PREDICTION OF MALARIA INCIDENCE IN BANGGAI REGENCY USING EVOLVING NEURAL NETWORK

Rita Rismala[1], Prof. The Houw Liong[2], Arie Ardiyanti Suryani[3]

[1]Magister Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Malaria merupakan penyakit endemis di sebagian besar wilayah Indonesia, terutama di daerah pedesaan dan terpencil. Banggai, salah satu kabupaten di provinsi Sulawesi Tengah, merupakan daerah endemis tinggi malaria, dengan Annual Parasite Incidence (API) pada tahun 2010 mencapai 7.88‰. Kejadian dan penyebaran malaria sangat dipengaruhi oleh sejumlah faktor lingkungan dan cuaca, terutama curah hujan dan suhu. Oleh karena itu pada studi ini dibangun suatu sistem prediksi kejadian malaria yang dikaitkan dengan faktor lingkungan dan cuaca agar bisa membantu Kementerian Kesehatan dalam pengendalian malaria. Adapun metode yang digunakan adalah Evolving Neural Network (ENN). Metode ini menggabungkan Jaringan Syaraf Tiruan (JST) dan Algoritma Genetika (AG).

Sistem prediksi yang dihasilkan dari studi ini menghasilkan performansi yang cukup bagus untuk memprediksi kejadian malaria berdasarkan faktor cuaca. Performansi terbaik saat memprediksi kejadian malaria pada tahun 2008 adalah MAPE 21.3%, akurasi 75%, dan F-value 84.21%, dimana sistem menghasilkan performansi terbaik saat melakukan prediksi di musim kemarau dengan MAPE 13.18%, akurasi 100%, dan F-value 100%. Sedangkan untuk memprediksi kejadian malaria pada tahun 2009, dihasilkan MAPE 15.29%, akurasi 75%, dan F-value 40%, dimana sistem menghasilkan performansi terbaik saat melakukan prediksi di musim hujan dengan MAPE 3.1%, akurasi 100%, dan F-value 100%. Hasil ini membuktikan bahwa ada korelasi yang cukup antara cuaca dan kejadian malaria.

ENN mengurangi proses trial-and-error dalam membangun arsitektur JST secara signifikan hingga 96%, dan memperbaiki performansi hingga 14.84% dalam MAPE, 25% dalam accuracy, dan 40% dalam F-value.

Kata Kunci : Malaria, Prediksi, Evolving Neural Network, Jaringan Syaraf Tiruan, Algoritma Genetika

Abstract

Malaria is an endemic disease in most of Indonesian area, especially in rural and remote areas. Banggai, one of regencies in Central Sulawesi province, is a high endemic area of malaria with Annual Parasite Incidence (API) in 2010 reached 7.88‰. The incidence and spreading of malaria were influenced by environmental and weather factors, particularly rainfall and temperature. Therefore this study would like to developed a malaria incidence prediction system based on environmental and weather factors, so that it may assist Indonesian Ministry of Health to control malaria. The method used to solve the problem was Evolving Neural Network (ENN). This method was a mixture between Artificial Neural Network (ANN) and Genetic Algorithm (GA).

The result of this study shows that the prediction system has acceptable performance for predicting malaria incidence based on weather factors. The best performance while predicting malaria incidence in 2008 was MAPE 21.3%, accuracy 75%, and F-value 84.21%. The best season to predict was dry season with MAPE 13.18%, accuracy 100%, and F-value 100%. As for predicting malaria incidence in 2009, was resulted MAPE 15.29%, accuracy 75%, and F-value 40%. The best season to predict was rainy season with MAPE 3.1%, accuracy 100%, and F-value 100%. These findings proved that there was a sufficient correlation between weather and malaria incidence. ENN reduced trial-and-error process in constructing ANN architecture very significaltly. The reduction was up to 96%. ENN also improved the performance of ANN up to 14.84% in MAPE, 25% in accuracy, and 40% in F-value.

Keywords : Malaria, Prediction, Evolving Neural Network, Artificial Neural Network, Genetic Algorithm

CHAPTER 1: THE PROBLEM

This study was dealing with the prediction of malaria incidence in Banggai Regency, Central Sulawesi, Indonesia. This section discusses the rationale, theoretical framework, conceptual framework/ paradigm, statement of the problem, hypothesis, assumption, scope and delimitation, importance of the study, and definition of terms of the study.

1.1 Rationale

Malaria is an endemic disease in most of Indonesian area, especially in rural and remote areas. On a national scale, malaria is one of diseases as a part of global commitment of Millenium Development Goals (MDG's). Its spreading and incidence are targeted to be stop and reduced in 2015 [11]. Banggai, one of regencies in Central Sulawesi province, is one area that has high endemic of malaria. Data in 2010 shows that malaria Annual Parasite Incidence (API) in Banggai Regency reached 7.88‰, this number were further above than targeted API in malaria control program that is API < 1‰ [6]. Several studies [9][12][17] show that malaria incidence and spreading are influenced by environmental and weather factors, particularly rainfall and temperature. This study was aimed at developing a malaria incidence prediction system based on environmental and weather factors. This was to provide information about prediction of malaria incidence, so that the Indonesian Ministry of Health might take strategic planning in order to control malaria.

Evolving Neural Network (ENN) is a method that integrating Evolutionary Algorithms (EAs) and Artificial Neural Network (ANN). In prediction system, ANN is widely used because of its accuracy and speed. Several studies on malaria prediction [12][17] use back-propagation (BP) algorithm for ANN learning, but there are two shortcomings of this algorithm, namely (1) correction weight which is limited to searching space of steepest descent method and less able to explore searching space which can lead to worse solution or early convergence, (2) sensitive to parameter setup such as number of

hidden layers, number of neuron in hidden layer, learning rate, etc, because the setup was very dependent to the problem encountered [5]. Because of those shortcomings, some studies recommend Genetic Algorithm (GA) as an EAs to replace commonly learning algorithm in ANN, like BP. The advantage of using GA for ANN learning are the definition of ANN parameters can be encoded genetically and evoluted in the same time, so that at the same time there are many solutions which can be selected [5][8]. A study [4] shows that the availability of multiple solutions provide learning performance which is more superior instead of only one solution avalaible.

1.2 Theoretical Framework

This study attempts to predict malaria incidence in Banggai, Indonesia based on weather factors. The input for this system is time series data of weather factors, while the output of this system is a number of malaria incidence in a particular month. In learning process, prediction model is constructed using ENN. First, GA is used to get the optimum ANN structures; then GA is reused to train the ANN. In order to prevent overfitting, cross-validation is performed to select the most optimum structure and weights. To measure performance of the prediction model, the model is tested using testing dataset. Finally, prediction model with best performance (most high accuracy) is used to predict malaria incidence.

1.3 Conceptual Framework/Paradigm

There are three variables applied to conduct measurement on this study, namely:

Variable	Variable's Information
Data composition	Composition of data used as training dataset, validation dataset, and testing dataset.
Number of time series	This variable indicates length of time series of past data that is used as input.
GA parameters	Parameter to setup GA, consists of population size, crossover probability, and mutation probability.

1.4 Statement of the Problem

The main problem discussed in this study was applying ENN to develop prediction system of malaria incidence in Banggai Regency based on time series data of weather factors.

1.5 Hypothesis

ANN is widely used for prediction in a number of areas because of its accuracy. The advantages of ANN for prediction are data error tolerance and ability to approximate complex non-linear mappings [20]. However, there are some difficulties in designing ANN. It does not have general method to determine the optimum structure for solving any problem and the correction weight is limited to searching space of steepest descent method which can lead to worse solution or early convergence. Those are impacted to the accuracy. Therefore this study proposed ENN which used GA to design ANN. ENN can generate many structures and weights (solutions) at the same time. The availability of multiple solutions provides learning accuracy which is more superior than only one solution available, because from those solutions can be selected the most optimum solution which produces the highest accuracy. Thus from those premises, using ENN can produce high accuracy in predicting malaria incidence.

1.6 Assumption

The assumptions of this study were:

1. This study did not conduct further analysis about correlation between weather factors and malaria incidence. All weather factors obtained from BPS-Statistical of Banggai Regency were assumed to have good correlation with malaria incidence in Banggai.
2. Weather forecasting was conducted using time series forecasting program from other study and it was assumed that the accuracy was reliable [16].

1.7 Scope and Delimitation

The scopes and delimitations of this study were:

1. The architecture of ANN was fully connected.
2. This study was conducted based on monthly weather data from BPS-Statistical of Banggai Regency and monthly malaria incidence data from Indonesian Ministry of Health from 2004 – 2009.
3. Maximum number of ANN hidden layer and hidden neuron were 2 and 16.
4. Number of time series of each weather factor was from 1 previous month to 4 previous month.

1.8 Importance of the Study

This study might reveal the ENN performance for predicting malaria incidence, especially ENN which applied GA to optimize ANN structure and to train ANN. Besides that, this study might be useful for the government of Indonesia, especially Ministry of Health, because it can be used to predict the malaria incidence in each regency based on environmental and weather factors at the regency, so that they could make a strategic planning to prevent outbreak and reduce the number of malaria incidence in order to control malaria incidence in Indonesia.

1.9 Definition of Terms

Malaria incidence	The number of malaria incidence that were recorded in Indonesian Ministry of Health.
Weather factors	Factors that indicate the weather condition in an area.
Outbreak	A condition characterized by the increasing in incidence of morbidity/ mortality that epidemiologically significant in a region within a certain period and can lead to the epidemic.

CHAPTER 4: PRESENTATION, ANALYSIS, AND INTERPRETATION OF DATA

This chapter is devoted to present and analysis the experiment result of the prediction system.

4.1 Presentation of Data

Experiment was conducted using 10 weather factors as input, i.e. rainfall, precipitation day, minimum temperature, maximum temperature, average temperature, average humidity, maximum wind velocity, average wind velocity, maximum direction of wind, and average length of daylight. Data of those factors were normalized and composed into n-time series dataset, where n = [1..4]. This means the weather data used as the input to predict malaria incidence in month m are weather data on month m–1, m–2, …, m–n. Table 4-1 and 4-2 are sample of dataset before and after normalized.

Table 4-1. Sample of Dataset 1-Time Series Before Normalized (Year 2004)

Weather Data on Month m-1										Malaria Incidence on Month m
Factor #1	Factor #2	Factor #3	Factor #4	Factor #5	Factor #6	Factor #7	Factor #8	Factor #9	Factor #10	
69	19	23.4	33.3	28.3	77	20	4	270	62	155
89	14	23.4	33.8	27.9	77	24	4	270	34	226
165	18	24	33.3	28.1	79	21	4	270	69	178
56	14	24.4	31.7	28	77	22	6	260	71	125
111	17	23.8	31.8	27.9	78	18	5	270	68	140
185	24	22.2	30.5	26.3	81	26	6	200	41	89
142	23	22.2	29.7	25.9	80	19	7	220	36	116
18	10	21	29.8	25.9	72	20	8	270	61	184
17	10	21.6	31	26.8	71	23	8	180	73	124
5	1	23	32	28.2	71	17	6	180	94	75
6	7	24	32.6	29	72	26	5	270	79	65

Table 4-2. Sample of Dataset 1-Time Series After Normalized (Year 2004)

Weather Data on Month m-1										Malaria Incidence on Month m
Factor #1	Factor #2	Factor #3	Factor #4	Factor #5	Factor #6	Factor #7	Factor #8	Factor #9	Factor #10	
0.1885	0.5765	0.4804	0.7051	0.7137	0.5392	0.4524	0.3000	0.6389	0.5196	0.2997
0.2142	0.4511	0.4804	0.7479	0.6453	0.5392	0.5794	0.3000	0.6389	0.2451	0.4047
0.3116	0.5514	0.5588	0.7051	0.6795	0.5784	0.4841	0.3000	0.6389	0.5882	0.3337
0.1718	0.4511	0.6111	0.5684	0.6624	0.5392	0.5159	0.5667	0.6111	0.6078	0.2554
0.2424	0.5263	0.5327	0.5769	0.6453	0.5588	0.3889	0.4333	0.6389	0.5784	0.2775
0.3373	0.7019	0.3235	0.4658	0.3718	0.6176	0.6429	0.5667	0.4444	0.3137	0.2021
0.2821	0.6768	0.3235	0.3974	0.3034	0.5980	0.4206	0.7000	0.5000	0.2647	0.2421
0.1231	0.3508	0.1667	0.4060	0.3034	0.4412	0.4524	0.8333	0.6389	0.5098	0.3426
0.1218	0.3508	0.2451	0.5085	0.4573	0.4216	0.5476	0.8333	0.3889	0.6275	0.2539
0.1064	0.1251	0.4281	0.5940	0.6966	0.4216	0.3571	0.5667	0.3889	0.8333	0.1814
0.1077	0.2755	0.5588	0.6453	0.8333	0.4412	0.6429	0.4333	0.6389	0.6863	0.1667

Notes

Factor #1: Rainfall

Factor #2: Precipitation Day

Factor #3: Minimum Temperature

Factor #4: Maximum Temperature

Factor #5: Average Temperature

Factor #6: Average Humidity

Factor #7: Maximum Wind Velocity

Factor #8: Average Wind Velocity

Factor #9: Maximum Direction of Wind

Factor #10: Average Length of Daylight

4.1.1 Experiment Result Using Data Scenario 1

In scenario 1, experiment was conducted using data 2004 – 2006 as training dataset, data 2007 as validation dataset, and data 2008 as testing dataset, with parameter scenario as described in sub-chapter 3-3. The results of training process using data scenario 1 are shown in Table 4-3. All architectures that resulted from training process then tested using testing dataset to measure their performance. Those architectures were tested to predict malaria incidences for 12 months that were divided into 4 seasons, i.e. rainy season, transition from rainy to dry season, dry season, and transition from dry to rainy season. Figure 4-1 is the testing result to predict malaria in 2008 using all architectures that resulted from training process.

Table 4-3. Training Result of Data Scenario 1

No	Pop. Size S	Pop. Size W	Pc	Pm	n-Time Series	Structure and Weight Training	
						Best Structure	Max Average Fitness
1	50	100	0.6	0.1	1	HL: 1; HN: 15	4.5374
2	50	100	0.6	0.1	2	HL: 1; HN: 8	4.7750
3	50	100	0.6	0.1	3	HL: 1; HN: 11	5.0571
4	50	100	0.6	0.1	4	HL: 1; HN: 12	5.2103
5	50	100	0.6	0.01	1	HL: 1; HN: 12	4.3307
6	50	100	0.6	0.01	2	HL: 1; HN: 16	4.6176
7	50	100	0.6	0.01	3	HL: 1; HN: 13	4.7481
8	50	100	0.6	0.01	4	HL: 1; HN: 15	4.9058
9	50	100	0.6	0.001	1	HL: 1; HN: 14	4.1384
10	50	100	0.6	0.001	2	HL: 1; HN: 13	4.1749
11	50	100	0.6	0.001	3	HL: 1; HN: 8	4.3102
12	50	100	0.6	0.001	4	HL: 1; HN: 11	4.6307
13	50	100	0.8	0.1	1	HL: 1; HN: 15	4.4931
14	50	100	0.8	0.1	2	HL: 1; HN: 14	4.7931
15	50	100	0.8	0.1	3	HL: 1; HN: 12	4.9857
16	50	100	0.8	0.1	4	HL: 1; HN: 16	5.1908
17	50	100	0.8	0.01	1	HL: 1; HN: 16	4.4367
18	50	100	0.8	0.01	2	HL: 1; HN: 11	4.6100
19	50	100	0.8	0.01	3	HL: 1; HN: 16	4.7875
20	50	100	0.8	0.01	4	HL: 1; HN: 11	4.7840
21	50	100	0.8	0.001	1	HL: 1; HN: 8	4.0986
22	50	100	0.8	0.001	2	HL: 1; HN: 16	4.4505
23	50	100	0.8	0.001	3	HL: 2; HN1:14; HN2:12	4.3136
24	50	100	0.8	0.001	4	HL: 1; HN: 14	4.9408
25	100	200	0.6	0.1	1	HL: 1; HN: 14	4.5298
26	100	200	0.6	0.1	2	HL: 1; HN: 16	4.8064
27	100	200	0.6	0.1	3	HL: 1; HN: 15	4.9977
28	100	200	0.6	0.1	4	HL: 1; HN: 15	5.2801
29	100	200	0.6	0.01	1	HL: 1; HN: 16	4.5014
30	100	200	0.6	0.01	2	HL: 1; HN: 14	4.5175
31	100	200	0.6	0.01	3	HL: 1; HN: 15	5.0663
32	100	200	0.6	0.01	4	HL: 1; HN: 14	5.0287
33	100	200	0.6	0.001	1	HL: 1; HN: 3	3.9924
34	100	200	0.6	0.001	2	HL: 1; HN: 12	4.5511
35	100	200	0.6	0.001	3	HL: 1; HN: 13	4.6617
36	100	200	0.6	0.001	4	HL: 1; HN: 15	4.8882
37	100	200	0.8	0.1	1	HL: 1; HN: 12	4.5315
38	100	200	0.8	0.1	2	HL: 1; HN: 11	4.8043
39	100	200	0.8	0.1	3	HL: 1; HN: 15	5.0321
40	100	200	0.8	0.1	4	HL: 1; HN: 13	5.2832
41	100	200	0.8	0.01	1	HL: 1; HN: 14	4.4167
42	100	200	0.8	0.01	2	HL: 1; HN: 15	4.7785
43	100	200	0.8	0.01	3	HL: 1; HN: 13	4.8910
44	100	200	0.8	0.01	4	HL: 1; HN: 7	4.9887
45	100	200	0.8	0.001	1	HL: 1; HN: 16	4.3316
46	100	200	0.8	0.001	2	HL: 1; HN: 10	4.6926
47	100	200	0.8	0.001	3	HL: 1; HN: 15	4.9252
48	100	200	0.8	0.001	4	HL: 1; HN: 15	4.8150

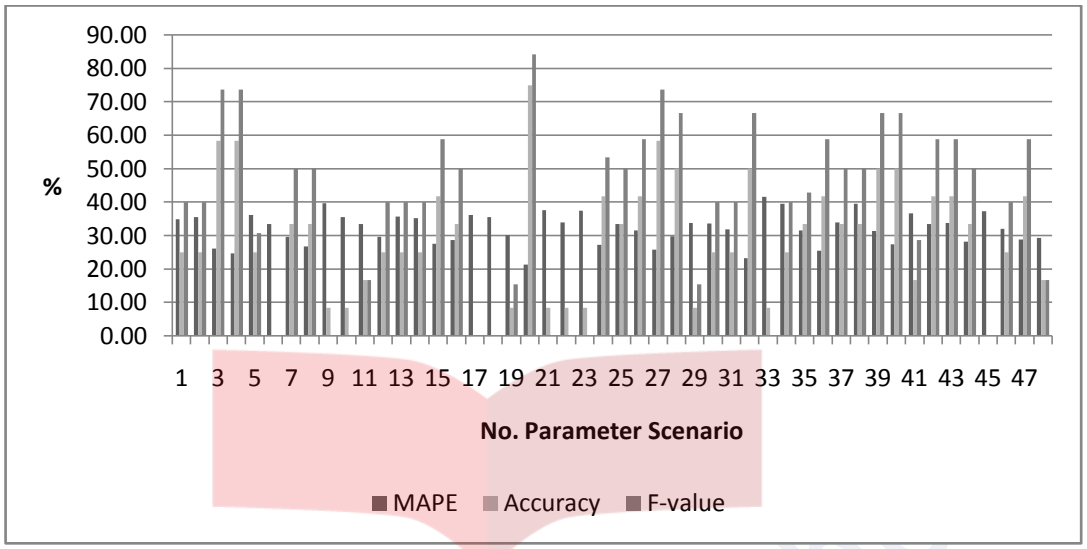

Figure 4-1. MAPE, Accuracy, and F-value of 12 Months Testing (Jan – Dec 2008)

Based on Figure 4-1, the scenario that resulted in the best testing MAPE, accuracy, and F-value was scenario number 20, where the combination of parameters are Pop Size S = 50, Pop Size W = 100, Pc = 0.8, Pm = 0.01, and n-Time Series = 4. That scenario obtained ANN structure: one hidden layer and 11 neurons in the hidden layer, along with its weights. Figure 4-2, Figure 4-3, and Table 4-4 are the visualization of best testing result. Similar architecture also used for testing by using weather forecast data. Figure 4-4, Figure 4-5, and Table 4-4 are the visualization of testing result using weather forecast data. While Figure 4-6 describes system performance that divided into four seasons. Based on that figure, system has the best performance while predicting malaria incidence in dry season.

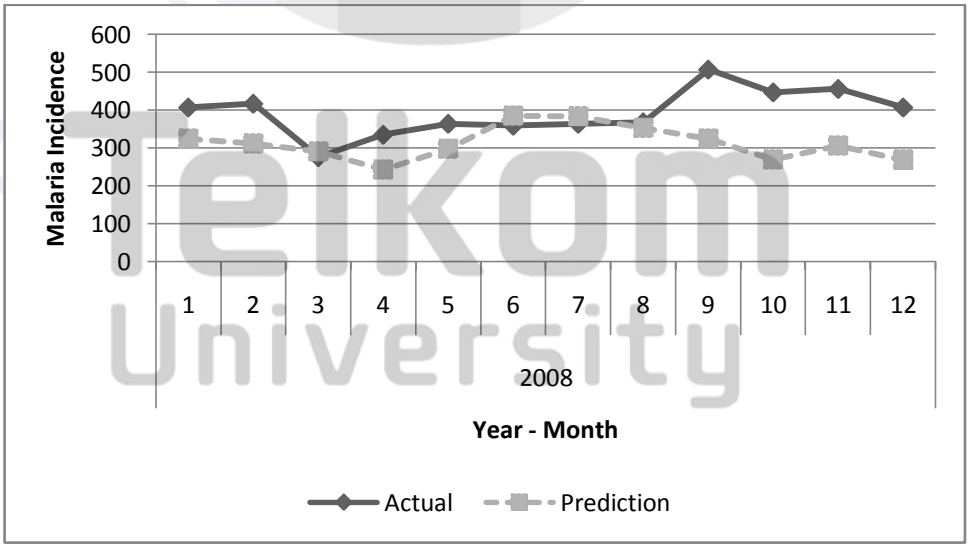

Figure 4-2. Best Testing Result Jan – Dec 2008 Based on Actual Weather Data

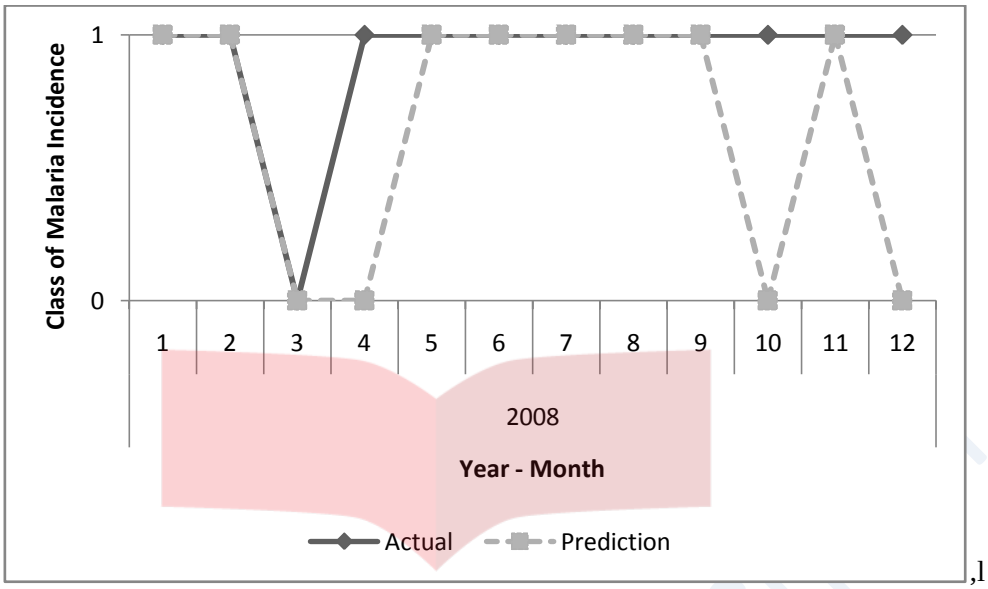

Figure 4-3. Classification of Best Testing Jan – Dec 2008 Based on Actual Weather Data

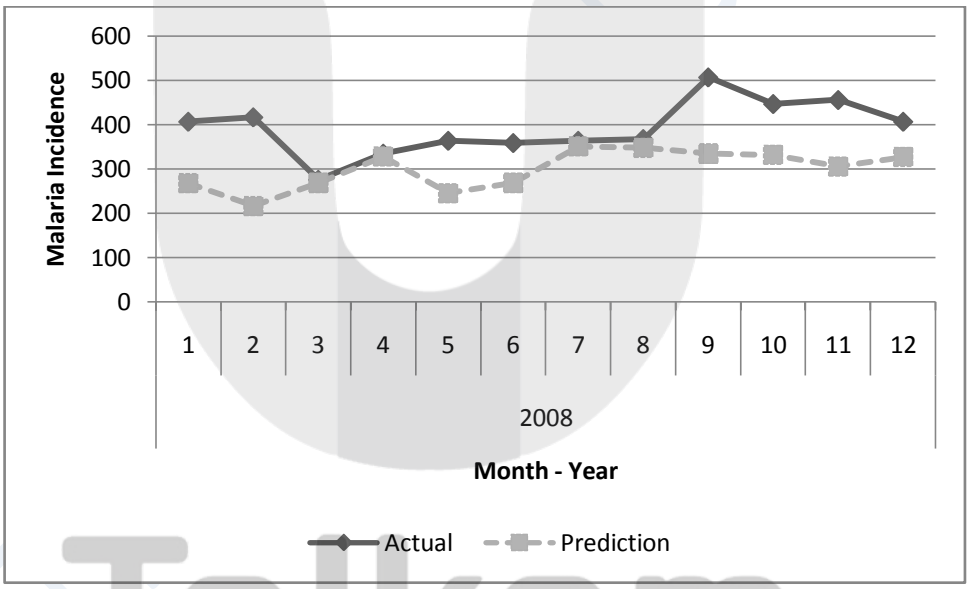

Figure 4-4. Best Testing Result Jan – Dec 2008 Based on Weather Forecast Data

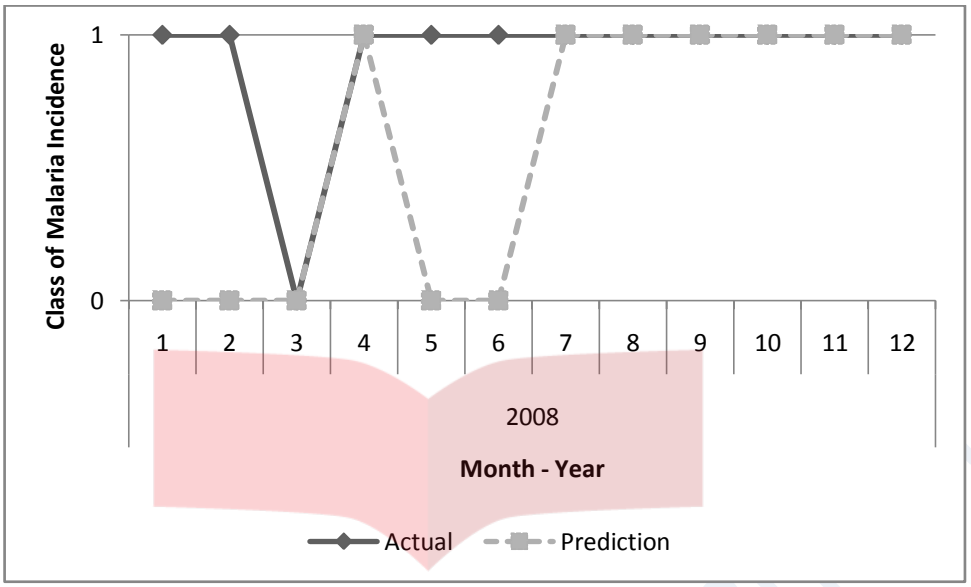

Figure 4-5. Classification of Best Testing Jan – Dec 2008 Based on Weather Forecast Data

Table 4-4. Best Testing Result on Year 2008

Month	Actual	Class	Testing Based on Actual Weather Data			Testing Based on Weather Forecast Data		
			Prediction	APE (%)	Class	Prediction	APE (%)	Class
1	407	1	324	20.39	1	268	34.15	0
2	417	1	312	25.18	1	217	47.96	0
3	276	0	291	5.43	0	269	2.54	0
4	335	1	243	27.46	0	328	2.09	1
5	364	1	298	18.13	1	246	32.42	0
6	359	1	385	7.24	1	269	25.07	0
7	364	1	383	5.22	1	351	3.57	1
8	368	1	352	4.35	1	348	5.43	1
9	507	1	325	35.90	1	335	33.93	1
10	447	1	270	39.60	0	332	25.73	1
11	456	1	306	32.89	1	306	32.89	1
12	407	1	269	33.91	0	327	19.66	1
			MAPE(%)	**21.30**		**MAPE(%)**	**22.12**	
			Accuracy (%)		**75**	**Accuracy (%)**		**66.67**
			F-value(%)		**84.21**	**F-value(%)**		**77.78**

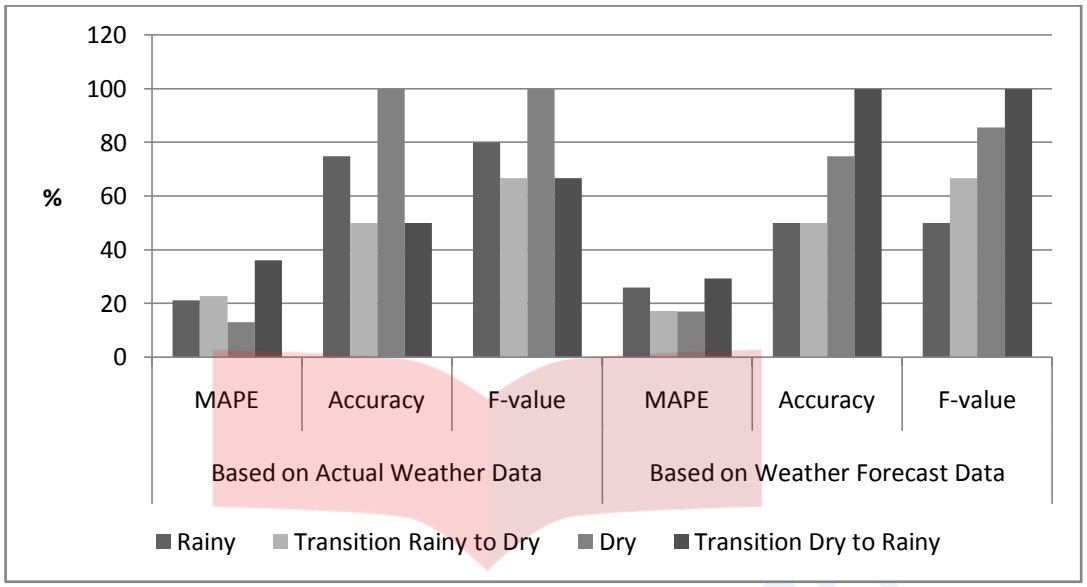

Figure 4-6. System Performance of Best Testing Jan – Dec 2008 in Four Seasons

4.1.2 Experiment Result Using Data Scenario 2

In scenario 2, experiment was conducted using data 2004 – 2007 as training dataset, data 2008 as validation dataset, and data 2009 as testing dataset, with parameter scenario as described in sub-chapter 3-3. The results of training process using data scenario 2 are shown in Table 4-5. All architectures that resulted from training process then tested using testing dataset to measure their performance. Those architectures were tested to predict malaria incidences for 12 months that were divided into 4 seasons, i.e. rainy season, transition from rainy to dry season, dry season, and transition from dry to rainy season. Figure 4-7 is the testing result to predict malaria in 2009 using all architectures that resulted from training process.

Based on Figure 4-7, the scenario that resulted in the best testing MAPE, accuracy, and F-value was scenario number 5, where the combination of parameters are Pop Size S = 50, Pop Size W = 100, Pc = 0.6, Pm = 0.01, and n-Time Series = 1. That scenario obtained ANN structure: one hidden layer and 10 neurons in the hidden layer, along with its weights. Figure 4-8, Figure 4-9, and Table 4-6 are the visualization of the best testing result. Similar architecture also used for testing by using weather forecast data. Figure 4-10, Figure 4-11, and Table 4-6 are the visualization of testing result using weather forecast data. While Figure 4-12 describes system performance that divided into four seasons. Based on that figure, system has the worst performance while predicting malaria incidence in transition dry to rainy season.

Table 4-5. Training Result of Data Scenario 2

No	Pop. Size S	Pop. Size W	Pc	Pm	n-Time Series	Structure and Weight Training	
						Best Structure	Max Average Fitness
1	50	100	0.6	0.1	1	HL: 1; HN: 13	3.9508
2	50	100	0.6	0.1	2	HL: 1; HN: 6	4.0989
3	50	100	0.6	0.1	3	HL: 1; HN: 14	4.2065
4	50	100	0.6	0.1	4	HL: 1; HN: 14	4.3338
5	50	100	0.6	0.01	1	HL: 1; HN: 10	3.9316
6	50	100	0.6	0.01	2	HL: 1; HN: 12	3.8307
7	50	100	0.6	0.01	3	HL: 1; HN: 11	3.9525
8	50	100	0.6	0.01	4	HL: 1; HN: 15	4.1806
9	50	100	0.6	0.001	1	HL: 1; HN: 10	3.6184
10	50	100	0.6	0.001	2	HL: 1; HN: 11	4.0225
11	50	100	0.6	0.001	3	HL: 1; HN: 14	4.0803
12	50	100	0.6	0.001	4	HL: 1; HN: 16	4.0447
13	50	100	0.8	0.1	1	HL: 1; HN: 14	4.0170
14	50	100	0.8	0.1	2	HL: 1; HN: 15	4.0468
15	50	100	0.8	0.1	3	HL: 1; HN: 10	4.1783
16	50	100	0.8	0.1	4	HL: 1; HN: 11	4.2700
17	50	100	0.8	0.01	1	HL: 1; HN: 13	3.8983
18	50	100	0.8	0.01	2	HL: 1; HN: 10	3.9759
19	50	100	0.8	0.01	3	HL: 1; HN: 11	3.9922
20	50	100	0.8	0.01	4	HL: 1; HN: 16	4.2588
21	50	100	0.8	0.001	1	HL: 1; HN: 15	3.7317
22	50	100	0.8	0.001	2	HL: 1; HN: 16	3.7539
23	50	100	0.8	0.001	3	HL: 1; HN: 9	3.9628
24	50	100	0.8	0.001	4	HL: 2; HN1: 13; HN2: 15	3.9030
25	100	200	0.6	0.1	1	HL: 1; HN: 16	3.9166
26	100	200	0.6	0.1	2	HL: 1; HN: 2	3.7333
27	100	200	0.6	0.1	3	HL: 1; HN: 16	4.1184
28	100	200	0.6	0.1	4	HL: 1; HN: 9	4.3746
29	100	200	0.6	0.01	1	HL: 1; HN: 16	3.968
30	100	200	0.6	0.01	2	HL: 1; HN: 14	4.0197
31	100	200	0.6	0.01	3	HL: 1; HN: 9	4.1333
32	100	200	0.6	0.01	4	HL: 1; HN: 16	4.0525
33	100	200	0.6	0.001	1	HL: 1; HN: 6	3.7096
34	100	200	0.6	0.001	2	HL: 1; HN: 13	4.0944
35	100	200	0.6	0.001	3	HL: 1; HN: 13	3.9672
36	100	200	0.6	0.001	4	HL: 1; HN: 14	4.2254
37	100	200	0.8	0.1	1	HL: 1; HN: 16	4.003
38	100	200	0.8	0.1	2	HL: 1; HN: 8	4.0729
39	100	200	0.8	0.1	3	HL: 1; HN: 10	4.1463
40	100	200	0.8	0.1	4	HL: 1; HN: 13	4.3178
41	100	200	0.8	0.01	1	HL: 1; HN: 12	3.9896
42	100	200	0.8	0.01	2	HL: 1; HN: 11	4.1335
43	100	200	0.8	0.01	3	HL: 1; HN: 9	4.1515
44	100	200	0.8	0.01	4	HL: 1; HN: 11	4.0242
45	100	200	0.8	0.001	1	HL: 1; HN: 7	3.7088
46	100	200	0.8	0.001	2	HL: 1; HN: 3	3.6993
47	100	200	0.8	0.001	3	HL: 1; HN: 10	3.9602
48	100	200	0.8	0.001	4	HL: 1; HN: 9	4.1886

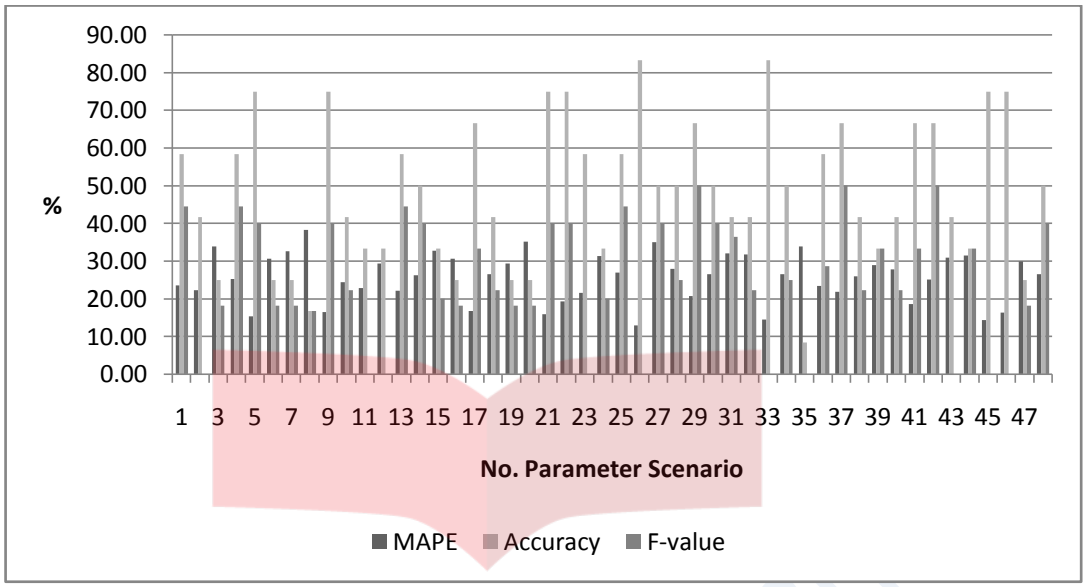

Figure 4-7. MAPE, Accuracy, and F-value of 12 Months Testing (Jan – Dec 2009)

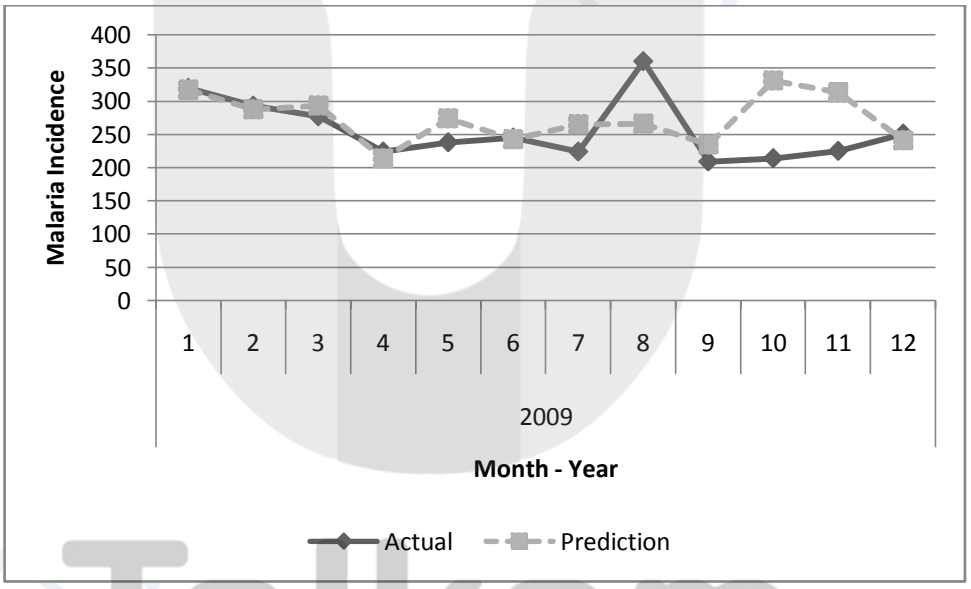

Figure 4-8. Best Testing Result Jan – Dec 2009 Based on Actual Weather Data

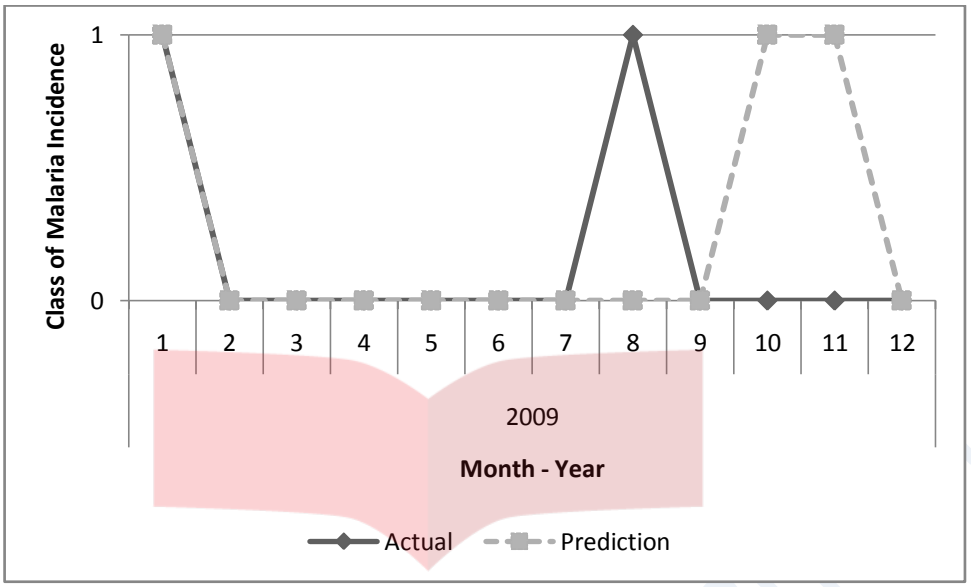

Figure 4-9. Classification of Best Testing Jan – Dec 2009 Based on Actual Weather Data

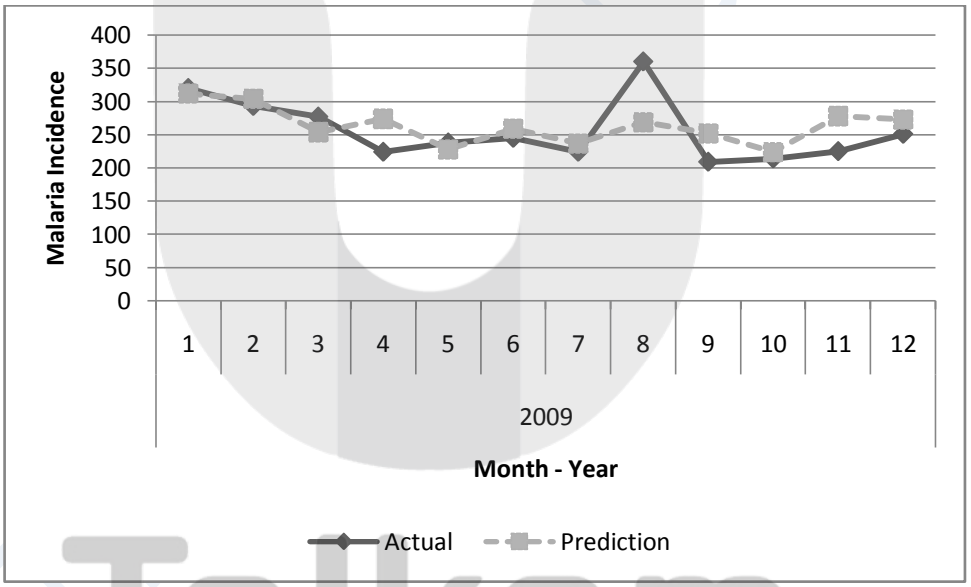

Figure 4-10. Best Testing Result Jan – Dec 2009 Based on Weather Forecast Data

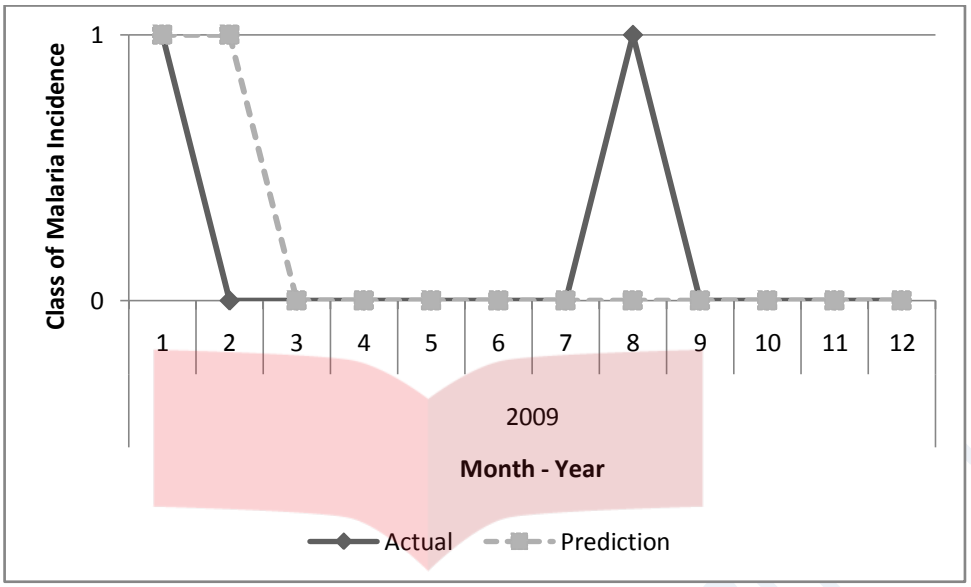

Figure 4-11. Classification of Best Testing Jan – Dec 2009 Based on Weather Forecast Data

Table 4-6. Best Testing Result on Year 2009

Month	Actual	Class	Testing Based on Actual Weather Data			Testing Based on Weather Forecast Data		
			Prediction	APE (%)	Class	Prediction	APE (%)	Class
1	320	1	317	0.94	1	312	2.50	1
2	293	0	288	1.71	0	304	3.75	1
3	277	0	293	5.78	0	254	8.30	0
4	224	0	214	4.46	0	274	22.32	0
5	238	0	274	15.13	0	228	4.20	0
6	245	0	243	0.82	0	259	5.71	0
7	224	0	265	18.30	0	237	5.80	0
8	360	1	266	26.11	0	269	25.28	0
9	209	0	235	12.44	0	252	20.57	0
10	214	0	331	54.67	1	224	4.67	0
11	225	0	313	39.11	1	278	23.56	0
12	251	0	241	3.98	0	273	8.76	0
			MAPE(%)	**15.29**		**MAPE(%)**	**11.29**	
			Accuracy (%)		**75**	**Accuracy (%)**		**83.33**
			F-value (%)		**40**	**F-value (%)**		**50**

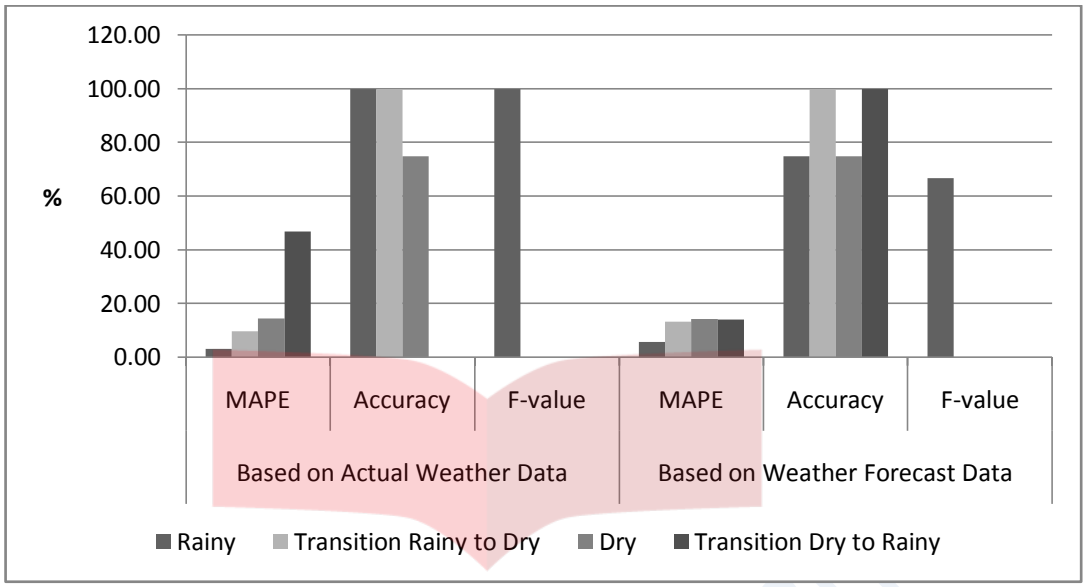

Figure 4-12. System Performance of Best Testing Jan – Dec 2009 in Four Seasons

4.1.3 Prediction of Malaria Incidence in 2010

In this stage, malaria incidence in 2010 was predicted using weather forecast data. The experiment showed that for predicting malaria incidence using weather forecast data, the prediction model from data scenario 2 resulted in better accuracy with reliable F-value, therefore it was used for predicting malaria incidence in 2010 as shown in Table 4-7. The most recommended seasons to predict were rainy, transition rainy to dry, and dry because those resulted in reliable performance whether while tested by using actual weather data or weather forecast data.

Table 4-7. Prediction of Malaria Incidence in 2010

Month	Malaria Incidence	Class
1	208	0
2	199	0
3	111	0
4	128	0
5	128	0
6	143	0
7	128	0
8	126	0
9	146	0
10	140	0
11	124	0
12	135	0

4.1.4 Experiment Result for Determining Threshold Value of Malaria Incidence Classification

Predicting "outbreak" class is more important than predicting "normal" class. Therefore besides resulting minimum error in predicting number of malaria incidence, the system also has to be accurate in predicting the outbreak. In order to achieve that goal, a proper threshold value is required. Because in this study "outbreak" was minority class then the performance was not only measured using accuracy but also using F-measure where performance in predicting "outbreak" was more focused. As for the experiment scenarios in determining threshold value are shown in Table 4-8.

Table 4-8. Threshold Experiment Scenario

No	Scenario	Value
1	mean + (2 * standard deviation)	479
2	mean + (1.5 * standard deviation)	424
3	mean + (1 * standard deviation)	370
4	mean + (0.6 * standard deviation)	326
5	mean + (0.5 * standard deviation)	315
6	mean + (0.3 * standard deviation)	293

where: mean = 259.71

standard deviation = 109.81

From the experiment result as shown in Figure 4-13, there was a trade-off between accuracy and F-value. Using threshold scenario number 1 and number 2, i.e. mean + (2 * standard deviation) and mean + (1.5 * standard deviation), the system obtained high accuracy, but the system could not predict the outbreak; while using threshold scenario number 6, i.e. mean + (0.3 * standard deviation), the system obtained lower accuracy but could predict the outbreak with acceptable performance. Therefore the proper threshold that was selected to distinguish "outbreak" class and "normal" class was threshold scenario number 6, because it gave acceptable result in accuracy and F-measure for both data scenario.

According to the previous description, accuracy was not the only key parameter to measure the system performance. The system performance also should be measured by considering the F-value which represented the system performance in predicting the outbreak, where the outbreak was determined by a threshold value. More detailed results of this experiment are shown in Appendix 4.

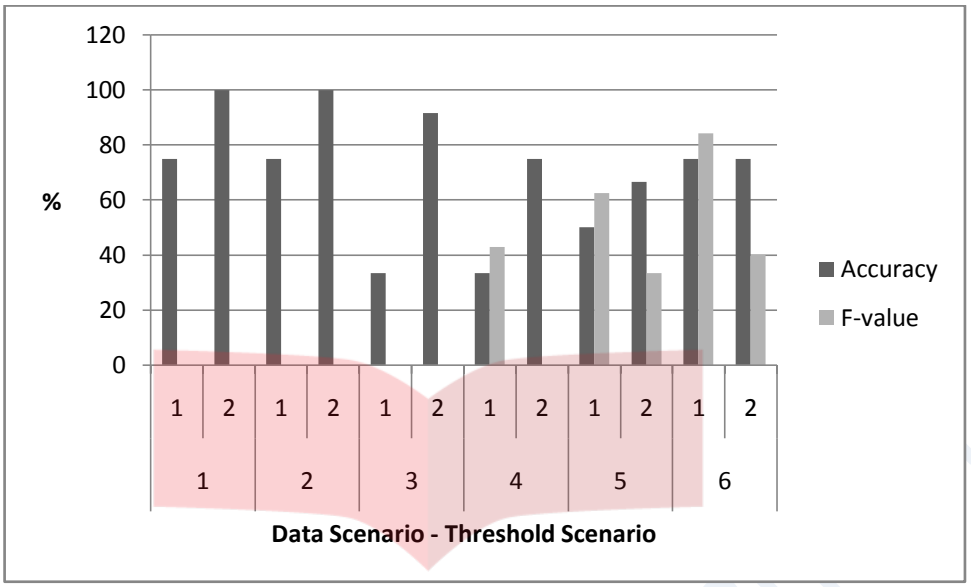

Figure 4-13. Accuracy and F-value for Each Threshold

4.1.5 Experiment Result for Determining Threshold Value of Malaria Incidence Classification Using Smoothed Malaria Data

In the experiment as described in sub-chapter 4.1.4, the prediction model was used to predict the actual malaria data. While in this experiment, the prediction model was used to predict the smoothed malaria data. The purpose of this experiment was to obtain a higher threshold value and improve the performance that resulted from experiment in sub-chapter 4.1.4. There were two scenarios of prediction model and three smoothing methods that were performed. The scenarios of prediction model were (1) using the best prediction model of data scenario 1 and data scenario; (2) retraining the system to generate model based on smoothed malaria data, using the same parameters that resulted in the best prediction model of data scenario 1 and data scenario 2. While the smoothing methods that were used are described in Table 4-10. Figure 4.14 and Figure 4.15 are the summary result of this experiment. Based on those figures, the system performance without smoothing in malaria data was better than with smoothing. This experiment also did not obtain a higher threshold value. Therefore, this study prefered to use the malaria data without smoothing. More detailed results of this experiment are shown in Appendix 5.

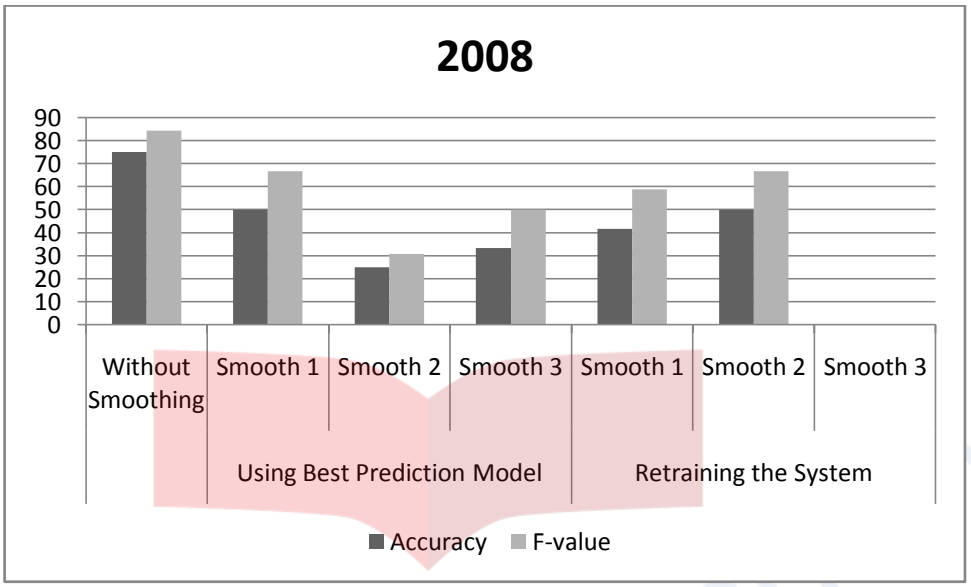

Figure 4-14. System Performance of Data Scenario 1 Before and After Malaria Smoothing

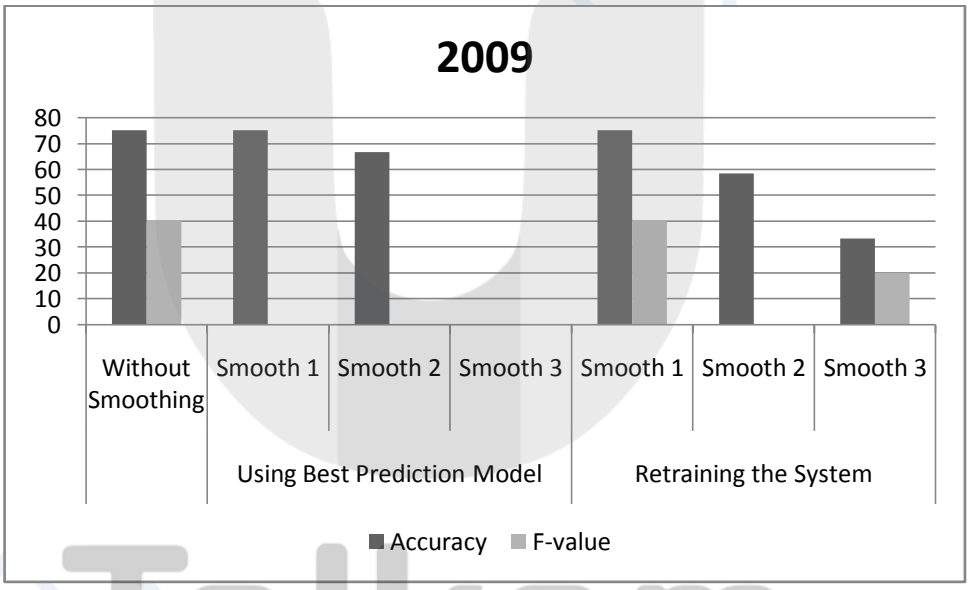

Figure 4-15. System Performance of Data Scenario 2 Before and After Malaria Smoothing

4.1.6 Weather Forecasting Result

Weather factors were forecasted by time series forecasting using Differential Evolution. Table 4-9 is MAPE of the forecasting result. More detailed results of weather forecasting are shown in Appendix 6.

Table 4-9. MAPE of Weather Forecasting

Weather Factors	MAPE (%)		
	2008	2009	2010
Rainfall	15.80	17.36	14.93
Precipitation Day	28.62	33.32	14.33
Minimum Temperature	2.18	1.95	5.12
Maximum Temperature	2.53	2.57	2.92
Average Temperature	1.68	1.21	2.23
Average Humidity	9.10	4.46	2.78
Maximum Wind Velocity	19.38	12.93	16.01
Average Wind Velocity	17.02	18.29	13.06
Maximum Direction of Wind	14.87	25.20	10.42
Average Length of Daylight	22.16	20.21	19.32

All weather factors were forecasted using actual data, except for the rainfall. Before forecasted, the rainfall data were smoothed (Figure 4-14). It was conducted to accommodate the very high fluctuation of rainfall because the data only available from one station, while the rainfall was very area-dependent. There were three smoothing methods used in this experiment as described in Table 4-10. Smoothing method resulted in the best performance was Moving Average 1. It resulted in better MAPE, accuracy, and F-value, especially in data scenario 1, as shown in Figure 4-15. As for data scenario 2, there was a decrease in MAPE but in a very small value, i.e. 0.07%. More detailed results of this experiment are described in Appendix 7.

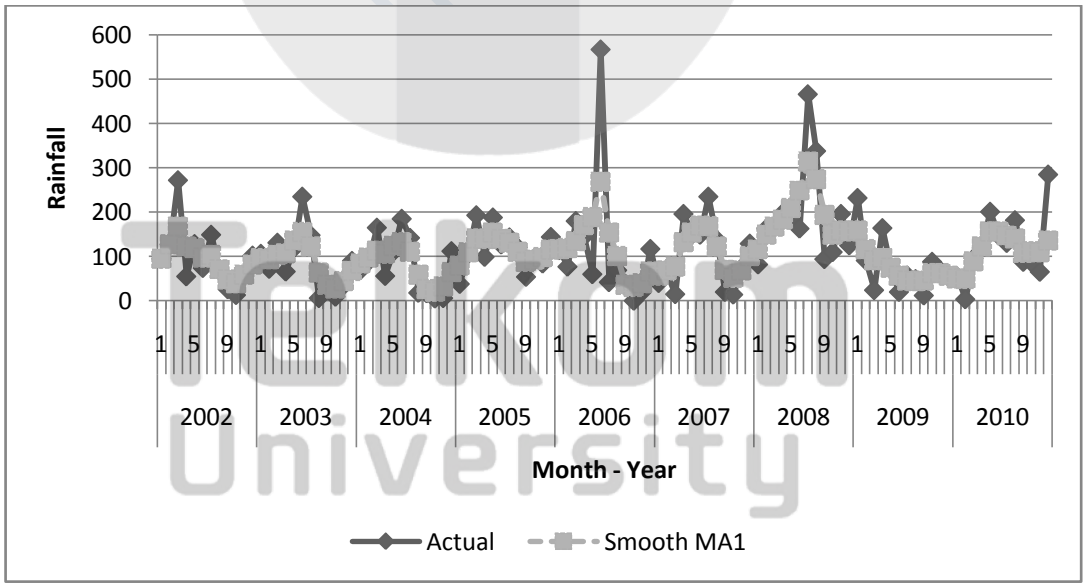

Figure 4-16. Rainfall Data Smoothing Using Moving Average 1

Table 4-10. Smoothing Methods

No	Methods	Equation
1	Moving Average 1 (MA1)	$S_t = \frac{A_{t-2} + (2 * A_{t-1}) + (4 * A_t) + (2 * A_{t+1}) + A_{t+2}}{10}$
2	Moving Average 2 (MA2)	$S_t = \frac{(2 * A_{t-1}) + (6 * A_t) + (2 * A_{t+1})}{10}$
3	Exponential Smoothing According to Hunter with damping factor α = 0.2	$S_t \begin{cases} A_1; \text{ if } t = 1 \\ \alpha A_{t-1} + (1-\alpha) S_{t-1}; \text{ if } t > 1 \end{cases}$

where: S_t = smooth value

A_t = actual value

t = time

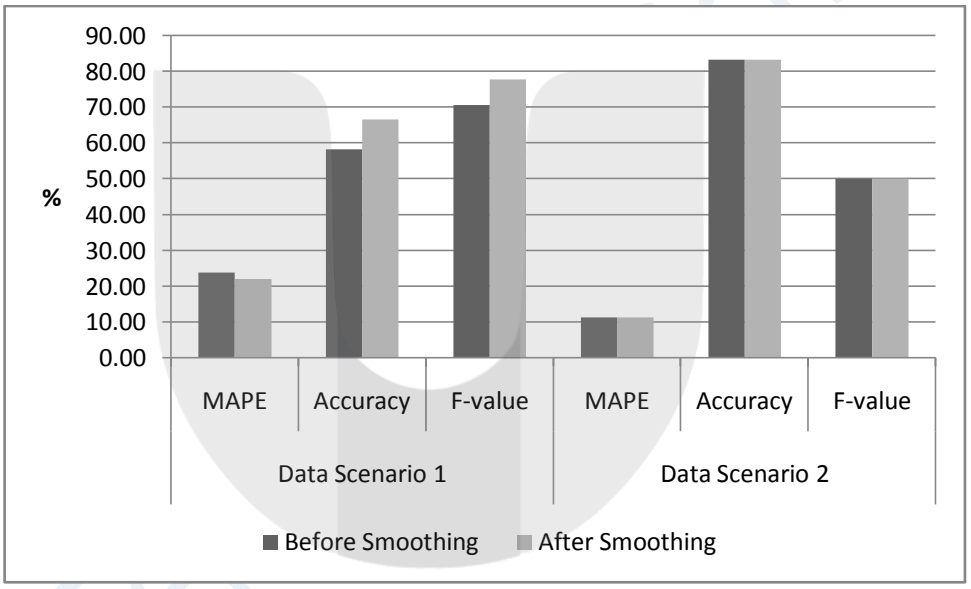

Figure 4-17. System Performance Before and After Rainfall Smoothing

4.1.7 Result of Prediction Using Back-Propagation Neural Network

Figure 4-16 shows the performance of prediction system using back-propagation neural network (BPNN) and Evolving Neural Network (ENN). The prediction system constructed by BPNN using the same data scenario, structure, and epoch as ENN best prediction model to predict malaria incidence in 2008 and 2009 using actual weather data. From Figure 4-16, ENN has better performance than BPNN in predicting malaria incidence both in 2008 and 2009. In 2009, BPNN obtained better accuracy than ENN, but the F-value was zero. It indicated that the BPNN's model was not able to predict the "outbreak" class,

whereas in this study predicting the outbreak was more important than predicting the "normal" class.

Besides that, another important point was ENN reduced trial-and-error process in constructing ANN architecture very significaltly. In this study, to find optimum architecture was used 2 data scenarios where for each data scenario was conducted experiment using 48 parameter combinations. Limitation in finding optimum architecture was maximum number of hidden layer = 2 and maximum number of neuron in each hidden layer = 16. If using BPNN then required:

- trial-and-error to try all possible structures from 1 Hidden Layer – 1 Hidden Neuron until 2 Hidden Layer – 16 Hidden Neuron 1 – 16 Hidden Neuron 2
- trial-and-error to set learning rate
- trial-and-error to try all possible number of time series, i.e. 1 – 4

Assumed only one learning rate used, then to try all the possibilities at least required experiment as many as (4*16)+(4*16*16) = 1088 experiments for each data scenario. Those amount was so much higher than number of experiment that need to be conducted if using ENN, i.e. only 48 experiments for each scenario. Therefore, ENN was a really worth effort for constructing NN architecture.

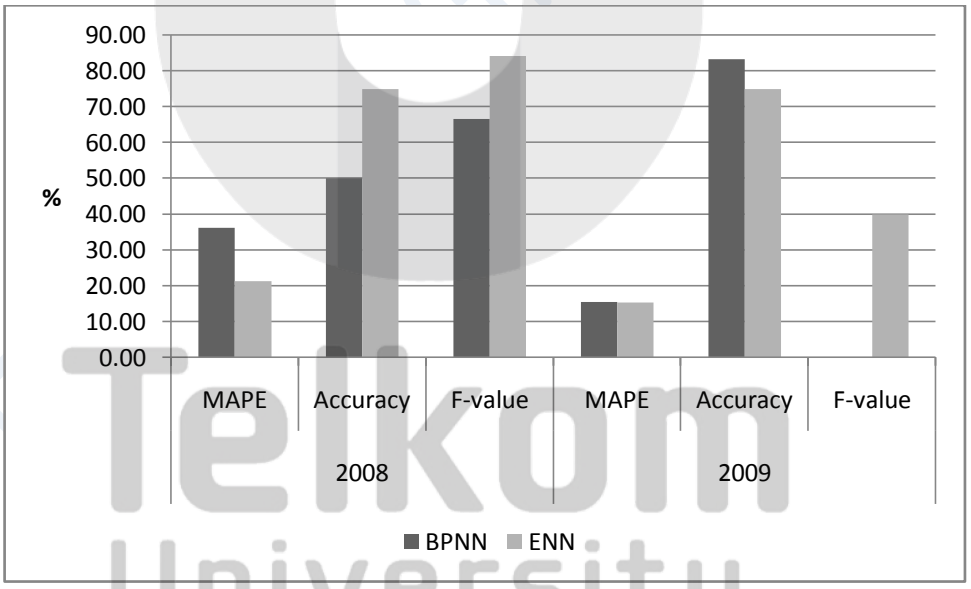

Figure 4-18. Comparison of System Performance Using BPNN and ENN

4.2 Analysis of the Data

The experiment results were analayzed into two types of analysis, namely training and testing analysis.

4.2.1 Training Analysis

These are the analysis from the training result.

1. Determining the number of hidden layer and number of neurons in hidden layer is definitely a difficult problem because there are no definite rules. So, usually the determination was conducted by trial-and-error. For constructing prediction model of this study was provided structure with maximum number of hidden layer = 2 and maximum number of neurons in each hidden layer = 16. The result was 98% optimum structures that resulted from structure optimization using data senario 1 and data scenario 2 used only one hidden layer. It indicated that mapping problem from weather data to malaria incidence was a complex problem, but there was a quite strong correlation, therefore the use of one hidden layer was sufficient for solving this problem. As for the number of neurons in each hidden layer, each scenario resulted in different result. It indicated that determining number of neurons in hidden layer was a more difficult problem. But from the results of this study can be said that for predicting malaria incidence based on weather factors was required at least two neurons in hidden layer. However based on the result that 90% optimum structures from structure optimization resulted the number of neurons more than 8, the most recommended was at least required 8 neurons in hidden layer for solving this problem, with the number of inputs 10 – 40.
2. For all parameter scenarios, training by using data scenario 1 was resulted in better fitness than data scenario 2. It indicated that easier to find function that can generalize pattern in data scenario 1 than data scenario 2. As the weather pattern was relatively more regular, based on visualizations of malaria incidence in Appendix 2 and Appendix 3 were seen that the pattern that has to approximated in data scenario 1 was easier than data scenario 2, the trendline was clearer, and because the number of data was fewer then the number of pattern that has to generalized was also fewer.
3. In training process, the most influential parameter was the number of time series. For both data scenario, the greater the number of time series was, the fitness value would be. The number of time series indicated number of information as input for the system. It meant that in the range of n-Time Series tested in this study, more

number of information was better for constructing the prediction model. As for the GA parameters, the most influential parameter was mutation probability. Mutation plays a role in generating diversity. From the experiment result, generally mutation probability Pm = 0.1 was resulted in better prediction model than Pm = 0.01 or Pm = 0.001. It indicated that mutation probability that is too small causes a decrease in probability to get more diverse model. The decrease in probability to get diverse model was, the probability to get optimum model would be. For the other GA parameters, i.e. population size and crossover probability, the influence of tested values were not significant.

4.2.2 Testing Analysis

The following are the analysis from the testing result.

1. In term MAPE and accuracy, best prediction model that generated using data scenario 2 resulted better performance than best prediction model that generated using data scenario 1, whether while predicting malaria incidence using actual weather data or weather forecast data. It because the number of training data in data scenario 2 was more than data scenario 1. More number of training data meant more information to be learned, therefore the performance was also better.
2. Testing dataset of data scenario 1 and data scenario 2 had different characteristic. In data scenario 1 the "outbreak" was majority class, while in data scenario 2 the "outbreak" was minority class. With those two different conditions, ENN was able to generate model that can predict the incidence with reliable performance.
3. Based on the testing that were divided into four seasons, both for data scenario 1 and data scenario 2, predicting malaria incidence in rainy, transition from rainy to dry, and dry season resulted in reliable performance. It proved by the stable performance wheter while predicting using actual weather data or weather forecast data. Generally, the outbreak prediction in dry season obtained good performance than the other seasons. It indicated that in dry season where the temperature was high the possibility for the outbreak was high. It was related to the life-cycle of the mosquito and the Plasmodium parasite where the warm temperature was good for sporogony process of Plasmodium and mosquito breeding. While the worst performance was prediction of malaria incidence in transition from dry to rainy season, particularly while predicting the incidence using actual weather data. Predicting malaria incidences in this season were predominantly influenced by weather conditions in

dry season. This result indicated that from dry season to transition dry to rain season occurred a fairly significant weather changes that caused the prediction misses. In this season, system tend to predict the outbreak but in reality the outbreak was not always occurred. It might happen because this season affected by two season. The warm temperature from dry season was good for sporogony process of Plasmodium and mosquito breeding, but effect of the rainfall from rainy season might washed the mosquito larvae, so that the outbreak was not occured.

4.3 Summary of Findings

This prediction system has a reliable performance for predicting malaria incidence based on weather factors. This result indicated that there was a sufficient correlation between weather and malaria incidence. The accuracy obtained for predicting malaria incidence in 2008 and 2009 was 75%. This showed that the system could predict 9 from 12 malaria incidences accurately with low MAPE. This system was also pretty good in identifying the "outbreak", it was proven by the F-value that reached 84.21% in 2008 and 40% in 2009. While derived into four seasons, in 2008 the system resulted in best performance while predicting malaria incidence in dry season and in 2009 the system resulted in best performance while predicting malaria incidence in rainy season. The maximum accuracy and F-value for those two seasons reached 100%, and resulted in the lowest MAPE. It means that for predicting malaria incidence in those seasons, the system was not only accurate in predicting the class but also resulted the prediction number that was approximate the actual number.

Compared with prediction system that constructed using BPNN with the same data scenario, structure, and epoch, the performance of prediction system using ENN was better. It was proven by lower MAPE, higher accuracy, and higher F-value. Besides that, ENN reduced trial-and-error process in constructing ANN architecture very significaltly, because by using ENN did not need to try all possible architectures. ENN would find optimum structure to solve the problem through learning process. Therefore, ENN was a really worth effort for constructing NN architecture.