HYPERPARAMETER TUNING METHOD OF EXTREME LEARNING MACHINE (ELM) USING GRIDSEARCHCV IN CLASSIFICATION OF PNEUMONIA IN TODDLERS

Pirjatullah[1], Dwi Kartini[2], Dodon Turianto Nugrahadi[3]
Fakultas Matematika dan Ilmu Alam Universitas Lambung Mangkuratama instansi
Jl. A. Yani Km. 36 Banjarbaru, Kalimantan Selatan, telp. (0511) 4773112
Pirjatullah14@gmail.com

Abstract

Pneumonia is a disease that is susceptible to attack toddlers. According to data from the Ministry of Health, the cause of under-five mortality due to pneumonia is number 2 of all under-five deaths. The dataset used is pneumonia disease data at the MTBS Health Center of East Martapura Health Center. The classification method in this study uses the Extreme Learning Machine (ELM) method. The classification process starts from SMOTE upsampling to balance the class, then parameter tunning is performed using GridsearchCV on the hidden layer neurons, then classification is carried out using the ELM method using the Triangular Basis activation function by comparing the test datasets 90:10, 80:20, 70:30, 60:40 and 50:50. This study provides the best performance results with an accuracy of 86.36%, the ratio of training and test data is 90:10 and 3 neurons hidden layer.

Keywords: *Toddlers, Pneumonia, Hyperparameter Tuning, GridsearchCV, Extreme Learning Machine.*

1. INTRODUCTION

Children under five years (toddlers) are the age vulnerable to disease. At that age, toddlers are very susceptible to disease from an unhealthy environment. According to KemenPPA data, the under-five mortality rate in Indonesia is 32 deaths per 1,000 live births[1]. According to data from the Ministry of Health of the Republic of Indonesia, the cause of under-five mortality due to pneumonia is no. 2 of all under-five deaths (15.5%)[2]. Therefore, we need a method that can be used to recognize the symptoms of Pneumonia and non-Pneumonia.

In conducting the classification, the data is first divided into test data and training data. To get maximum results, the data used must of course use balanced data. If the data used is not balanced, then a balancing method is needed such as undersampling or oversampling. As research conducted by Siringoringo, the SMOTE method as a method to balance the data and the results obtained that SMOTE increases the average G-Mean from 53.4% to 81.0% and the F-Measure average from 38.7% to 81, 8%[3].

Classification techniques carried out using computers can be solved by various methods, one of which is the Extreme Learning Machine (ELM) method. Multazam et al conducted a classification of the types of hepatitis which got an average

accuracy of 80.00% in the ELM method, this accuracy result is better than the KNN and Naïve Bayes methods.[4]. Huang et al mentioned that the input and hidden weight parameters can be chosen randomly so that ELM has a fast learning speed and can produce a good performance. This method has a mathematical model that is simpler and more effective than a feedforward neural network[5].

With biased hidden layer neuron parameters, a search method is needed to determine the optimal number of neurons. One method for tuning parameters is GridSearchCV. Yong Shuai and Huang proved that using parameter tuning using GridSearchCV to optimize support vector engine parameters and build an SVM classification model, proved to provide reliable and more effective optimization recommendations.[6].

Through the problems and explanations that have been mentioned, it can be seen that ELM is one of the algorithms that can be used for classification and can produce fairly good accuracy. The bias parameter in the ELM is also one of the testing factors in this study. Therefore, this study will optimize testing by performing hyperparameter tuning, which in this study will use GridsearchCV and the Triangular Basis activation function for data classification of Pneumonia in Children Under Five Years (Toddler).

2. RESEARCH METHOD

This research stage uses data mining stages. The stages of the research consisted of collecting data obtained from the East Martapura Health Center, preprocessing data, classification, and performance testing using a confusion matrix table. The following stages of the research used are shown in Figure 1.

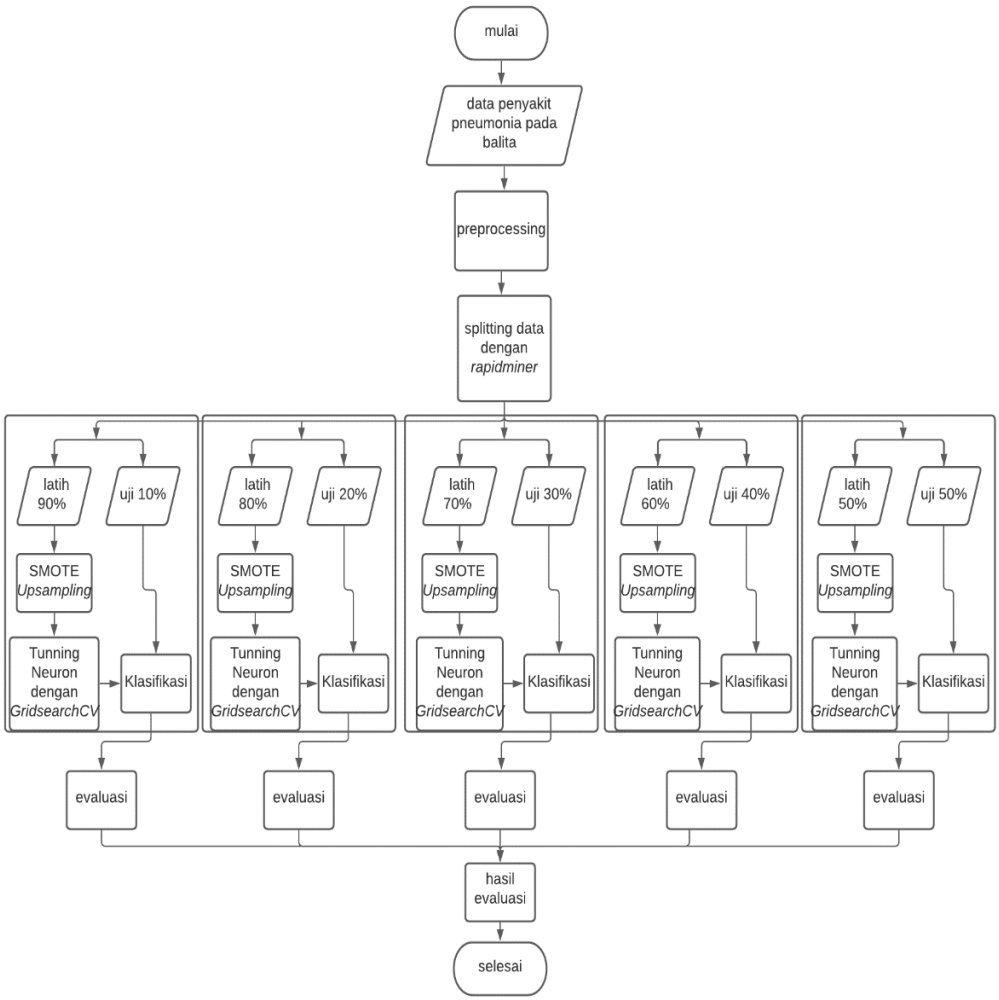

Figure 1. Research Flow

2.1 Data Collection

Collection of Pneumonia and Non-Pneumonia Cough datasets in toddlers obtained at the MTBS polyclinic of the East Martapura Health Center in 2020. The dataset obtained consisted of 217 disease data and 13 attributes. With a total data of 167 Non-Pneumonia Cough and 50 Pneumonia patients. The details of the dataset attributes can be seen in Table 1.

Table 1. Attributes of Patient Data

Variable	Information
Age	Toddler Age (Months)
Gender	L/P
Weight	Toddler Weight (Kg)

Height	Toddler Height / Length (Cm)
Body temperature	Toddler Body Temperature (°C)
Breath/second	Toddler Breath Count Results (Seconds)
Fever	Yes/No
Cough	Yes/No
Cold	Yes/No
Diarrhea	Yes/No
Itchy	Yes/No
Other Problems/Complaints	Additional symptoms or other complaints
Disease (Class)	Disease decision class

The following sample data is shown in Table 2.

Table 2. Data on Pneumonia in Toddlers

Number	Patient Number	New/Old Patient	Age	Gender	Weight	...	Disease
1	3057	L	4,5 TH	P	16	...	Cough Not Pneumonia
2	994	B	1,40 TH	L	8,8	...	Pneumonia
3	525	L	8,5 BLN	P	7,8	...	Cough Not Pneumonia
4	353	L	3,3 TH	L	11	...	Cough Not Pneumonia
5	175	L	11,5 BLN	P	7,2	...	Cough Not Pneumonia
6	5795	B	4,3 TH	L	13	...	Cough Not Pneumonia
7	34	L	4,2 TH	L	22	...	Cough Not Pneumonia

8	34	L	2,4 TH	P	10	…	Cough Not Pneumonia
9	3123	L	6 BLN	L	8,7	…	Cough Not Pneumonia
…	…	…	…	…	…	…	…
217	6010	L	1 TH	P	7,3	…	Pneumonia

2.2 Preprocessing

At this stage, the missing value data is deleted and preparation is changing the string data to numeric data. For gender, it is labeled 0 and 1 for descriptions of male and female, and on the symptom row, it is labeled 0 and 1 for statements of yes and no. Then the data were normalized using the Minmax Scaler. This is done to make it easier for the system to process data for classification processing. The sample data is shown in Table 3.

Tabel 3. Data Preprocessing

Number	Age	Gender	Weight	Height	Body Temperature	Breath/second	…	Disease
1	54	0	16	107	38,2	20	…	0
2	16,8	1	8,8	60	34	53	…	1
3	8,5	0	7,8	68	36,5	32	…	0
4	39,6	1	11	95	35	35	…	0
5	11,5	0	7,2	69	35	40	…	0
6	51,6	1	13	94	35,4	28	…	0
7	50,4	1	22	108,5	36,5	28	…	0
8	28,8	0	10	77	35,9	32	…	0
9	6	1	8,7	78	36,3	48	…	0
…	…	…	…	…	…	…	…	…
217	12	0	7,3	69	36	50	…	1

From the table above, the dataset used consists of can be seen in table 4.

Table 4. Total Disease Data

Variable	Information
Cough Not Pneumonia	167
Pneumonia	50
Total	217

2.3 Data Mining Model

In this stage, dataset distribution and data sampling are carried out, namely SMOTE upsampling on the dataset used, because there is a class imbalance between the majority class and the minority class. In this stage, parameter tuning is also carried out using GridsearchCV on the training data to determine the number of hidden layer neurons before entering the classification stage.

2.4 Classification

At this stage, the classification uses the Extreme Learning Machine method. By testing the activation function of the triangular basis and using the number of hidden layer neurons that have been determined by the previous GridsearchCV method.

2.5 Evaluation

The last stage is testing the performance of the ELM method in classifying Pneumonia in Toddlers using a confusion matrix table. The confusion matrix method in this study is to test the accuracy. The confusion matrix table can be seen in Table 5.

Table 5. Table Confusion Matrix

		True Condition	
	Total Population	*Condition Positive* (Cough Class Not Pneumonia)	*Condition Negative* (Pneumonia Class)
Predicted Condition	*Prediction Condition Positive* (Cough Class Not Pneumonia)	*True Positive* (TP)	*False Positive* (FP)
	Predicted Condition Negative (Pneumonia Class)	*False Negative* (FN)	*True Negative* (TN)

The formula used in calculating accuracy (1) is as follows:

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

Information:

1. True Positive (TP) is a class worth predicting correctly.
2. False Positive (FP) is a class that does not deserve to be predicted wrong.
3. True Negative (TN) is a class that does not deserve to be predicted correctly.
4. While False Negative (FN) is a class that deserves to be predicted wrong.

3. RESULTS AND ANALYSIS

3.1 Results

Before classifying the data, the data is divided into two, namely training data

and test data. The distribution of data in this study uses *rapidminer* tools. Divide the data into, 90:10, 80:20, 70:30, 60:40 and 50:50. The training data will be used as processing in machine learning to recognize patterns from the data and test data will be used as a result of learning from the training data.

a. Dataset Balancing

Due to the data imbalance, the SMOTE (Synthetic Minority Oversampling Technique) process was carried out to balance the data. Data balancing is only done on training data as processing in machine learning to recognize data. Before entering the classification stage, the data will be oversampled, namely to balance the minor data with the major data, which is to balance the pneumonia class data (minor) with the non-pneumonia cough class (major). After the data is balanced, the amount of data between classes will be the same or have the same amount of data. The dataset ratio can be seen in Table 6.

Table 6. Dataset Class Ratio

Number	**Sampling Method**	**Number of Instances in Training data**		**Number of Instances on Test data**		***Dataset***
		Cough Not Pneumonia	Pneumonia	Cough Not Pneumonia	Pneumonia	
1	Original	150	45	17	5	90:10
2	SMOTE	150	150	-	-	
3	Original	134	40	33	10	80:20
4	SMOTE	134	134	-	-	
5	Original	117	35	50	15	70:30
6	SMOTE	117	117	-	-	
7	Original	100	30	67	20	60:40
8	SMOTE	100	100	-	-	
9	Original	84	25	83	25	50:50
10	SMOTE	84	84	-	-	

b. Comparison of the ratio of training data and test data

The comparison test of the ratio of training data and test data was carried out to determine the effect of the ratio of training data and test data on the accuracy results. This test was carried out with the training data and test data in each ratio determined, namely the ratios of 90%: 10%, 80%: 20%, 70%: 30%, 60%: 40%, and 50%: 50%. This test is done by using Triangular Basis activation. The results of the comparison test of the ratio of training data and test data can be seen in Figure 2.

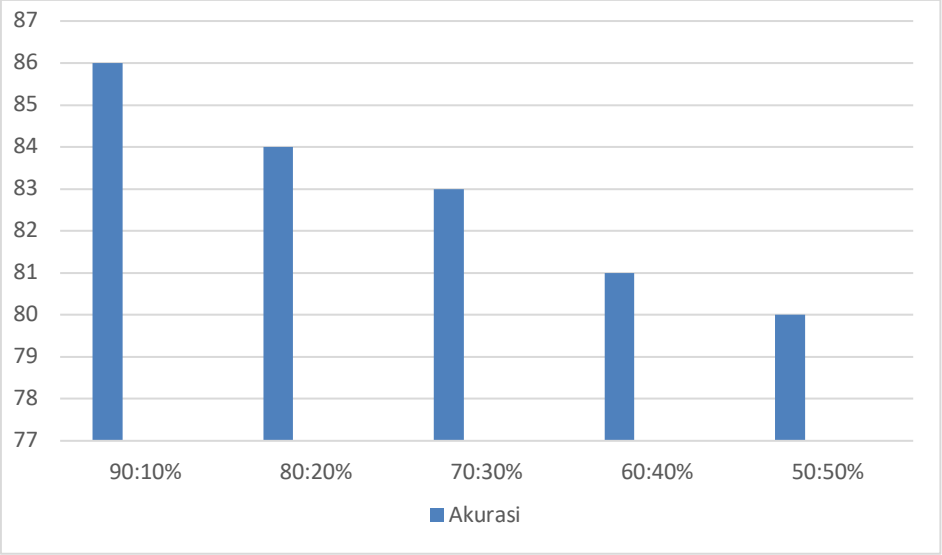

Figure 2. Comparison Testing of Training Data Ratios and Test Data

Figure 2. Testing Based on the test results that have been presented in Figure 2, it can be seen that the results of the comparison test of the ratio of training data and test data produce fairly good accuracy. The best accuracy results produced using the ratio of training data and test data with a comparison of 90:10 with an average accuracy value of 86.36%. Comparison of the Ratio of Training Data and Test Data

c. Performance evaluation

Performance testing is done to find out how good the level of accuracy is using the confusion matrix table. An Evaluation was carried out on the activation function of the Triangular Basis with 90:10 data sharing. The best prediction results from the classification using the architecture obtained from previous tests can be seen from the Confusion Matrix table in table 7.

Table 7. Best Prediction Results Method

Confusion matrix		**Actual**	
		TP	FP
Prediction	FN	17	0
	TN	3	2

The confusion matrix table presented in table 6, there are 17 data with disease class Non-Pneumonia Cough and no Pneumonia disease class that is predicted as Non-Pneumonia Cough disease. And there are 3 class data for Non-Pneumonia Cough and 2 Pneumonia class data which are predicted as Non-Pneumonia Cough disease class. Based on the confusion matrix table, classification performance evaluation is carried out in the form of accuracy with calculations as in table 8.

Table 8. Calculation of the accuracy of the best prediction results using the ELM method

Calculation Accuracy	
Formula	Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ x 100%
Result (%)	$\frac{17+2}{17+0+3+2}$ x 100 = 86.36%

In the table above, it is known that the evaluation of the classification of the Extreme Learning Machine method using the confusion matrix produces an accuracy of 86.36%.

3.2 Analysis

Classification is done by testing the Extreme Learning Machine method and testing the 90:10, 80:20, 70:30, 60:40, and 50:50 datasets. The performance sought is the accuracy of the classification into the compared variables. The data used is pneumonia in children under five obtained from the East Martapura Health Center with a periode of January to June 2020. The amount of data used is 217 with the data used being unbalanced between classes, so this stage is carried out by balancing the data due to data imbalances.

The data balancing technique uses SMOTE (Synthetic Minority Oversampling Technique) Upsampling to make the number of minority data equal to the majority data. In this study, hyperparameter tuning was also carried out, namely testing the hidden layer neurons using *gridsearhcv* to determine the optimal hidden layer neurons. Optimal hidden layer neurons are tested for predictive data using the confusion matrix in the ELM method so that the performance of the overall prediction results in each training and test data distribution is obtained, as well as the optimal number of hidden layer neurons from each data division. The best accuracy performance of each dataset on the classification of pneumonia in toddlers using the Extreme Learning Machine method produces an accuracy value of 86% at 90:10, 84% at 80:20, 83% at 70:30, 81% at 60:40, and 80% at 50:50

4. CONCLUSION

Based on the test results, the use of hyperparameter tuning on the Extreme Learning Machine (ELM) method can classify pneumonia in toddlers with good accuracy results. The best accuracy performance is 86.36%, with the number of hidden layer neurons as many as 3 neurons, and using a ratio of training and test data of 90:10%. Therefore, the use of the GridsearchCV hyperparameter tuning to determine the optimal hidden layer neurons in the Extreme Learning Machine method can be used to determine the classification of Pneumonia in Toddlers, because it has a good performance based on accuracy.

REFERENCES

[1] Kemenppa RI, *Profil Kesehatan Anak Indonesia Tahun 2018*, vol. 5, no. 1. 2018.

[2] Kemenkes RI, *Buletin Jendela Epidemilogi Pneumonia Balita*, vol. 3. 2010.

[3] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[4] S. Multazam, I. Cholissodin, and S. Adinugroho, "Implementasi Metode

Extreme Learning Machine pada Klasifikasi Jenis Penyakit Hepatitis berdasarkan Faktor Gejala," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 3, pp. 789–797, 2020.

[5] G. Bin Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," *IEEE Int. Conf. Neural Networks - Conf. Proc.*, vol. 2, pp. 985–990, 2004.

[6] Y. Shuai, Y. Zheng, and H. Huang, "Hybrid Software Obsolescence Evaluation Model Based on PCA-SVM-GridSearchCV," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2018-November, pp. 449–453, 2019.