.4.2 Clinical progress of dengue diseases

Dengue is relatively diverse. The disease can occur suddenly in a normal person and progress from mild to severe quickly through 3 phases: febrile, critical, and recovery.

a. Febrile phase:

Clinical:
- Sudden, high fever
- Severe headache, anorexia, nausea
- Skin congestion
- Pain in muscle and joints, pain in the backside of the eyes
- Positive test for ligation
- Usually have spots of bleeding under the skin, teeth root bleeding or nosebleeds

Subclinical:
- Hematocrit is normal
- Normal or gradually decreased number of platelets (PLT)
- The number of white blood cells (WBC) usually decreases.

b. Critical phase: usually happen on the 3rd-7th day of the sickness

Clinical:
- The patient may have still or reduce fever
- Symptom of plasma drainage due to increased vascular permeability
- Pleural effusion, interstitial lung, peritoneal cavity can be painful, hepatomegaly
- If plasma drainage is high blood pressure will be stuck (the difference of maximum and minimum blood pressure ≤ 20 mmHg), hypotension or unmeasurable blood pressure, low urination
- Bleeding under the skin
- Mucosa bleeding
- Internal bleeding (digestive system, lungs, brain) is a severe symptom
- Some cautionary cases may show serious organ failures such as severe hepatitis, encephalitis, myocarditis.

Subclinical:
- Increasing HCT
- PLT < 100.000/mm³
- Ultrasound or X-ray can detect peritoneal fluid and pleural effusion.

c. Recovery phase:

Clinical:
- After 24-48 hours of the critical phase, there is a gradual re-absorption of fluid from interstitial tissue into the inner vein. This period lasts 48-72 hours
• Patients no longer have fever, overall condition is better, appetite, stable blood pressure and frequent urination
• May have low heart beats and electrocardiographic (ECG) changes
• During this time, excessive infusion can cause pulmonary edema or heart failure.

Subclinical:
• HCT returns to usual or may be lower due to blood thinners when the fluid is reabsorbed
• The number of WBC usually increases early after the fever goes down
• The number of PLT gradually returns to normal, later than the number of WBC.

2.4.3 Diagnosis

Dengue disease is divided into 3 levels (World Health Organization, 2009):
• Dengue
• Dengue with warning signs
• Severe dengue
  a. Dengue:

Clinical:
Suddenly high fever, continuously for 2-7 days and have at least 2 of following signs:
• Bleeding under the skin
• Headache, anorexia and nausea
• Muscle pain.

Subclinical:
• Normal or increased HCT
• Normal or gradually decreased PLT
• WBC usually decreases.

b. Dengue with warning signs:

Clinical: includes clinical symptoms of dengue, along with the following warning signs
• Lethargic, tired, fatigued
• Abdominal pain around liver area or liver pain
• Hepatomegaly > 2 cm
• Frequent nausea
• Mucosa bleeding
• Low urination.

Subclinical
• Increasing HCT
• Rapidly decreased PLT.

  c. Severe dengue:
• Severe plasma drainage leads to shock
• Severe bleeding
• Organ failures

3 Building medical diagnostic support system

3.1 Database model

After collecting data, we need to build the database to store necessary information for the controller unit as follows:

![Diagram showing database model](image)

3.1.1 Medical records in data warehouse

Digital medical records include patient’s personal information like gender, age, name…and treatment information.

a. Objectives:

Samples: all new patients, who are over 15 years old, are initially examined and possibly having dengue with variable levels.

Criteria of sample selection: Dengue; dengue with warning sign; severe dengue

The data set contains information of infected patients under therapy in the hospital, including all above criterias.
b. Data collection:

Collected data is stored in Excel files. Patient’s information includes: Cardinal number (Num), Name, age, gender (female: 0, male: 1) and BMI. BMI formula is calculated as follows:

\[
BMI = \frac{\text{Weight}(kg)}{\text{Height}^2(m)}
\]

<table>
<thead>
<tr>
<th>BMI</th>
<th>Physique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>18.5 - 22.9</td>
</tr>
<tr>
<td>Lightweight</td>
<td>&lt;18.5</td>
</tr>
<tr>
<td>Overweight</td>
<td>≥23</td>
</tr>
</tbody>
</table>

Clinical symptoms:
- Pulse: Slow, normal, fast, too fast
• Temperature: Medium, High
• Blood pressure: DBP, SBP
• Functional symptoms: headache, myalgia, abdominal pain, vomit, cough, diarrhea
• Others: Jaundice (yellowish skin), shock, hemorrhage.

Subclinical symptoms:
• Ultrasound: Liver is large or normal
• HCT: Normal, high, low
• WBC: Normal, high, low
• PLT: Normal, high, low
• Antigen NS1: Positive or negative
• IgM antibodies: Positive or negative
• IgM antibodies: Positive or negative

3.1.2 Preprocessing data

Collected data will be saved to an excel file and then imported into the database of datadengue.mdb. Data is entered on a tuple, including attributes Num, Age, Gender, SBP, DBP, HCT, WBC, PLT, Heart rate, HCT level, WBC level, PLT level, NS1, IGM, IGG, Myalgia, Headache, Hemorrhage, Abdominal pain, Vomit, Cough, Hepatomegaly, Jaundice, Diagnosis, DiagnosisID.

In order to implement the association rule model, we need to re-edit data and remove unnecessary attributes:
• Remove properties that data is missing or too much noisy
• Remove the attribute “Num”, “Gender” because these attributes are not used in the model. In the other words, this is attribute filtering
• Discrete data:

Discrete HCT data:

<table>
<thead>
<tr>
<th>ID</th>
<th>Meaning</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>High</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Low</td>
<td>0</td>
<td>35</td>
</tr>
</tbody>
</table>

Discrete PLT data:
Table 7. Discrete PLT data

<table>
<thead>
<tr>
<th>ID</th>
<th>Meaning</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal</td>
<td>140,000</td>
<td>400,000</td>
</tr>
<tr>
<td>2</td>
<td>High</td>
<td>400,000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Low</td>
<td>0</td>
<td>140,000</td>
</tr>
</tbody>
</table>

Discrete WBC data

Table 8. Discrete WBC data

<table>
<thead>
<tr>
<th>ID</th>
<th>Meaning</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal</td>
<td>5,000</td>
<td>10,000</td>
</tr>
<tr>
<td>2</td>
<td>High</td>
<td>10,000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Low</td>
<td>0</td>
<td>5,000</td>
</tr>
</tbody>
</table>

Table 9. Data table of attributes

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Types</th>
<th>Examples</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Num</td>
<td>Numeric</td>
<td>13.04436</td>
<td>Identify patients</td>
</tr>
<tr>
<td>2 Age</td>
<td>Numeric</td>
<td>36</td>
<td>Age of patients</td>
</tr>
<tr>
<td>3 Gender</td>
<td>Nominal</td>
<td>True, False</td>
<td>Patients’ gender</td>
</tr>
<tr>
<td>4 SBP</td>
<td>Numeric</td>
<td>12</td>
<td>Pressure of vessels when heart beats</td>
</tr>
<tr>
<td>5 DBP</td>
<td>Numeric</td>
<td>8</td>
<td>Pressure of vessels when heart rests between beats</td>
</tr>
<tr>
<td>6 HCT</td>
<td>Numeric</td>
<td>54.1</td>
<td>RBC capacity</td>
</tr>
<tr>
<td>7 WBC</td>
<td>Numeric</td>
<td>3.1</td>
<td>Leukocyte index</td>
</tr>
<tr>
<td>8 PLT</td>
<td>Numeric</td>
<td>89</td>
<td>Platelet index</td>
</tr>
<tr>
<td>9 Heartrate</td>
<td>Numeric</td>
<td>1, 2, 3</td>
<td>Pulse pressure</td>
</tr>
<tr>
<td>10 HCTlevel</td>
<td>Numeric</td>
<td>1, 2, 3</td>
<td>Normal, High, Low</td>
</tr>
<tr>
<td>11 WBClevel</td>
<td>Numeric</td>
<td>1, 2, 3</td>
<td>Normal, High, Low</td>
</tr>
<tr>
<td>12 PLTlevel</td>
<td>Numeric</td>
<td>1, 2, 3</td>
<td>Normal, High, Low</td>
</tr>
<tr>
<td>13 NS1</td>
<td>Nominal</td>
<td>True, False</td>
<td>Positive, Negative</td>
</tr>
<tr>
<td>14 IGM</td>
<td>Nominal</td>
<td>True, False</td>
<td>Positive, Negative</td>
</tr>
<tr>
<td>15 IGG</td>
<td>Nominal</td>
<td>True, False</td>
<td>Positive, Negative</td>
</tr>
<tr>
<td>16 Musclepain</td>
<td>Nominal</td>
<td>True, False</td>
<td>Positive, Negative</td>
</tr>
<tr>
<td>17 Headache</td>
<td>Nominal</td>
<td>True, False</td>
<td>Yes, No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>18</td>
<td>Hemorrhage</td>
<td>Nominal</td>
<td>True, False</td>
</tr>
<tr>
<td>19</td>
<td>Abdominalpain</td>
<td>Nominal</td>
<td>True, False</td>
</tr>
<tr>
<td>20</td>
<td>Vomit</td>
<td>Nominal</td>
<td>True, False</td>
</tr>
<tr>
<td>21</td>
<td>Cough</td>
<td>Nominal</td>
<td>True, False</td>
</tr>
<tr>
<td>22</td>
<td>Hepatomegaly</td>
<td>Nominal</td>
<td>True, False</td>
</tr>
<tr>
<td>23</td>
<td>Jaundice</td>
<td>Nominal</td>
<td>True, False</td>
</tr>
<tr>
<td>24</td>
<td>Diagnosis</td>
<td>Nominal</td>
<td>Dengue, Warning, Severe</td>
</tr>
<tr>
<td>25</td>
<td>Diagnosticclassification</td>
<td>Numeric</td>
<td>1, 2, 3, 4, 5, 6</td>
</tr>
</tbody>
</table>

After the pre-processing step, the data will be entered into the system as shown:
3.1.3 Analysis of digital medical records

Analyze digital medical records to find diagnostic rules provided with information about patients’ clinical and subclinical symptoms. The algorithms used to build the system are decision tree with C4.5 algorithm using gain entropy as the measurement to choose the best attribute.

The decision tree will be handled as follows:

a. The data sheet here is the dengue patient data set taken from medical records, handwritten data… The data will then be transferred into useful information for the system (datadengue.mdb data set).
b. Create function (T).
c. Function S is originally function (T)
The program will start working as follows:
d. If all data in S has the same class then it will end
   • If all data in S has the same class it will be all categorized as dengue.
   • End the program.
e. For each intermediate attribute in S data set, we would calculate the value of each
   • These are the attributes related to the diagnosis for patients such as: pulse, blood pressure, temperature or test indicators
   • We put those attributes’ indicators into the program to calculate the probability.
f. Select the attribute with the best measurement
   For instance: We have 4 attributes for the decision tree to build such as pulse pressure, IgG, IgM and NS1, we get
   \[
   \begin{align*}
   \text{RE1} & \rightarrow \text{pulse pressure} \\
   \text{RE1} & \rightarrow \text{IgM} \\
   \text{RE1} & \rightarrow \text{IgG} \\
   \text{RE2} & \rightarrow \text{NS1} \quad (\text{NS1 is selected to be the best attribute})
   \end{align*}
   \]
   NS1 has the best measurement, so we take NS1 as the root node with 2 values: True, False.
g. If the attribute has K means, it will turn K \rightarrow S, then use recursion to separate on that domain value.
   With the example above, the program will divide NS1 into 2 components (2 branches) to consider recursion until the final result is assigned the leaf node and we have a diagnostic subclass.

![Figure 8. NS1 is selected due to the best measurement](image)

Most of the learning machine systems try to create a tree as small as possible because the smaller the tree, the easier it is to achieve higher predictive accuracy.
Because it is impossible to ensure the minimum value of decision tree, C4.5 will be based on optimization research and the choice of division to have the measurement for selecting the attributes that reach maximum value.

Two measurements used in C4.5 are information gain or gain ratio.

### 3.1.4 Diagnostic rules

The diagnostic rules are built by the classification algorithm of decision tree in data mining. These rules are derived from the training data set, with each leaf node is a rule. After obtaining clinical and subclinical information, the doctor will rely on it to provide diagnostic results for patients.

- **Subclinical**
  
  Patients who are infected with dengue virus has the test of HCT increased by more than 20% of normal HCT values, decreased PLT, signs of haemorrhage, and being positive for NS1 and IgM…

- **Exclusion criteria**
  
  Patients with a history of heart, lung, liver, kidney diseases, etc. or have the pathology of dyslipidemia, coronary artery disease which is known in advance or just being detected during hospitalization.

### 3.1.5 Sample medical records

Sample medical records are patients' medical records that include information about patients' clinical, subclinical symptoms and diagnostic conclusions. Sample medical records are used to insert patients' clinical and subclinical information into the system. This information will then be analyzed based on the diagnostic rules of the system.

### 3.1.6 Diagnose

After inserting information from the sample medical record into the system, the program will reply on the rules drawn during the training process to categorize the sample medical record into the corresponding class of the system to produce the diagnostic result. This result will be compared with the diagnostic conclusions in the sample medical record.

### 3.2 Building the application

### 3.2.1 Introducing to the program

The program is designed and operated according to the requirements of the research topic. The technology part of the program is based on the strengths of data mining with
decision tree classification techniques and C4.5 algorithms to explore data on dengue and extract some diagnostic rules of dengue fever.

The program will create a decision tree, going from the root node to the leaf node of a certain branch of the tree, and let us know a diagnostic rule. After the decision tree is created, it will be automatically saved in the “*.xml” form to use for the later diagnoses without having to learn the entire data to create the tree again. This will shorten the diagnosis time. The program only re-learns when there is any change or updates.

The program is built in C# and Net Framework 4.5 languages.

3.2.2 How to operate the program

Step 1: the program is started from the exe file

![Diagnostic support system](image)

Figure 9. Starting screen

Step 2: Select the link to the digital medical data storage by clicking the “Load Data” Button and select the data file to put into the program.
Select file datadengue.mdb, click “Open”

Step 3: Select the data table. After choosing the data table, all the data will be inserted into the system.

Step 4: Select the attributes to build the decision tree

Select the attributes such as: Pulse, IgM, IgG, NS1, …

Pulse pressure is the difference between diastolic blood pressure and systolic blood pressure. For example: diastolic blood pressure is 11, systolic blood pressure 7, hence, the pulse pressure is 4. When selecting the attributes, it is essential to pay attention to choose the type of data accordingly (table 8).

Step 5: Choose the % training for the machine to learn and click “Set” button.

Step 6: Choose the tab “Decision tree”, click “Create” button to proceed building the decision tree from the data selected in the steps 4 and 5. The data that was downloaded from
the digital medical storage and the configuration of the attributes are selected to build the decision tree.

Click “Create” button, the program will proceed the classification algorithms to create the decision tree as follows
After building the tree, the program will automatically save the tree in the form of XML file so that it can be uploaded on the program for later use. This will reduce the learning time for the machine and the machine only learns when there is any data update. Double clicking on any leaf node of the tree, the system will provide the rule corresponding to that leaf node.

Click on the “Select” button to get a previously created decision tree (saved as *.xml) to classify for new attributes.

Click on the tick mark button to view the rule drawn after training.

Enter the number of rules that need to be listed in the “Number of rules” box and click “Find” button. For example, we need to find 10 rules, the program will list 10 rules with the greatest number of occurrences in the rule set and list them on the screen.
Step 7: Select the “Testing result” tab to conduct the evaluation of the newly built model. In order to test the results, we must have the test set. This set can be imported from excel file or from the untrained original data. The training and testing processes have the following model:
If using 70% of training set, the test set is 30%. Then fill the box “30%” and click on the “Data test” button in the “Testing result” tab to conduct the test. The test results tell us the percentage corresponding to the model that has just been created. Thereby, we are able to evaluate the newly built tree model.

Click the “Data test” button to conduct the test.
In addition, it is possible to get the test data set from excel file to test the model. The test data sets are also preprocessed to have the same structure as the training data set. The results also tell us the percentage corresponding to the current model of system.

Steps to get excel file to test as follows:

a. Click “Access data” and select the excel file that needs to be checked.

b. Click “Test data” button.

The program will tell us:

- The total number of tested data.
- The number of right results (%).

In case of building a model and cross-checking with cross validation, we also perform from steps 1 to 5, then select the tab “Cross validation”. The screen of cross validation will be as follows:

![Figure 14. Data test result](image-url)
Enter "K" number into the "Folds" box.
Enter the percentage of learning data into "Training data" box.
Click the "Train" button to proceed with cross validation. The model with the highest test result will be selected.

The result of 82.82% is the best rate so it will be chosen as the model in this training.

Step 8: Select tab “Diagnosis”. A list of diagnostic information (attributes) is already pre-established corresponding to the attributes selected in step 4 (this information is pre-established and saved as data.xml, this file contains the attributes of a sample medical records). Enter the clinical and subclinical results that need to be checked or the sample medical record in the “Value” column corresponding to the “Attribute” column, the program will base on the model of the newly built decision tree to classify (diagnose) the data that needs to be checked. The design of patients’ sample information with XML file is flexible. If the structure of digital medical record changes later, we just need to modify the XML file accordingly so that the program can operate as normal without having to re-edit the source code of the program.
The diagnosis screen has 2 parts: “Attribute” and “Value”. For instance, enter the values for the attributes:

- Pulse pressure: 2, 3, 4, 5.
- Symptoms: Headache, Muscle pain, Abdominal pain, Vomit. Choose True or False.

After entering the values, click “Diagnose” button to classify the attributes that have just been entered.

In case we do not create a decision tree but want to use the previously created tree, then click the “Test Tree” button so that the attributes (new medical records) we want to check will be detected in the previous trees and give the classification results. The machine accomplishes this task relatively quickly. The diagnostic result of this case will be shown as follows:
After diagnosing, the program will tell us the classification result and which model we use (decision tree has been saved).

Step 9: After the system has given the classification results (diagnosis), click the button “Treatment” to open the treatment regimen corresponding to the diagnosis of the system (the available treatment regimen files correspond to the diagnosis and are saved in a word file format).

For example, we have the treatment regimen of dengue (Thu vien phap luat, 2011). Most cases are treated with outpatient care and followed-up at health centers, mainly are symptomatic treatment and close monitoring to have early detection of shocks for timely management.

a. Symptomatic treatment
   • If high fever ≥ 39°C, give patients antipyretic medicine, loosen their clothes and cool them with warm water.
   • With antipyretic medicines, we only use monotherapy Paracetamol, dose 10-15mg/kg of weight/time, take every 4-6 hours.
   • Attention:
     Total dose does not exceed 60mg/kg of weight/24 hours.
     Do not take aspirin (acetyl salicylic acid), analgin, ibuprofen for treatment because it can cause hemorrhage and acidosis.

b. Oral rehydration therapy:
• Encourage patients to drink plenty of Oresol or water, juice (coconut, orange, lemon, etc.) or dilute porridge with salt.

Table 10. Table of classification statistics and diagnostic time support (the results of conducting the application).

<table>
<thead>
<tr>
<th>Number of attributes</th>
<th>Classification time</th>
<th>Number of rules</th>
<th>Results</th>
<th>Diagnostic support time</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>3 minutes 14 seconds</td>
<td>341</td>
<td>98%</td>
<td>&lt; 1 second</td>
</tr>
<tr>
<td>10</td>
<td>1 minute 13 seconds</td>
<td>276</td>
<td>95%</td>
<td>&lt; 1 second</td>
</tr>
<tr>
<td>8</td>
<td>39 seconds</td>
<td>226</td>
<td>93%</td>
<td>&lt; 1 second</td>
</tr>
</tbody>
</table>

The statistic table clearly shows that if the patient provides the medical staffs with sufficient clinical symptoms and the medical staff performs the necessary subclinical tests for the patient, the diagnosis result will be highly effective (in case the number of attributes is 16 and the result is 98%. If the attributes are 10, the result will be reduced to 95%). Thereby, we find it is logical to reality.

4 Test and evaluation

The thesis will conduct tests in order to verify and evaluate the testing methods as well as the collected actual results.

The evaluation of the effectiveness of the program is entirely based on the collected information during the test. This evaluation is based on a number of criteria:

• The usability of the application.
• The accuracy of collected information.
• Shortened time of diagnosis.

4.1 Test

4.1.1 Test the data set with few attributes

With some clinical and sub clinical attributes (pulse pressure; platelet level, hematocrit level, white blood cell level, haemorrhage, NS1, IgM) in the data set, some rules are explored as follows:
• If PLTlevel = “Low” ∧ Pulsepressure = 5 ∧ IgM = “Yes” ∧ Hemorrhage = “No” ∧ HCTlevel = “Normal” ∧ WBClevel = “Low” ∧ NS1 = “No” → Belong to subclass “Dengue”.
• If PLTlevel = “Low” ∧ Pulsepressure = 6 ∧ IgM = “No” ∧ NS1 = “Yes” ∧ Hemorrhage = “No” ∧ WBClevel = “Low” ∧ IgM = “Yes” → Belong to subclass “Dengue”.
• If PLTlevel = “Low” ∧ Pulsepressure = 4 ∧ Hemorrhage = “Yes” ∧ IgM = “No” ∧ WBClevel = “Low” ∧ HCTlevel = “Normal” ∧ NS1 = “Yes” → Belong to subclass “Dengue”.
• If PLTlevel = “Low” ∧ Pulsepressure = 3 ∧ HCTlevel = “Low” ∧ Hemorrhage = “No” ∧ NS1 = “Yes” ∧ WBClevel = “Normal” ∧ IgM = “Yes” → Belong to subclass “Dengue”.
• If PLTlevel = “Low” ∧ Pulsepressure = 3 ∧ HCTlevel = “High” ∧ IgM = “Yes” ∧ Hemorrhage = “No” ∧ WBClevel = “Normal” ∧ NS1 = “Yes” → Belong to subclass “Severe dengue”.
• If PLTlevel = “Low” ∧ Pulsepressure = 6 ∧ IgM = “Yes” ∧ HCTlevel = “High” ∧ WBClevel = “Normal” ∧ NS1 = “Yes” ∧ Hemorrhage = “No” → Belong to subclass “Dengue with warning sign”.
• If PLTlevel = “Low” ∧ Pulsepressure = 3 ∧ HCTlevel = “Normal” ∧ Hemorrhage = “Yes” ∧ IgM = “Yes” ∧ Cough = “Yes” ∧ WBClevel = “Normal” ∧ Pulsepressure = 5 → Belong to subclass “Dengue with warning sign”.
• If PLTlevel = “Low” ∧ Jaundice = “Yes” ∧ Hemorrhage = “Yes” ∧ IgM = “Yes” ∧ Cough = “Yes” ∧ WBClevel = “Normal” ∧ Pulsepressure = 7 ∧ HCTlevel = “Normal” → Belong to subclass “Dengue with warning sign”.
• If PLTlevel = “Low” ∧ Jaundice = “Yes” ∧ Hemorrhage = “Yes” ∧ IgM = “No” ∧ Cough = “Yes” ∧ WBClevel = “High” ∧ NS1 = “Yes” → Belong to subclass “Dengue with warning sign”.

4.1.2 Test the data set with full attributes

With full clinical and subclinical attributes in the data set, some rules are explored as follows:
• If PLTlevel = “Low” ∧ Pulsepressure = 3 ∧ HCTlevel = “Normal” ∧ Hemorrhage = “Yes” ∧ WBClevel = “Normal” → Belong to “Dengue”.
• If PLTlevel = “Low” ∧ Jaundice = “Yes” ∧ Hemorrhage = “Yes” ∧ IgM = “Yes” ∧ Cough = “Yes” ∧ WBClevel = “Normal” ∧ Pulsepressure = 5 → Belong to subclass “Dengue with warning sign”.
• If PLTlevel = “Low” ∧ Jaundice = “Yes” ∧ Hemorrhage = “Yes” ∧ IgM = “Yes” ∧ Cough = “Yes” ∧ WBClevel = “Normal” ∧ Pulsepressure = 7 ∧ HCTlevel = “Normal” → Belong to subclass “Dengue”.
• If PLTlevel = “Low” ∧ Jaundice = “Yes” ∧ Hemorrhage = “Yes” ∧ IgM = “Yes” ∧ Cough = “Yes” ∧ WBClevel = “High” → Belong to subclass “Dengue”.
• If PLTlevel = “Low” ∧ Jaundice = “Yes” ∧ Hemorrhage = “Yes” ∧ IgM = “Yes” ∧ Cough = “No” ∧ HCTlevel = “Low” ∧ IgG = “Yes” ∧ Pulsepressure = 4 ∧ WBClevel =
"Normal" ^ NS1 = "Yes" ^ Musclepain = "No" ^ Headache = "Yes" ^ Abdominalpain = "Yes" ^ Vomit = "No" ^ Hepatomegaly = "Yes" ^ Diarrhea = "No" → Belong to subclass “Dengue”.

- If PLTlevel = “Low” ^ Jaundice = “Yes” ^ Hemorrhage = “Yes” ^ IgM = “No” ^ WBClevel = “Low” ^ Abdominalpain = “No” ^ HCTlevel = “Normal” ^ Pulsepressure = 5 ^ Cough = “No” ^ NS1 = “Yes” ^ Musclepain = "No" ^ IgG = “No” ^ Headache = “Yes” ^ Vomit = “No” ^ Hepatomegaly = “Yes” ^ Diarrhea = “No” → Belong to subclass “Dengue with warning sign”.

Draw a sample decision tree:

Draw a subclass “Dengue with warning sign” with the attributes: If PLTlevel = “Low” ^ Jaundice = “Yes” ^ Hemorrhage = “Yes” ^ IgM = “Yes” ^ Cough = “Yes” ^ WBClevel = “Normal” ^ Pulsepressure = 5 → Belong to the subclass “Dengue with warning sign”.

Draw a subclass “Dengue” with the attributes: If PLTlevel = “Low” ^ Hepatomegaly = “Yes” ^ Jaundice = “No” ^ NS1 = “Yes” ^ Pulsepressure = 3 ^ Vomit = “No” ^ Abdominalpain = "Yes" → Belong to the subclass “Dengue”.

Figure 19. Image of dengue with warning sign
Draw a subclass “Severe dengue” with the attributes: If PLT level = “Low” \(\land\) Hepatomegaly = “Yes” \(\land\) Jaundice = “No” \(\land\) NS1 = “Yes” \(\land\) Pulsepressure = 2 \(\rightarrow\) Belong to the subclass “Severe dengue”.

Figure 20. Image of dengue
4.2 Evaluation

The ultimate objective of this evaluation is to assess the strengths and weaknesses of the prototyping model and the impact of its deliverables. After presenting the features of the application and deploying the test, the system was highly appreciated by specialist doctors. However, there are still errors, unexpected and incorrect results.

5 Conclusion

5.1 Conclusion

This thesis provides a way to look at the integration of information technology into medical field. The result of the thesis is a community based diagnostic support system, which will
support a lot in terms of expertise for the areas without a team of doctors with high professional qualifications, lack of medical equipment as well as remote heath routes. In addition, for medical students and young doctors, the diagnostic support system also helps them review the knowledge of infectious diseases in tropical area in general and the patients with dengue will know the level of disease for precise treatment direction in particular.

The results of the research are still not really good. The diagnosis results of the diagnostic support system still have many cases that are not quite correct. However, after demonstrating the test for specialists at Pham Ngoc Thach hospital, the results were highly appreciated. According to these doctors, if investment and effort are continued and developed, the diagnostic support system will be very helpful for doctors in diagnosing and treating patients.

Besides, the thesis can even be developed for faster and more accurate diagnosis if the diagnostic support system is directly connected to the information systems of medical management at the medical facility. Moreover, the diagnostic support system can be applied to many different diseases.

5.2 Limitations

In terms of technology, the using of C4.5 algorithm has some limitations on data processing. In case the data has too many layers, the algorithm will easily cause errors. The more the data the longer the training time.

In term of practical issues, the data of this thesis was collected at a random time, hence, it was impossible to observe the patients’ progress. In order to achieve better results, we need to collect data from when the patient has the initial symptoms until the disease breaks out and cured. At the same time, it is necessary to get more data of the patients’ treatment process, which will give us a better understanding of the disease process and help the program achieve high practical results.

5.3 Development direction

Data mining is a topic that many researchers are interested in because it is widely applied in many fields as well as contains many different directions of expansion. However, in order to expand the application and put it into practice, we need to do some more tasks:

- Data of patients must be collected more.
- Collect data in terms of clinical and subclinical progress.
- Collect progress of treatment regimen.
- Better data processing to increase program performance.
- Build a disease diagnosis system for many different diseases.
• Need professional cooperation from both information technology and medical specialists/professionals.
• Learn other algorithms such as C5.0, Bayesian network and Neuron network if they are more effective.
• Learn the ILA algorithm to compare the results achieved with C4.5 algorithm.
References


