Detecting Hand, Foot and Mouth Disease in Earlier Stage Using C4.5 Algorithm as Expert System Based on Android

Farisa Hafida Syahrial¹, Budhi Irawan², and Anggunmeka Luhur Prasasti³
¹College Student, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia.
²,³Lecturer, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia.

Abstract: Hand, Foot and Mouth Disease (HFMD) is an infectious diseases caused by enterovirus virus 71 (EV 71). The symptoms of HFMD is similar to several other disease that caused by a virus, especially disease that have a fever and rash symptoms which people usually underestimate diseases that have early symptoms like that. Therefore, in this system we classify the HFMD with the intention of detecting the disease from an earlier stage. And we use Android based application since at this present time, smartphone is the closest device that is always used by many people. The classification used in this paper is Decision Tree C4.5 Algorithm. Dataset used in this research is as many as 256 which divided into training data and testing data, that formed based on symptoms that had previously been validated by the doctor. The result shows that data partitions of 90%:10%, 80%:20% and 70%:30% has accuracy, precision and recall value are 100%. Thus, data partition 70%:30% has the best result because this partition has less training data but can still classify diseases effectively.

1. Introduction

Hand, Foot and Mouth Disease (HFMD) is a disease that usually attacks infants and children under the age of 10 years. This disease has been prevalent throughout the world. In tropical and subtropical countries, this epidemic often occurs throughout the year. China is one of the countries with the most cases of this disease, which is around 500,000-1,000,000 cases [1]. This disease is included in infectious diseases. The cause of this disease is the group of enterovirus A viruses. Although classified as a mild disease, HMFD can also cause complications that lead to death if caused by the 71 enterovirus virus (EV 71) [2]. The initial symptoms of the disease are generally fever, mouth sores and the appearance of vesicles or rashes accompanied by blisters on the palms and soles of the feet. Although this disease is usually suffered by children under the age of 10 years, since it is an infectious disease, this disease can also be suffered by adults who do not have a strong immune system [3]. The symptoms of HFMD is similar to several other disease that caused by a virus, especially disease that have a fever and rash symptoms which people usually underestimate diseases that have early symptoms like that. Therefore, in this system we classify the HFMD with the intention of detecting the disease from an earlier stage. The purpose of this system is not to replace the doctor but to help people know how much they’re exposed to HFMD before consulting a doctor. And we use Android based application since at this present time, smartphone is the closest device that is always used by many people. This system detects HFMD based on rash conditions on the palms and symptoms experienced by the user. The system detects rashes on the palm of the hand using image processing with the Discrete Wavelet Transforms method[4].
The classification used in this paper is Decision Tree C4.5 Algorithm. This method is often used in classifying diseases. This decision tree method is proven to be able to classify diseases with good accuracy and performance. A research on C4.5 algorithm has been done before by various people and with various diseases data set such as thrombosis collagen diseases, diabetes, breast cancer, liver disease, etc. Based on a survey conducted by B. Padmapriya and T. Velmurgan, C4.5 algorithm has better accuracy and performance in classifying breast cancer than other algorithms [5]. Another survey on C4.5 has also been done by D. Sindhuja and R. Jemina Priyadarsini. They were doing a survey on classification techniques to analyze liver disease disorder by comparing various classification techniques. And it was found that C4.5 gave better results in diagnosing the disease at earlier stage [6]. Furthermore, another research used diabetes dataset to comparing C4.5 and KNN to predict the diabetes disease. This research has been done by Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan. In this research the results show that the C4.5 algorithm has better performance than KNN [7]. C4.5 algorithm also used in a research that has been done by Sarah A. Soliman, Safia Abbas and Abdel-Badeeh M. Salem. The research implements C4.5 to diagnose the thrombosis [8].

2. Materials and Method

This section will explain about proposed system design, the data set, decision tree C4.5 algorithms and accuracy measures.

Proposed System Design

The system divided into two process. The first one is train the dataset using C4.5 Algorithm by building a decision tree. After getting the decision tree rules from the training process, the system can start testing process by following decision tree rules to detect HFMD. The output of this system is information on whether or not HFMD is detected. The output is derived from the user's answer to the question about the symptoms they have experienced. For rash detection seen in the user’s palm, the results obtained from a different system, where the results will be entered into the user's answer list.

Hand, Foot and Mouth Disease Dataset Description

The data set used is formed based on symptoms of HFMD that had previously been validated by the doctor. The questions that will be asked to the user is designed based on these symptoms, which are shown in Table 1.
Table 1. Symptoms of HFMD and Questions for The System

| Symptoms                          | Questions                                                                 |
|-----------------------------------|---------------------------------------------------------------------------|
| Fever                             | Before the rash appears, have you ever experienced a fever?                |
| Fever for 3 days                  | Have you had the fever for more than 3 days?                              |
| Rashes                            | (not included to the interview because the result is from a different system) |
| Rashes fill with liquid           | Does the rash on your palm contain white liquid?                          |
| Red rashes                        | Is the rash in your palm red?                                             |
| Rashes also appears in Mouth or Feet | Does the rash that you have on your palm also found in the mouth or feet? |
| Sore rashes                       | Does your rash hurt?                                                     |
| Itchy rashes                      | Does the rash feel itchy?                                                 |

The data set contains 256 instances, 8 attributes and a target attribute with two possible values. All of the attribute are discrete valued, which are shown in Table 2.

Table 2. Description of The Attributes in The HFMD Dataset

| Attribute           | Value                |
|---------------------|----------------------|
| Fever               | Yes, No              |
| Three days          | Yes, No              |
| Rashes              | Yes, No              |
| Liquid              | Yes, No              |
| Red                 | Yes, No              |
| Mouth/Feet          | Yes, No              |
| Sore                | Yes, No              |
| Itchy               | Yes, No              |
| Class               | Detected, Not detected |

Decision Tree C4.5 Algorithms

Decision tree is the most commonly used method for classification because it has a simple structure to understand. Decision tree consists of nodes connected by edge to form rooted trees. The decision tree starts with a root node that does not have an incoming edge. Other nodes have one edge entered. Nodes that have an edge that comes out are called branches, while nodes that do not have an edge are called leaf or decision nodes. In a decision tree, each branch divides the edge into two or more sub-spaces according to the attribute values that are in the branch. Each leaf refers to a class that represents the most suitable to target value.

The C4.5 algorithm was developed in 1993 by Ross Quinlan to overcome the limitations of the previous algorithm, ID3 algorithm [9]. This algorithm can handle attributes with continuous or discrete values. This algorithm can also handle attributes whose values are unknown. The C4.5 algorithm, choose a root based on the attribute that has the highest Gain Ratio value.
Fig 2. Flowchart of Decision Tree C4.5 Algorithm

The first thing to do when building a decision tree with the C4.5 algorithm is to count total cases for each attribute, Entropy of all the attribute, Information Gain, Splitting Info and Gain Ratio [10].

\[
\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \text{Entropy}(S_i)
\]  

(1)

Where, Entropy(S) is:

\[
\text{Entropy}(S) = \sum_{i=1}^{n} p_i \log_2 p_i
\]  

(2)

And Gain Ratio is:

\[
\text{Gain Ratio}(S,A) = \frac{\text{Gain}(S,A)}{\text{SplitInfo}(S,A)}
\]  

(3)

Where SplitInfo(S,A) is:

\[
\text{SplitInfo}(S,A) = \sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S}
\]  

(4)

Where:

- \( S \): Set of cases
- \( A \): Attribute
- \( n \): number of partition attributes A
- \( |S_i| \): number of cases on the i partition
- \( |S| \): number of cases in S
- \( p_i \): proportion of \( S_i \) to S
The attribute that has the highest Gain Ratio will become the root of the decision tree. And the values of the attribute will be the branch. If one or more value of the attribute is 0, then the branch that has the value automatically becoming a leaf that refers to one of the value on the target attribute. The other value that does not equal 0, is going to be the next node of the decision tree. And this step is done until all the branches have a leaf or decision nodes[11].

### Accuracy Measures

Confusion matrix is needed to determine classification performance. In this classification measurements were made on Accuracy, Precision, Recall to determine classification performance to classify the data used in this experiment. Precision is used to see how many output of the system classified as "Detected as HFMD" is correctly classified. Accuracy is used to see the exact classification results that are classified correctly by the Decision Tree C4.5 algorithm. Recall is used to see how many classification result in dataset that indicates as "Detected as HFMD", that are correctly classified. Confusion matrix for the dataset is presented in Table 3.

| Actual | Prediction | Not Identified as HFMD | Identified as HFMD |
|--------|------------|------------------------|-------------------|
| Not HFMD | A | B |
| HFMD | C | D |

The following is a formula for calculating Accuracy, Precision, Recall:

\[
\text{Accuracy} = \frac{(A+D)}{(A+B+C+D)} \times 100\%
\]  \hspace{1cm} (5)

\[
\text{Precision} = \frac{D}{(B+D)} \times 100\%
\]  \hspace{1cm} (6)

\[
\text{Recall} = \frac{D}{(C+D)} \times 100\%
\]  \hspace{1cm} (7)

3. **Result and Discussion**

The following are results of calculating total cases, entropy, gain, splitinfo and gain ratio after applying the C4.5 Algorithm to the training data with 205 training data:

| Attributes | Total Cases | Undetected | Detected |
|------------|-------------|------------|----------|
| Total      | 205         | 187        | 18       |
| Fever      |             |            |          |
| Yes        | 107         | 89         | 18       |
| No         | 98          | 98         | 0        |
| Three days |             |            |          |
| Yes        | 106         | 97         | 9        |
| No         | 99          | 90         | 9        |
| Rashes     |             |            |          |
| Yes        | 77          | 59         | 18       |
| No         | 128         | 128        | 0        |
| Liquid     |             |            |          |
| Yes        | 104         | 92         | 12       |
| No         | 101         | 95         | 6        |
| Red        |             |            |          |
| Yes        | 106         | 94         | 12       |
| No         | 99          | 93         | 6        |
| Mouth Feet |             |            |          |
| Yes        | 105         | 87         | 18       |
| No         | 100         | 100        | 0        |
| Sore       |             |            |          |
| Yes        | 107         | 95         | 12       |
| No         | 98          | 92         | 6        |
| Itchy      |             |            |          |
| Yes        | 110         | 98         | 12       |
| No         | 95          | 89         | 6        |
Where, Entropy Total is 0.4291, used for calculating the Gain for each attribute.

Table 5. Calculate Entropy, Gain, SplitInfo and Gain Ratio for Each Attribute

| Attributes  | Entropy (Yes) | Entropy (No) | Gain       | Split Info | Gain Ratio |
|-------------|---------------|--------------|------------|------------|------------|
| Fever       | 0.653627      | 0            | 0.08793885 | 0.998609   | 0.088061   |
| Threedays   | 0.419233      | 0.439497     | 0.00008101 | 0.999159   | 0.0000811 |
| Rashes      | 0.784519      | 0            | 0.134426989| 0.954882   | 0.140779   |
| Liquid      | 0.515947      | 0.325082     | 0.007188896| 0.999846   | 0.00719    |
| Red         | 0.509516      | 0.329846     | 0.006351606| 0.999159   | 0.006357   |
| MouthFeet   | 0.660962      | 0            | 0.090558347| 0.999571   | 0.090597   |
| Sore        | 0.506366      | 0.332287     | 0.005952594| 0.998609   | 0.005961   |
| Itchy       | 0.497168      | 0.339855     | 0.004833269| 0.996134   | 0.004852   |

As seen, the Rashes attributes have the highest Gain Ratio. In this kind of case, the attribute that will be selected as root is Rashes. After all of the gain ratio values for each attributes is calculated, and every branches have a leaf or decision nodes, the next step is to built the decision tree. The decision tree is shown in Fig 3.

Fig 3. Decision Tree

Based on the decision tree, a decision tree rules can be formed which will be used to classify the answers entered by user.

Table 6. Case 1

| Attributes | User’s Answer |
|------------|---------------|
| Fever      | Yes           |
| Threedays  | Yes           |
| Rashes     | Yes           |
| Liquid     | Yes           |
| Red        | Yes           |
| MouthFeet  | Yes           |
| Sore       | Yes           |
| Itchy      | Yes           |

Table 7. Case 2

| Attributes | User’s Answer |
|------------|---------------|
| Fever      | Yes           |
| Threedays  | Yes           |
| Rashes     | Yes           |
| Liquid     | No            |
| Red        | Yes           |
| MouthFeet  | No            |
| Sore       | No            |
| Itchy      | Yes           |

The results of the user's answers are shown in Table 6. After the answers are classified using a decision tree rules, the results is as shown in Fig 4 which is the user is detected HFMD. The results of the user's answers
are shown in Table 7. After the answers are classified using a decision tree rules, the results is as shown in Fig 5 which is the user is not detected HFMD.

The Analysis of Decision Tree Result

The dataset is divided into 4 partitions, which is 90% training data : 10% test data, 80% training data : 20% test data, 70% training data : 30% test data, 60% training data : 40% test data.
Decision tree shown in Fig 8 is a decision tree produced by data partition of 70%:30%. The decision tree produces the following rules:

- IF (Rashes=Yes) & (Fever=Yes) & (MouthFeet=Yes) & (Liquid=No) & (Red=Yes) THEN decision= “Detected”
- IF (Rashes=Yes) & (Fever=Yes) & (MouthFeet=Yes) & (Liquid=No) & (Red=No) THEN decision= “Not Detected”
- IF (Rashes=Yes) & (Fever=Yes) & (MouthFeet=Yes) & (Liquid=Yes) THEN decision= “Detected”
- IF (Rashes=Yes) & (Fever=Yes) & (MouthFeet=No) THEN decision= “Not Detected”
- IF (Rashes=Yes) & (Fever=No) THEN decision= “Not Detected”
- IF (Rashes=No) THEN decision= “Not Detected”

The decision tree generated by data partition of 70%:30% seen in Fig 8. The amount of branches produced by data partition of 70%:30% has the same amount as the data partition of 80%:20%. The difference is in the order of branches. At data partition of 80%:20%, the branch after Fever is Liquid, while at data partition of 70%:30%, the next branch is Red. The rules for Fig 8 are:

- IF (Rashes=Yes) & (MouthFeet=Yes) & (Fever=Yes) & (Red=No) & (Liquid=Yes) THEN decision= “Detected”
The decision tree generated by data partition of 60%:40% shown in Fig. 9. The decision tree has fewer branches than the decision tree built by other partition data. The rules produced are also slightly more different, namely as follows:

- IF (Rashes=Yes) & (MouthFeet=Yes) & (Fever=Yes) & (Red=No) & (Liquid=No) THEN decision= “Not Detected”
- IF (Rashes=Yes) & (MouthFeet=Yes) & (Fever=Yes) & (Red=Yes) THEN decision= “Detected”
- IF (Rashes=Yes) & (MouthFeet=No) & (Fever=Yes) THEN decision= “Not Detected”
- IF (Rashes=Yes) & (MouthFeet=No) THEN decision= “Not Detected”
- IF (Rashes=No) THEN decision= “Not Detected”

Based on rules formed from the decision tree on each partition data, the testing process is carried out. The testing process produces Precision, Recall, and Accuracy as shown in Table 8. The Accuracy produced by data partition of 90%:10%, 80%:20%, 70%:30% has the same result that is equal to 100%, so are Precision and Recall. While for data partition of 60%:40%, the results are 75.76% Accuracy, 27.27% Precision, 100% Recall.

Table 8. Evaluation Table

| Data Partition | Precision | Recall | Accuracy |
|----------------|-----------|--------|----------|
| 90% : 10%      | 100%      | 100%   | 100%     |
| 80% : 20%      | 100%      | 100%   | 100%     |
| 70% : 30%      | 100%      | 100%   | 100%     |
| 60% : 40%      | 27.27%    | 100%   | 75.76%   |

4. Conclusion

From the test results, data partition of 90%:10%, 80%:20%, and 70%:30% have Accuracy, Recall, and Precision with a value of 100%. While for data partition of 60%:40% has a smaller Accuracy than other data partition, which is 75.76%. Data partition of 60%:40% also has a very small Precision value of 27.27%, but Recall value obtained by data partition of 60%:40% is 100%. The data partitions of 60%:40% has the smallest value of Precision due to many outputs from the system that indicate the user is detected as HFMD, but should not be detected as HFMD in the actual dataset. These results were obtained because the data partition of 60%:40% has fewer branches than the others. This causes the trees produced to have fewer rules.

Data partition 70%:30% has the best result because this partition has less training data but can still classify diseases effectively. It can be concluded that the C4.5 algorithm can detect HFMD in earlier stage based on the symptoms found in this system. It can also be concluded that, the performance of this algorithm is determined by the quality of training data and the form of the decision tree produced by the training data.

5. References

[1] J. F. Wang et al., “Hand, foot and mouth disease: Spatiotemporal transmission and climate,” Int. J. Health Geogr., vol. 10, pp. 1–10, 2011.
[2] World Health Organization, A Guide to Clinical Management and Public Health Response for Hand, Foot and Mouth Disease (HFMD). 2011.
[3] K. Kaminska, G. Martinetti, R. Lucchini, G. Kaya, and C. Mainetti, “Coxsackievirus A6 and hand, foot
and mouth disease: Three case reports of familial child-to-immunocompetent adult transmission and a literature review,” *Case Rep. Dermatol.*, vol. 5, no. 2, pp. 203–209, 2013.

[4] M. N. Aziz, T. W. Purboyo, and A. L. Prasasti, “A Survey on the Implementation of Image Enhancement,” *Int. J. Appl. Eng. Res.*, vol. 12, no. 21, pp. 11451–11459, 2017.

[5] B. Padmapriya and T. Velmurugan, “A survey on breast cancer analysis using data mining techniques,” *2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014*, 2015.

[6] D. Sindhuja and R. J. Priyadarsini, “A survey on classification techniques in data mining for analyzing liver disease disorder,” *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 5, pp. 483–488, 2016.

[7] E. K. Hashi, M. S. Uz Zaman, and M. R. Hasan, “An expert clinical decision support system to predict disease using classification techniques,” *ECCE 2017 - Int. Conf. Electr. Comput. Commun. Eng.*, pp. 396–400, 2017.

[8] S. A. Soliman, S. Abbas, and A. B. M. Salem, “Classification of thrombosis collagen diseases based on C4.5 algorithm,” *2015 IEEE 7th Int. Conf. Intell. Comput. Inf. Syst. ICICIS 2015*, pp. 131–136, 2016.

[9] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, 1993.

[10] R. K. Amin, Indwiarti, and Y. Sibaroni, “Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region),” *2015 3rd Int. Conf. Inf. Commun. Technol.*, pp. 75–80, 2015.

[11] Ivana Herliana W and S. Hansun, “Rancang Bangun Sistem Pakar untuk Deteksi Dini Katarak Menggunakan Algoritma C4 . 5,” *Ultim. Comput.*, vol. VII, no. Desember, pp. 48–58, 2015.

**Acknowledgments**

We thank dr. Mahmud Muhyiddin from Al-Islam Hospital Bandung for being an expert on this work which has validated the dataset based on the symptoms of HFMD so that it can be used as a dataset.