International Journal on Electrical Engineering and Informatics - Volume 14, Number 1, March 2022

Dataset Feasibility Analysis Method based on Enhanced Adaptive LMS method with Min-max Normalization and Fuzzy Intuitive Sets

Sri Arttini Dwi Prasetyowati[1], Munaf Ismail[1], Eka Nuryanto Budisusila[1], De Rosal Ignatius Moses Setiadi[2], Mauridhi Hery Purnomo[3]

[1]Universitas Islam Sultan Agung Semarang, Indonesia
[2]Dian Nuswantoro University, Semarang, Indonesia
[3]Institut Teknologi Sepuluh Nopember Surabaya, Indonesia
arttini@unissula.ac.id, munaf@unissula.ac.id, ekanbs@unissula.ac.id, moses@dsn.dinus.ac.id, hery@ee.its.ac.id
*Corresponding Author: arttini@unissula.ac.id

Abstract: A good dataset was required for attaining good accuracy in machine learning, especially in prediction, so that prediction accuracy was high. The imbalanced or too small dataset was a common problem in machine learning. This study proposed a method for determining the dataset's quality. If the dataset is not feasible, preprocessing can be performed to improve the dataset's quality before making predictions. Adaptive Least Mean Square (LMS) was merged with Min-max Normalization and Fuzzy Intuitive Sets (FIS) algorithms to create the proposed technique. This method might assess the value of uncertainty and information, which will influence the dataset's feasibility. If the dataset has an uncertainty value closed 1.5 and an information value of less than 0.5, it is usable. The method has been tested on both public and private datasets. According to all experiments conducted, the uncertainty value and information value on the stated threshold can have prediction accuracy above 70% with various methodologies.

Keywords: Min-max Normalization; LMS Adaptive; Fuzzy Intuitive Sets; Uncertainty; Information

1. Introduction

A good dataset, one that is normal, balanced, and contains enough records for the learning process, was required to produce accurate prediction results. A faulty dataset has a significant impact on the overall prediction process. However, obtaining a solid dataset was sometimes challenging due to factors such as minor disjuncts, noise, and overlapping data. [1]. In addition, a lot of predictive research only focused on improving the method and did not check the dataset feasibility. One of the biggest problems in prediction was imbalanced data, because most algorithms assumed a balance class in their design [2]. In imbalanced data, the prediction accuracy could be very low, so it is necessary to validate the dataset to determine the feasibility of the dataset used to make predictions.

Dataset feasibility analysis means analyze the feasibility of the dataset that will be used for prediction process, whether the dataset is feasible for processing or not. The analysis is carried out using Least Mean Square Predictive, where the data has been normalized before. The results of the least mean square predictive are then used as input for the fuzzy intuitive set, and the uncertainty values and information values are analyzed.

A prediction is strongly influenced by the feasibility of the dataset. If the dataset is bad, the prediction will be bad. With the dataset feasibility test, at least it can be determined the next steps that must be taken by the decision maker. For example, if there is an unbalanced dataset, it must be balanced first, or in other words, do preprocessing first so that the dataset becomes better.
Research on checking the quality of the dataset has never been found, especially in prediction. Most statistical analysis on datasets was about analyze the data and determine the type of distribution be Gaussian distribution, uniform distribution, or another distribution [3][4]. The importance of this study is to minimize the use of not feasible data in decision making. If the

Received: October 27th, 2021. Accepted: March 23rd, 2022
DOI: 10.15676/ijeei.2022.14.1.4

only data is data that is not feasible, then some of these analyzes can be used to consider preprocessing.

Adaptive LMS was an algorithm widely used for prediction, identification or recognition of unknown systems, filter configuration, and noise canceling configuration with two inputs. The advantage was simple in computation, while the disadvantage was long in iterations. Adaptive LMS was a robust algorithm, because the algorithm had a simple computation and sooner or later in achieving the target depended on the step size taken. Predicting imbalanced data was a big challenge [5]. However, processing imbalanced data directly with an adaptive LMS would cause errors in the calculations. In this condition, min-max normalization was required. Adaptive LMS was performed after the imbalanced data was processed using the min-max normalization algorithm.

According to research, the fundamental LMS algorithm proved to be highly effective for weak and monotonous ECG readings. However, in the presence of interference, the LMS algorithm has to be adjusted to reduce interference in the ECG signal, resulting in a more accurate signal reading [6]. Other studies found that the adaptive LMS algorithm was fairly effective to use for random or monotonous data, but that it needed to be modified to improve the outcomes even further. [7]. However, the adaptive LMS algorithm can work optimally. Under changing target conditions, this algorithm was quite strong to reach the target because of its simple characteristics, and the pace of the target achievement depends on the value of the step size [8]. Some of these studies had confirmed that the adaptive LMS algorithm was suitable for use in data that has irregularities.

This study aimed to exclude the imbalanced data that was unprocessable due to high uncertainty and low information. It would be a waste of time to continue study with such data. As a result, the Adaptive LMS algorithm was chosen, which is simple to calculate but effective in terms of results. After the data has been normalized with the Min-max Normalization algorithm, the Adaptive LMS algorithm is utilized to calculate the prediction error value. In order to calculate the estimate of uncertainty, the prediction error was used as a benchmark. After that, as a component of the FIS process.

This research took data from the KEEL repository [9], which consisted of data Ecoli0_vs_1, Glass 0, Hamberman and Pima. Many calculation errors appeared when the data were processed using the LMS adaptive algorithm. The data were still in an imbalanced data condition and must be processed into balanced data. After becoming a balance data, the problem was whether the four data could be processed for decision making required by the authorized parties or not, the amount of uncertainty in the data, and the amount of information that can be obtained from the data. Perhaps there were invalid data in one of them and should be excluded. This study used min-max normalization to balance out the imbalanced dataset. Min-max normalization was a straightforward normalization method that used only the maximum and minimum variables. The normalized could be determined using this way. The LMS adaptive method then process the normalized data.. The LMS algorithm was chosen as the major reference because, despite having many iterations, it was straightforward to compute. Therefore, it is easy to modify the weight changes. When compared to real data, it appeared that the prediction results still had an error. An accountable analysis was required to determine whether to accept data with minor errors or to reject data with major errors. The value of uncertainty and the contribution of information provided from the data that would be eliminated were determined using a Fuzzy Intuitive Sets (FIS) analysis [10][11].

The novelty of this research was analyzing the feasibility of the datasets to be used in prediction or other data processing using an enhanced adaptive LMS method with min-max normalization and fuzzy intuitive sets so that a decision maker was not wrong in choosing valid data, especially for imbalanced data. This paper also gave the solution if the only data was the data that was not feasible

2. Related research

Previous research related to imbalanced data only discussed classification problems and techniques for handling data. Imbalanced classification problem have become one of the challenges in data mining community, and have been widely studied in recent years, due to their complexities and huge impacts in real-world applications [12]. It introduced two kinds of decision tree ensembles for imbalanced classification problems, extensively utilizing properties of α-divergence. First, a novel splitting criterion based on α-divergence was shown to generalize several well-known splitting criteria such as those used in C4.5 and CART. When the α-divergence splitting criterion was applied to imbalanced data, one can obtain decision trees that tend to be less correlated (α-diversification) by varying the value of α. This increased diversity in an ensemble of such trees improves AUROC values across a range of minority class priors. The second ensemble uses the same alpha trees as base classifiers but uses a lift-aware stopping criterion during tree growth. The resultant ensemble produces a set of interpretable rules that provide higher lift values for a given coverage, a property that is much desirable in applications such as direct marketing. Experimental results across many class-imbalanced datasets, including BRFSS, and MIMIC datasets from the medical community and several sets from UCI and KEEL are provided to highlight the effectiveness of the proposed ensembles over a wide range of data distributions and of class imbalance.

Imbalanced data, defined as a significant discrepancy in observation frequency between classes, has got much attention in data mining research. Because most classifiers assume the class distribution is balanced or the penalties for different types of classification errors are equal, prediction performance frequently degrades as classifiers learn from unbalanced data. The researchers suggested a new framework called model-based synthetic sampling (MBS), which combined modeling and sampling techniques to generate synthetic data to deal with imbalanced situations. The results of the experiments showed that the proposed strategy was not only comparable but also stable. It also included thorough examinations and visualizations of the proposed strategy to empirically show why it may provide good data samples. [13].

Another research related to imbalanced data discussed the pre-processing technique for handling imbalanced data using k nearest neighbor or kNN [12]. Under sampling method in dealing with class-imbalanced problems, this method only employed a subset of the majority class and thus was very efficient. The main deficiency was that many majority class examples are ignored [14], and many more research that was essentially about the problem of classification and technique for handling imbalanced data.

Research that discussed the validity of imbalanced data so that it was feasible to be processed has not existed in previous studies. This research was intended to develop a method to determine the validity of imbalanced data so that decision makers do not use the wrong data for processing. This method combined min-max normalization, adaptive LMS and intuitive fuzzy sets. Fuzzy Intuitive Sets is required to determine the value of uncertainty and information of the imbalanced data. This was to ensure that the data was classified as valid data or not. Some research that used fuzzy intuitive sets in its method can be mentioned as follows.

Methods of calculating the value of uncertainty and information had begun to emerge. Some of them were studied using patient data to diagnose a disease suffered by a patient [10].
Discussion about fuzzy soft set decision problems and a new algorithm based on grey relational analysis is presented. The evaluation bases of the new algorithm were multiple. There was more information in a decision result based on multiple evaluation bases, which was more easily accepted and logical to one's thinking [15].

Another research was the needs for the application of a multicriteria approach to solve problems in which solution consequences cannot be estimated on the basis of a single criterion, it involves the necessity of analyzing a vector of criteria, and problems that may be considered on the basis of a single criterion but their unique solutions are not achieved because the uncertainty of information produces so-called decision uncertainty regions. (Ekel, 2002a). The paper shows the inconsistent intuitionistic fuzzy sets, picture fuzzy sets and neutrosophic fuzzy sets are representable by Interval-valued intuitionistic fuzzy sets, which themselves are representable by an ordered pair of the standard Intuitionistic fuzzy sets.[16]. Intuitionistic fuzzy interpretations of the processes of multi-person and of multi-measurement tool were discussed in this article. Multi-criteria decision makings are also discussed in this paper.[17], Linear regression analysis in an intuitionistic fuzzy environment using intuitionistic fuzzy linear models with symmetric triangular tuitionistic fuzzy number (STriIFN) coefficients was introduced in this article. [18]. Table 1 shows the list of previous research related to this research.

Table 1. Related Reseach

No	Author	**FIS**	**Min-Max**	**Adaptive LMS**	**Description**	Constribution
1	Kulicka [10]	Y	N	N	FIS for decision-making based on one main alternative	logic of thought
2	P. Y. Ekel [19]	Y	N	N	Application of fuzzy sets for a multicriteria approach	Comparative analysis for the equation
3	Kong et al [15]	N	N	Y	Application of fuzzy soft set in decision-making based on grey theory	Comparative analysis for the method
4	K, Atanassov, Vassilev[16]	Y	N	N	Showing the inconsistent intuitionistic fuzzy sets	Comparative analysis
5	K. Atanassov et. al [17]	Y	N	N	Intuitionistic fuzzy interpretations of the processes of multi-person and of multi-measurement tool	logic of thought, Additional insight
6	Parvathi et. al [18]	Y	N	N	Linear regression analysis in an intuitionistic fuzzy environment using intuitionistic fuzzy linear models	Additional insight
7	Y.Park, J. Ghosh [12]	N	N	N	about classification problems and technique for handling Imbalanced data	Additional insight
8	C.L. Liu, P. Hsieh, 2020 [13]	N	N	N	Explain about Imbalanced data was characterized by the severe difference in observation frequency between classes and proposed a novel framework called model-based synthetic sampling (MBS) to cope with imbalanced problems	Comparative analysis and additional insight
9	P. Nair, I. Kashyap, 2019[20]	N	N	N	pre-processing technique for handling imbalanced data using k nearest neigbour or kNN	Comparative analysis and additional insight
10	X.Y. Liu, J.Wu, Z. Zhou, 2009 [14]	N	N	N	Research about the main deficiency was that many majority class examples are ignored	Comparative analysis and additional insight

3. Proposed Scheme

The LMS Adaptive Algorithm had frequently been modified in previous studies to obtain the desired result. Min-max normalization was used to modify the adaptive LMS method in this study. Even the prediction error findings from the adaptive LMS Algorithm had to be evaluated using FIS before a judgment could be made.

Pseudocode of the proposed method was shown in Code Snippet 1 and the entire step can be described in the flowchart in Fig. 1. It begins with Imbalanced data input from KEEL repository [12], consisting of data Ecoli0_vs_1, Glass0, Haberman, and Pima, then normalized with min-max normalization data became balanced data. After that, the adaptive LMS process was executed and the errors were noted. The error was then used to estimate the α value to start the FIS process.

```
Input: KEEL repository = X[n]
Initialization:
        Desired output = d[n]
        Weight value = W[n]
        Output = Y[n]
        Error LMS adaptive = e[n]
Do, while last X[n]
        Count Min_max normalization
        X[n]_norm ← X[n]
        LMS Adaptive
           Input: X[n]_norm = d[n]
           X[n]_norm delayed = X[n}
           Do, while last X[n]
                Count W[n]
                Count Y[n]
                Count e[n]
           Count Uncertainty
           Count Information value
        Error_ X[n]_norm ← X[n]_norm
        Count uncertainty and information
           value
Conclusion
```

Code Snippet 1. Pseudocode of Proposed method

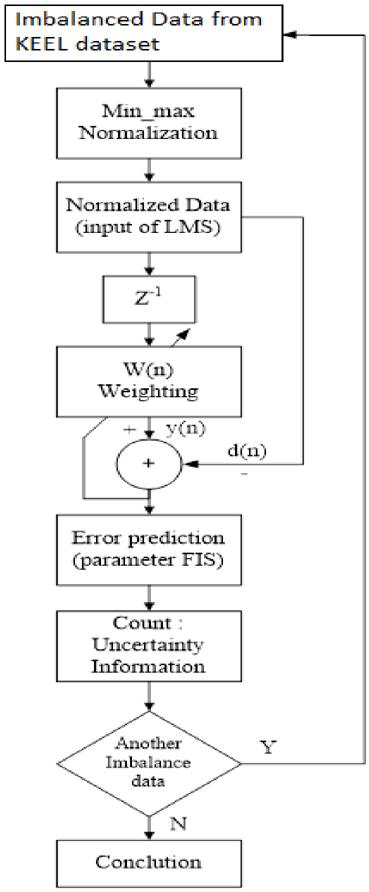

Figure 1. Proposed combined scheme

In detail, the steps proposed were as follows:

1. The input must be more than two imbalanced data because this process was used to select data to be excluded.
2. The normalization using min-max algorithm was carried out to obtain data in the certain range, see section 3.1 for more details.
3. Normalized data was used as input for adaptive LMS algorithm.
4. Prediction with Adaptive LMS was implemented, where input for LMS were normalized data and delayed normalized data see section 3.2 for more details.

5. The error prediction was the difference between the prediction results and the desired output, it was expected as small as possible.
6. The biggest error prediction will be a parameter in FIS to calculate uncertainty and its information
7. Count the uncertainty and its Information. See 3.3 for more details.
8. The process can be repeated for other imbalanced data

A. Min-max Normalization

The min-max algorithm looks for new data based on the maximum and minimum value. Eq. 1 shows the min-max algorithm formula.

$$v' = \frac{v - \min}{\max - \min}(New\_max - New\_min)\ \_\ New\_min \quad (1)$$

Where v' is output, v is input, max and min are the maximum and minimum values of data, respectively. New_max and New_min are the newly defined maximum and minimum values.

B. LMS Adaptive

The block diagram of the adaptive predictive least mean square (LMS) was shown in the block diagram in Fig. 2.

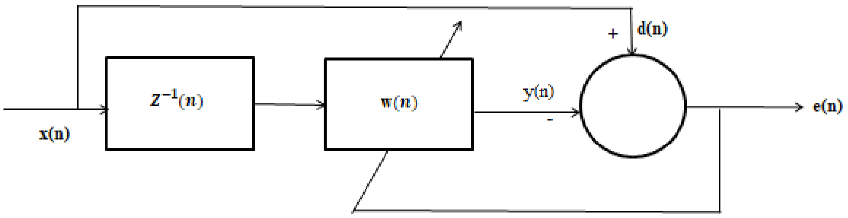

Figure 2. Adaptive LMS Block Diagram

The values of d and x are targets and inputs, respectively. Both are the same data, but input x is the target that is delayed by one sample [7] [8] [21].

Least Mean Square Adaptive algorithm was used to solve linear estimation problems such as the one shown Fig. 3, where the input vector $x_k = [x_{1k} x_{2k} \cdots x_{Lk}]^T$ and desired response $d_k \in R$ are jointly stationary random processes. The equations for weight, output and error are shown in equations (2), (3). and (4) below.

The weight vector w_k :

$$w_k = [w_{1k} w_{2k} \cdots w_{Lk}]^T \quad (2)$$

Output $y_k = x_k^T w_k \quad (3)$

and error $\varepsilon_k = d_k - y_k \quad (4)$

The Mean Square Error (MSE) is defined as

$$\xi k = E[\varepsilon_k{}^2] \quad (5)$$

and it is a quadratic function of the weight vector. The optimal weight vector that minimizes ξk is given by $w* = R^{-1}p \quad (6)$

where $R = E[x_k x_k^T] \quad (7)$

Is the input autocorrelation matrix (assumed to be full rank), and

$$p = E[x_k d_k] \quad (8)$$

is the cross-correlation vector. The minimum MSE (MMSE) was obtained using $w*$ which is denoted by $\xi*$.

Often in practice, $R^{-1}p$ cannot be calculated due to the lack of knowledge of the statistics R and p. However, when samples of x_k and d_k are available, they can be used to iteratively adjust the weight vector to obtain an approximation of $w*$. The simplest and most widely used algorithm for this is LMS. It performs instantaneous gradient descent adaptation of the weight vector:

$w_{k+1} = w_k + 2\mu\varepsilon_k x_k$. (9)

The parameter of step size μ and the initial weight vector w_0 is arbitrarily set by the user. The MSE sequence ξ_k corresponding to the sequence of adapted weight vectors w_k is commonly known as the learning curve. The adaptive linear combiner is shown in Fig. 10 [5],[22],[21],[23].

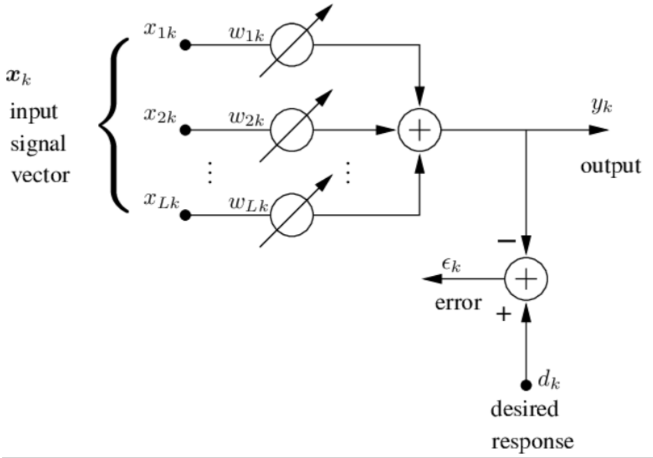

Figure 3. Adaptive Linear Combiner

C. Fuzzy Intuitive Sets (FIS)

The simplest model of decision-making assumes just one of two relevant variants. There was α and β. The fuzzy set is:

$$\underline{A} = \{\alpha/\mu(\alpha); \beta/\mu(\beta), \mu(\alpha) + \mu(\beta) \leq 1\}. \quad (10)$$

Where rate of plausibility $\mu(\alpha), \mu(\beta) \in [0,1]$.

The uncertainty for fuzzy set (10) can be estimated by the following equation (11) with the choice of independent variations

$$H(\underline{A}) = \min\{\mu(\alpha); 1 - \mu(\alpha)\} + \min\{\mu(\beta); 1 - \mu(\beta)\} \quad (11)$$

The information $I(\alpha)$ can be determined from the following relationship:

$$I(\alpha) = 1 - H(\underline{A}) \quad (12)$$

$$I^s(\alpha) = \begin{cases} -I(\alpha), for\ 0 \leq \mu(\alpha) \leq 0{,}5 \\ I(\alpha), for\ 0{,}5 \leq \mu(\alpha) \leq 1 \end{cases} \quad (13)$$

Of course, $-1 \leq I^s(\alpha) \leq 1$. Then for uncertainty modeled by eq.10 under **the** condition

$$\mu(\alpha) + \mu(\beta) = 1 \quad (14)$$

can be estimated: $H(\underline{A})(\alpha) = -\mu(\alpha).log_2\mu(\alpha) - (1 - \mu(\alpha)).log_2(1 - \mu(\alpha))$ (15)

Then $0 \leq H(\underline{A}) \leq 1$. From eq. (11) dan eq. (14), another simple expression for the uncertainty $H(\underline{A})$ became: $H(\underline{A}) = 2\min\{\mu(\alpha); 1 - \mu(\alpha)\};\ 0 \leq H(\underline{A}) \leq 1$. (16)

When the decision-maker has a counter selecting bias towards some of the variants or one does not know how to choose. It can decide in this case to apply the use of intuitionist fuzzy set in order to characterize the decision-making as in the form:

$$\mathcal{F} = \{\alpha/((\mu(\alpha); v(\alpha)); \beta//((\mu(\beta); v(\beta))\} \quad (17)$$

where $\mu(\alpha), v(\alpha)\ \epsilon\ [0,1]$ and $\mu(\beta), v(\beta)\ \epsilon\ [0,1]$ and $\mu(\alpha) + \mu(\beta) \leq 1$ both α and β.
Deciding about the variant α and also according about the variant β is thus divided into three aspects:

1. Assessing the degree of acceptance of the variant α (estimated $\mu(\alpha)$)
2. Assessing the degree of disacceptance of the variant α (estimated $\mu(\alpha)$)
3. Assessing the degree of indecision for some variant of the α (defined by 1- $\mu(\alpha)$-$v(\alpha)$).

Corresponding fuzzy sets to the above types of decisions are:

$\underline{A}_1(\alpha) = \{\alpha/\mu(\alpha); \neg\alpha/(1 - \mu(\alpha))\}$;

$\underline{A}_2(\alpha) = \{\alpha/v(\alpha); \neg\alpha/(1 - v(\alpha))\}$;
$\underline{A}_3(\alpha) = \{\alpha/\pi(\alpha); \neg\alpha/(1 - \pi(\alpha))\}$

(18)
Where $\pi(\alpha) = 1 - \mu(\alpha) - v(\alpha)$

(19)
and the uncertainties $H(\underline{A}_1(\alpha))$, $H(\underline{A}_2(\alpha)), H(\underline{A}_3(\alpha))$ determined according to (15) or (16).
From (17) $\mathcal{F}$ can be introduced the uncertainty $H(\alpha)$.
$H(\alpha) = -(\mu(\alpha).\, log_2\mu(\alpha) + v(\alpha).\, log_2 v(\alpha) + \pi(\alpha).\, log_2\pi(\alpha))$

(20)
And then $0 \leq H(\alpha) \leq log_2 3$

(21)
For uncertainty $H(\alpha)$ and uncertainty $H(\underline{A}_1(\alpha))$, $H(\underline{A}_2(\alpha)), H(\underline{A}_3(\alpha))$:
$H(\underline{A}_1(\alpha)) \leq\ H(\alpha);\ H(\underline{A}_2(\alpha)) \leq\ H(\alpha);\ H(\underline{A}_3(\alpha)) \leq\ H(\alpha)$

(22)
$I_\alpha(\mathcal{F})$=$log_2 3 -\ H(\alpha)$

(23)
$I_\alpha(\mathcal{F})_{\text{norm}}$ = 1- $H(\alpha)/log_2 3$

(24)
Corresponding relationship for the semantisation:
$I_\alpha \mathcal{F}^s_{norm} \begin{cases} -I_\alpha(\mathcal{F})_{norm}\,; \mu(\alpha) \leq 0{,}5 \\ I_\alpha(\mathcal{F})_{norm}\,; 0{,}5 \leq \mu(\alpha) \leq 1 \end{cases}$

(25)
where $-j \leq I_\alpha \mathcal{F}^s_{norm} \leq 1$

$H(\mathcal{F})(\alpha) = H(\underline{A}_1(\alpha))$+ $H(\underline{A}_2(\alpha)) +\ H(\underline{A}_3(\alpha))$

(26)
and $0 \leq H(\mathcal{F})(\alpha) \leq 3$ so that $I(\mathcal{F})(\alpha) = 3 -\ H(\mathcal{F})(\alpha)$

$$I_\alpha(\mathcal{F})_{norm} = 1 - H(\mathcal{F})(\alpha)/3$$

(27)

Choose $I(\mathcal{F}^s(\alpha) = \begin{cases} -I(\mathcal{F})_{norm}\,; \mu(\alpha) \leq 0{,}5 \\ I(\mathcal{F})_{norm}\,; 0{,}5 \leq \mu(\alpha) \leq 1 \end{cases}$
and Eq.22 , Eq 23 and Eq 24. [10]

4. Experimental results and analysis

This section described the details of the experiments performed to evaluate the capabilities of the proposed method and then compared it to other results from the journal that have calculated the accuracy of each dataset first.

The dataset was taken from the KEEL repository, where there were four data that were considered representative good, medium, and poor accuracy. The four data were Ecoli0_vs_1, Glass0, Haberman, and Pima datasets.

4.1. The Results of the Min-Max Algorithm

Fig. 4 showed the results of the Min-max normalization algorithm for Ecoli0_vs_1, Glass0, Haberman, and Pima datasets. The pattern before and after min-max was similar; only the scale changed. This showed that the normalization was performed correctly.

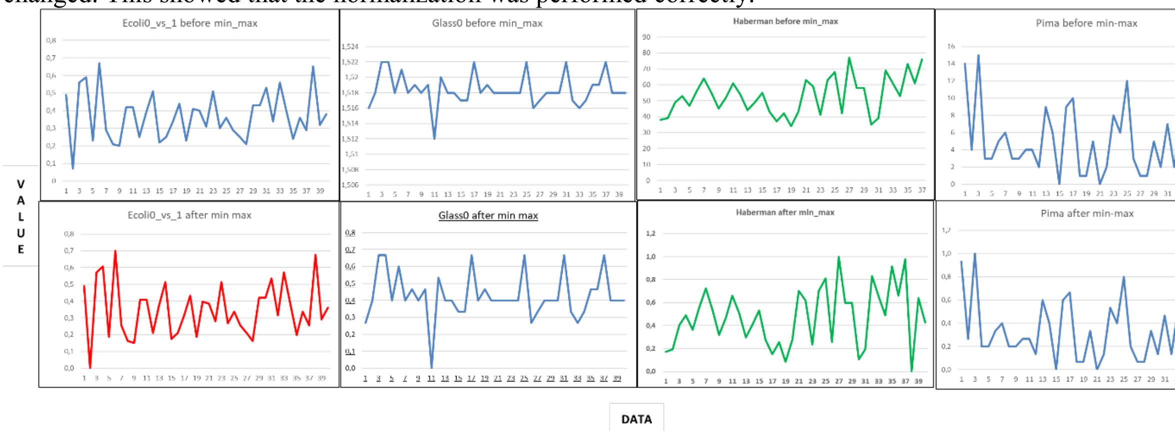

Figure 4. Ecoli0_vs_1, Glass0, Haberman, and Pima datasets before and after min-max normalization

Next, predictions were made to observe which dataset was closest to the real data and how significant the predictive error was.

4.2 The Result of LMS Adaptive

An Adaptive LMS process was carried out for normalized data in Fig.4. Meanwhile, Fig. 5 shows Ecoli0_vs_1 after normalization and LMS predictive.

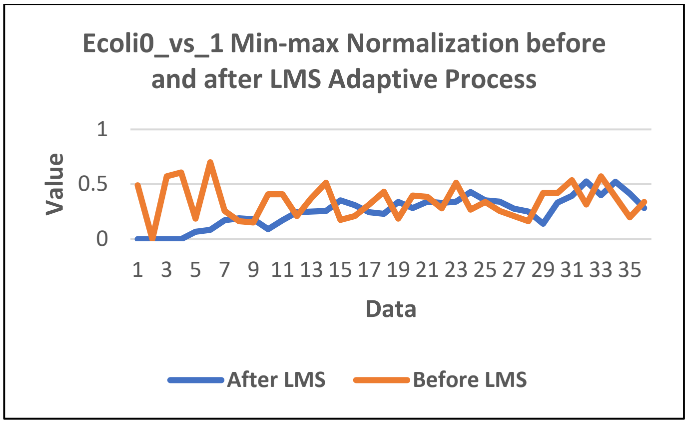

Figure 5 Normalization Results (blue) and Adaptive LMS Results (red) for Ecoli0_vs_1

Even if the normalized data and the predicted normalization data appeared to be similar in the 26th data, the error analysis must still be performed to ensure the data was valid. Glass0, Haberman, and Pima data were all subjected to the same procedure. Figures 6 through 8 illustrate the results.

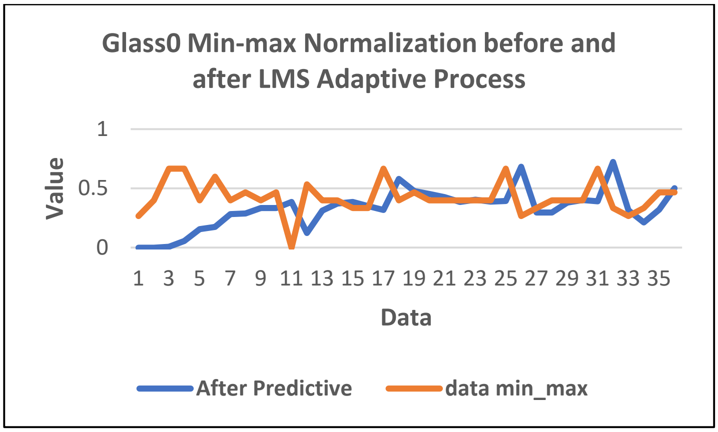

Figure 6. Normalization Results (blue) and Adaptive LMS Results (red) for glass0 data.

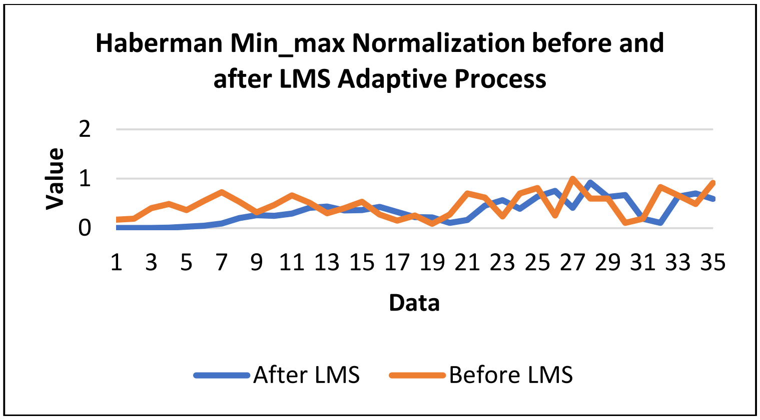

Figure 7. Normalization Results (blue) and Adaptive LMS Results (red) for Haberman data.

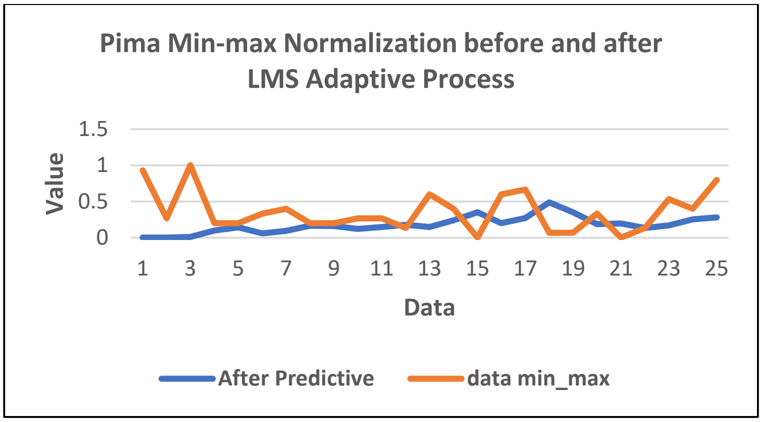

Figure 8. Normalization Results (blue) and Adaptive LMS Results (red) for Pima data.

Even though Ecoli0_vs_1 had undergone a slow convergence process (Fig.5), but in the 26th data, it looked close to the original data. Glass0 was more predictable (Fig. 6). The pattern of the prediction results closed to the min-max data since 16th iteration, although it looked like there were some delays. The Haberman dataset is closed to the min-max value but there are quite large oscillations (Fig.7). Therefore, it required more data to achieve convergence. The Pima predictive data was not close to the real data (Fig.8). The error was calculated from the four normalized and predicted normalized data in Table 2.

Table 2. Error Percentage of Ecoli0_vs_1, Glass0, Haberman and Pima Datasets

Dataset	Error Average	Procent Error
Ecoli0_vs_1	0,171166935	17%
Glass0	0,174309234	17%
Haberman	0,280025054	28%
Pima	0,251379461	25 %

Table. 2, it was clear that Haberman dataset was the data with the highest error rate. Meanwhile, the Pima dataset still had a high error rate. The other two datasets, Ecoli0_vs_1 and Glass0 have relatively the same error. From these results, we cannot immediately decide that the Haberman and Pima datasets should be discarded because they are not suitable for processing. It must be seen how much uncertainty and information value contained in the datasets first. Here the role of FIS appears to determine uncertainty and information value. Subchapter 4.3 described the FIS algorithm applied in each dataset.

4.3 The Result of Fuzzy Intuitive Sets

The simplest model of decision-making assumes just one of two relevant variants. There was α and β. From Eq. (10) the fuzzy set is, $\underline{A} = \{\alpha/\mu(\alpha); \beta/\mu(\beta), \mu(\alpha) + \mu(\beta) \leq 1\}$. where rate of plausibility $\mu(\alpha), \mu(\beta) \in [0,1]$. First, it will be expressed in a mathematical statement as follows:

α :"The first data, that was the data that would be excluded" were declared valid as $\mu(\alpha); \nu(\alpha)$

β_1:"The second data, that is the other data that would not be excluded," were declared valid as $\mu(\beta_1); \nu(\beta_1)$

β_2:" The third data, that is the other data that would not be excluded too" were declared valid as $\mu(\beta_2); \nu(\beta_2)$

The value "$\mu(\alpha)$"will be replaced with the prediction error of each dataset. The results using Fuzzy Intuitive Sets can be seen in Table 3.

Table 3. Uncertainty value and information value of some datasets

Equation	Uncertainty H(α)				Information *I(F)*			
	Ecoli0_vs_1	Glass0	Pima	Haberman	Ecoli0_vs_1	Glass0	Pima	Haberman
Eq.15	2,14017363	2,14017363	2,21434177	2,2824884	-	-	-	-
Eq. 20	1,27023139	1,27023139	1,23616025	1,2740023	-	-	-	-
Eq. 25	-	-	-	-	-0,19857322	-0,19857322	-0,22007	-0,196194049
Eq. 26	2,14017363	2,14017363	2,21434177	2,2824884	-	-	-	-
Eq. 27	-	-	-	-	-0,28660879	-0,28660879	-0,26189	-0,239170525
Eq. 16 and Eq. 26	1,08	1,08	1,4	1,52	-	-	-	-
Eq. 16 and Eq.27	-	-	-	-	-0,64	-0,64	-0,53333	-0,493333333
Eq. 15	0,65770478	0,65770478	0,81127812	0,8554508	-0,34229522	-0,34229522	-0,188722	-0,14454919
conclusion	1,45765669	1,45765669	1,57522439	1,642886	-0,4	-0,4	-0,3	-0,3

Table 3 was obtained from the fuzzy intuitive set formulas in chapter 3. Each calculation result was accompanied by the equation used. Ecoli0_vs_1 and Glass0 had the same uncertainty value and information value because the predictive error value were 0.171166935 and 0.174309234, respectively, then rounded to 0.17, and set the value $\mu(\alpha)$ as 0,17.

Uncertainty tolerance value H(α) was $0 \leq \text{H}(\alpha) \leq 1{,}5$, while the tolerance value for information I(F) was $-1 \leq \text{I(F)} \leq 1$. The value of I(F) was positive and negative depending on the range of $\mu(\alpha)$. From Eq. 13, Eq.25 and Eq. 27 the value of Information was positive for $0 \leq \mu(\alpha) \leq 0{,}5$ and negative for $0{,}5 \leq \mu(\alpha) \leq 1$. Because the value of $\mu(\alpha)$ in this paper was less than 0,5 so that the value of information was always negative. Information values equal to -0.5 and +0.5 of course, indicated the existence of significant information. However, such a large value was quite difficult to achieve for imbalanced data.

The results obtained in Table 3 showed that the value of H(α) for the Ecoli0_vs_1 and Glass0 datasets could still be tolerated because the value was still less than 1.5, while H(α) for the Pima and Haberman datasets had already exceeded the tolerance value. Therefore, from the results of H(α), the imbalanced data Pima and Haberman were declared not valid for processing. For the value of Information, it turned out that none of them reached -0.5. The information values for Ecoli0_vs_1 and Glass0 were closed to -0.5. Meanwhile, the information values for Pima and Haberman were both in the range of -0.3. These values were obtained after rounding to determine the proximity to -0.5. So, the Pima and Haberman imbalanced datasets were declared not valid for processing. Ecoli0_vs_1 and Glass0 were feasible to be processed and could be used for decision making.

To confirm the validity of the conclusions of this research, the accuracy of each dataset on the previous journal will be shown. The previous research had been carried out to calculate the accuracy of imbalanced data taken from the KEEL repository. A total of forty-four (44) data were studied, in which there were datasets Ecoli0_vs_1, Glass0, Haberman, and Pima. The results obtained can be explained in Table 4. [24]. That research studied the simple principle of Data Gravitation Classification (DGC), which classifies data samples by comparing the gravitation between different classes. However, the calculation of gravitation was not a trivial problem due to the different relevance of data attributes for distance computation, the presence of noise or irrelevant attributes, and the class imbalance problem.

Table 4. AUC (Area Under Curve) Results for Imbalanced Datasets

Dataset	DGC+ None	DGC None	ADAC2 CS	NN CS	CSVM CS	C 4.5 CS	C 4.5 RUS	C 4.5 SMT	C 4.5 SMT-TL
Ecoli0_vs_1	0,9799	0,9642	0,9692	0,9796	0,9671	0,9832	0,9796	0,9832	0,9761
Glass0	0,865	0,8553	0,8101	0,6792	0,5074	0,8212	0,8206	0,7754	0,8039
Haberman	0,6213	0,5062	0,5604	0,6245	0,5382	0,5752	0,6423	0,6539	0,6203
Pima	0,7394	0,5274	0,7114	0,7175	0,7289	0,7125	0,7235	0,7134	0,6948

Table 4. contains some of the data from previous studies. Haberman had the lowest average of AUC, so that Haberman had the lowest accuracy. These results were corroborated with the results of this research that Haberman had the highest uncertainty value, so the accuracy was low.

Table5. showed different methods in finding the accuracy of a dataset. It used the Wilcoxon rank-sum test for AUC.

Table 5. Results of the Wilcoxon rank-sum test for AUC

Dataset	DGC+ None	DGC None	ADAC2 CS	NN CS	CSVM CS	C 4.5 CS	C 4.5 RUS	C 4.5 SMT	C 4.5 SMT-TL
Ecoli0_vs_1	0,9652	0,9436	0,9305	0,9598	0,9479	0,9695	0,9598	0,9695	0,9504
Glass0	0,7037	0,7248	0,5812	0,2941	0,0181	0,5942	0,5942	0,5113	0,5431
Haberman	0,1977	0,0182	0,0946	0,2399	0,0840	0,1110	0,2614	0,2627	0,1787
Pima	0,4504	0,0682	0,3878	0,3943	0,4551	0,3976	0,4226	0,4064	0,3486

From Table 5, it could be read that the accuracy for the Haberman data was the lowest, while Ecoli0_vs_1 had the highest average accuracy. This confirmed that the method presented in this paper, method that combined between Min-max Normalization, LMS and Fuzzy Intuitive set was the valid method. This method was also easy to use and robust.

Then, after finding an invalid dataset, what can be done if only invalid data was owned. The answer can be found in sub-chapter 4.4.

4.4. Solutions offered to anticipate data with high uncertainty and low information

If the only available data was not feasible, yet the circumstances demanded that the data be processed, there was still a solution. One of the solutions was to carry out the data cleaning process. [25][26][27]. The solution was to clean the data. Showing that data cleaning can reduce uncertainty and increase information, this section showed the results of a simple data cleaning process on not feasible data. The process was performed by deleting data that had a sharp increase or decreas because that is one of the reasons for the imbalance dataset. Although the results were less than optimal, they were sufficient to reduce uncertainty and increase the information. Fig.9 and Fig.10 were Pima data and its predictions before and after the data reduction process.

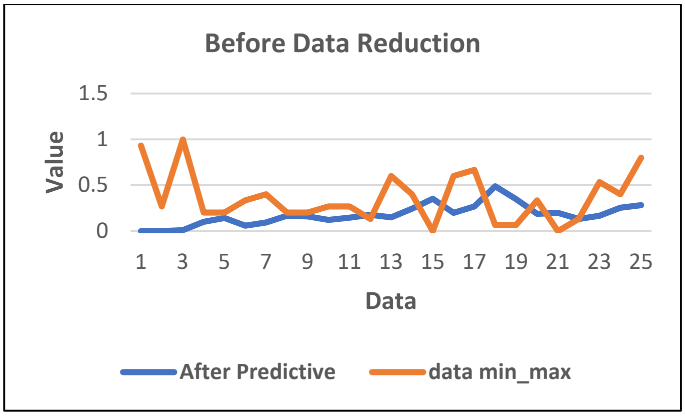

Figure 9. Pima data before data reduction

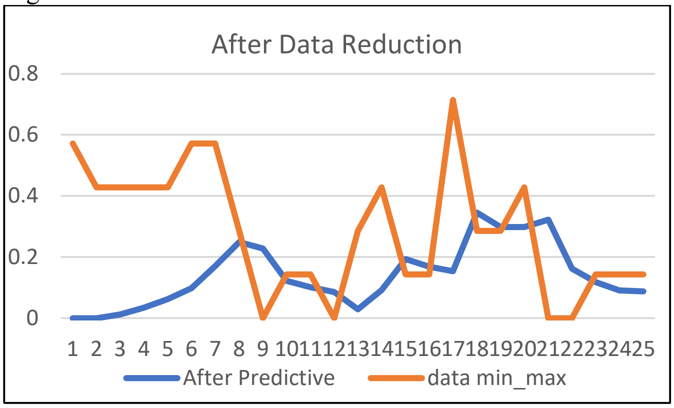

Figure 10. Pima data after data reduction

In this way, the error was reduced, although only slightly (about 3%). However, it was sufficient to show that data cleaning could be a solution if only available data were not feasible data. In Fig.10, the pattern of the predicted results was close to the pattern of the original data, although it is still not effective. Hopefully, by employing a well-researched algorithm, the results were more effective.

4.5. Applied for another imbalanced data

It took Covid-19 data in Indonesia, which consisted of data involving people who were infected, deceased, and recovered from the virus because the Covid-19 pandemic is an ongoing crisis and is colloquially known as the corona virus pandemic. Because there is no specific treatment protocol available for this viral infection, social distancing is considered as one of the remedies to prevent the infection [28]. The number of covid data was constantly changing, determined by human activity and a variety of other factors. There could be a substantial increase at one point and a significant decline at another point. As a result of this circumstance, the covid dataset became imbalanced.

The Indonesian population that is infected, deceased and recovered from Covid-19 can be shown in Fig.11 [29]. This data was taken from Indonesia *Coronavirus Information and Stats* from March to May 2020. The data was relatively small but sufficiently representative to predict other data for similar cases.

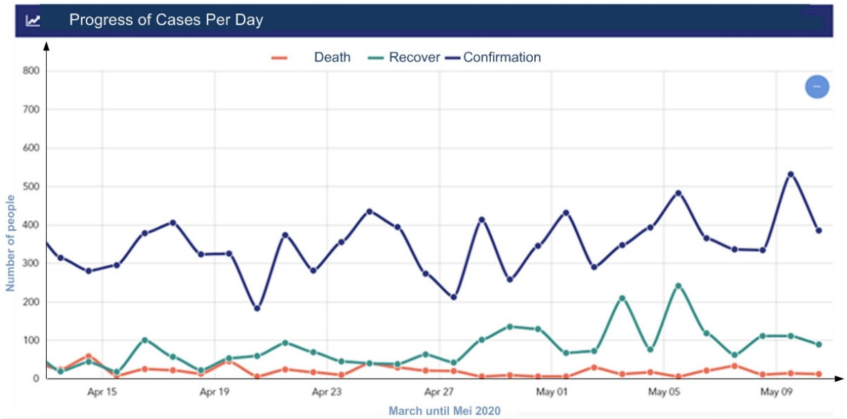

Figure 11. Comparison of infected (confirmation), dead and recovered cases

The difference between the minimum and maximum numbers is significant. It's still challenging to process such data because it's imbalanced. However, the data pattern may be noticed, with death data being the smallest. The rise in infected data was not accompanied by a rise in mortality records. Because infected data, the highest one, will be split into recovered and death data. It was still difficult to make a decision in the situation. The most crucial phase has become normalization.

4.5.1 Application of the Min-Max Algorithm to Covid-19 data

Fig. 12 showed the results of the min-max normalization algorithm for infected, death and recovered data. The pattern before and after min-max was similar, only the scale was changed. This showed that the normalization was performed correctly.

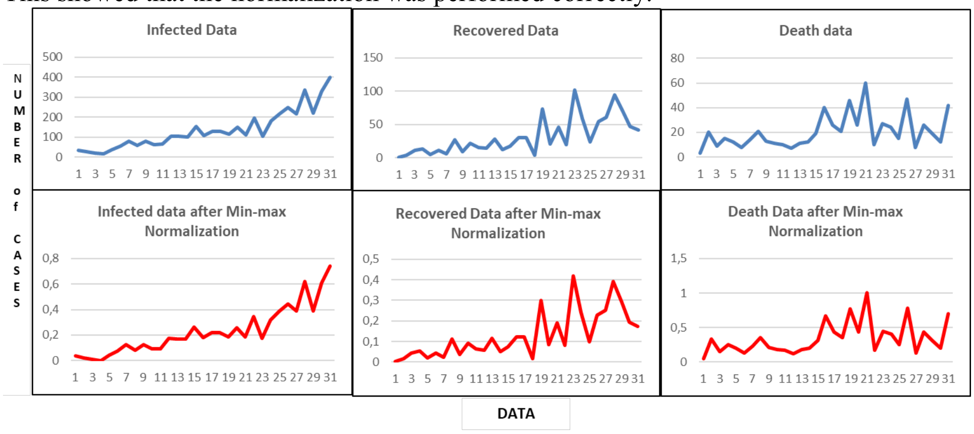

Figure 12. Infected, recovered, dan death data before and after min-max normalization

Then would be processed using LMS adaptive to examine the errors that occur from the outputs compared to the real conditions.

4.5.2 The Result of LMS Adaptive

An Adaptive LMS process was carried out for normalized data in Fig.13. Meanwhile, Fig. 14 shows infected data after normalization and LMS predictive.

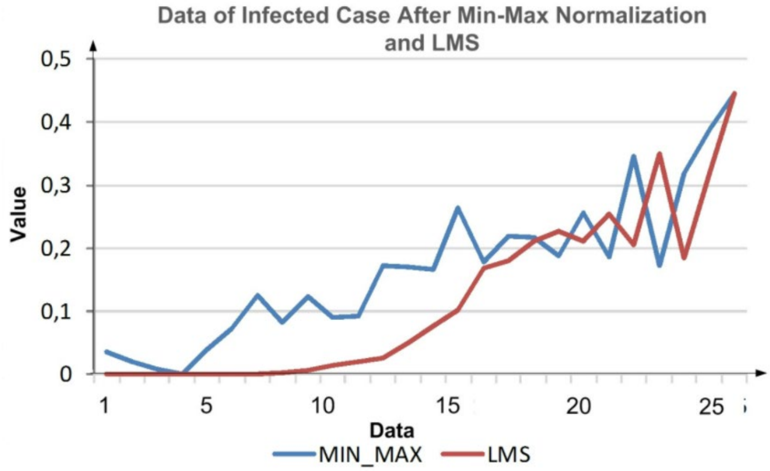

Figure 13. Normalization Results (blue) and Adaptive LMS Results (red) for Infected data.
Even though in the 26th data, the normalized and predicted normalization data appeared similar, the error analysis must still be performed to ensure the validity of the data. The process was also carried out for data "Died due to Covid-19" and "Recovered from Covid-19" data. The results are shown in Fig.15 and Fig.16.

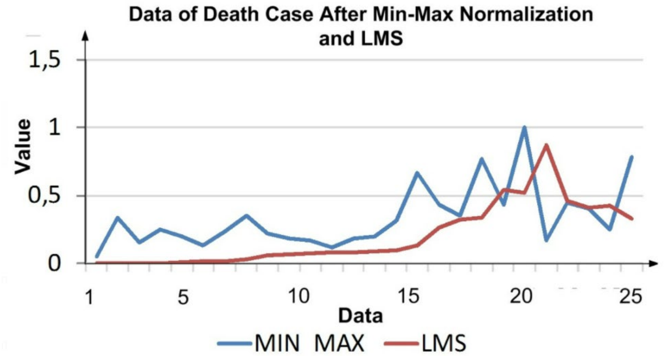

Figure 14. Normalization Results (blue) and Adaptive LMS Results (red) for death data.

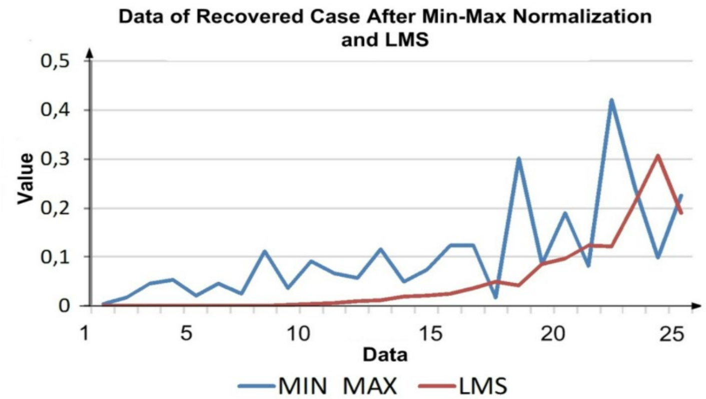

Figure 15. Normalization Results (blue) and Adaptive LMS Results (red) for recovered data.
Infected data (Fig. 13) was more predictable, it looked at the 16th data (16th iteration). The pattern of the results of LMS process closed to the min-max data, although it looked like there were some delays.
Death data (Fig.14) was difficult to predict because of the oscillations until the twenty-fifth iteration. From the initial iteration until the twentieth iteration, it seems that there was no result

that closed to the min-max result. The process towards convergence appeared in the twenty-first iteration.
The recovered data (Fig.15) closed to the min-max value but there were quite large oscillations. So, it required more data to be convergence. The error was shown in Table 6.

Table 6. Error Percentage of Infected, Death and Recovered Data

Data	Error Prediction (Percentage)
Infected	8 %
Dead	21 %
Recovered	7 %

4.5.3 The Result of Fuzzy Intuitive Sets
There were errors in the calculation of infected, death, and recovered data in the previous section. The error percentage is shown in Table 6.
Evidently, the "Death" data, had the biggest error, "Infected" and "Death" data still had error measurement less than 10 %. To know its validity, that the data showing the number of deceased people can be excluded or death data was invalid data, the FIS can be used to make the decision. The simplest model of decision-making assumes just one of two relevant variants. There was α and β. From Eq. (10) the fuzzy set is, $\underline{A} = \{\alpha/\mu(\alpha); \beta/\mu(\beta), \mu(\alpha) + \mu(\beta) \leq 1\}$. where rate of plausibility $\mu(\alpha), \mu(\beta) \in [0,1]$.
First, it will be expressed in a mathematical statement as follows:
α :"Dead" data were declared valid (could be used) $\mu(\alpha)$; $\nu(\alpha)$
β_1:"Infected" data were declared valid (could be used) $\mu(\beta_1); \nu(\beta_1)$
β_2:"Recovered" data were declared valid (could be used) $\mu(\beta_2); \nu(\beta_2)$
The "death" data had the biggest error. Therefore, α declaring that it is valid with μ(α) estimation should be excluded. Then $\beta = \beta_1 \mathrm{v}\, \beta_2$.
Because error values greater than 20% was not accurate, it can be agreed that the "Dead" data could not be used. The estimation of uncertainty is realized due to α. Because α should be mainly excluded from the possibility of data correction, therefore $\beta = \beta_1 \mathrm{v} \beta_2$. Table 7 showed the result of FIS on Covid-19 Imbalanced Data.
Table 7. Uncertainty value and information value of death data

Equation	Uncertainty H(α)	Information *I*(F)
Eq. 15	2,072214588	
Eq. 20	1,156779649	
Eq. 25		-0,270153301
Eq. 26	2,072214588	
Eq. 27		-0,309261804
Eq. 16 and Eq. 26	1,2	
Eq. 16 and Eq. 27		-0,6
Eq. 15	0,721928095	- 0,278071905
Conclusion	1,444627384	-0,364371753

It could be seen that the uncertainty of α is quite significant, which is more than 0.70. In addition, the information of α is very small, because it did not reach -0,5. So that, the "dead" data could not be used because it was invalid data.

In June 2021, this result was processed. Following reports from the Ministry of Communications and Information Technology in August 2021 that erroneous death data had been discovered, the government has resolved to rectify the data [30].

5. Conclusion

By combining the min-max normalization, adaptive LMS and FIS algorithm, invalid data could be detected and would be decided to be excluded and could not be processed for decision making. Data sets were taken from the KEEL repository. It can be concluded that Haberman and Pima datasets had a high uncertainty, 1,64 and 1,58 respectively, with the information value -0,3 for both Haberman and Pima data sets. Meanwhile, Ecoli0_vs_1 and Grass0 had the same uncertainty and information value, there were 1,46 and -0,4, respectively. Therefore, the datasets Ecoli0_vs_1 and Glass0 were more suitable for processing as decision making than Haberman and Pima datasets. To prove the truth of the results, the results were compared with the previous studies regarding the accuracy of the dataset, especially for Ecoli0_vs_1, Glass0, Haberman and Pima. It showed that Haberman and Pima datasets had the lowest accuracy.

However, the Haberman and Pima datasets can still be processed using data cleaning to reduce the uncertainty value and increase the information value if needed. Although the result only lowered error by 3%, it was enough to demonstrate that data cleaning might be a viable option if only not feasible data was the only available data. This research could be utilized in the future to perform data augmentation on prediction or classification systems to increase the results' accuracy.

This method was not only applied to public data but also applied to Covid-19 data in Indonesia as a private data. The death data had an uncertainty value around 1,4 and an information value around -0,4, so that the death data was not valid. The death data results, obtained around June 2021, had been confirmed by an announcement from the Indonesian government in August 2021, that the death data was not valid and should be corrected.

References

[1] Y. Lu, Y. M. Cheung, and Y. Y. Tang, "Bayes Imbalance Impact Index: A Measure of Class Imbalanced Data Set for Classification Problem," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 9, pp. 3525–3539, 2020, doi: 10.1109/TNNLS.2019.2944962.

[2] S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat, "Variance Ranking Attributes Selection Techniques for Binary Classification Problem in Imbalance Data," *IEEE Access*, vol. 7, pp. 24649–24666, 2019, doi: 10.1109/ACCESS.2019.2899578.

[3] Rahul Varma, "Statistical analysis on a dataset you don't understand," *Towar. Data Sci.*, 2020, [Online]. Available: https://towardsdatascience.com/statistical-analysis-on-a-dataset-you-dont-understand-f382f43c8fa5.

[4] B. Timothy C. Heeren, PhD, Professor of Biostastics, Jacqueline N. Milton, PhD, Clinical Assistant Professor, "Basic Statistical Analysis Using the R Statistical Package," *Bost. Univ. Sch. Public Heal.*, [Online]. Available: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R-Manual/R-Manual_print.html.

[5] A. Flores and B. Widrow, "Assesment of the efficiency of the LMS algorithm based on spectral information," *Conf. Rec. - Asilomar Conf. Signals, Syst. Comput.*, vol. 1, no. 1, pp. 120–124, 2004, doi: 10.1109/acssc.2004.1399102.

[6] A. Ren, Z. Du, J. Li, F. Hu, X. Yang, and H. Abbas, "Adaptive interference cancellation of ECG signals," *Sensors (Switzerland)*, vol. 17, no. 5, pp. 1–15, 2017, doi: 10.3390/s17050942.

[7] S. A. D. Prasetyowati and A. Susanto, "Multiple Processes for Least Mean Square Adaptive Algorithm on Roadway Noise Cancelling," *Int. J. Electr. Comput. Eng.*, vol. 5, no. 2, pp. 355–360, 2015, [Online]. Available: http://ijece.iaescore.com/index.php/IJECE/article/view/5704/4765.

[8] S. A. D. Prasetyowati, H. Abbas, and N. B. S. Eka, "Solution for Vehicles Noise Cancellation With Modification of LMS Adaptive Algorithm," *Int. J. Comput. Sci. Eng. (IJCSE*, vol. 4, no. 5, pp. 929–937, 2012, [Online]. Available: https://www.researchgate.net/publication/266462670_Solution_for_Vehicles_Noise_Cancellation_With_Modification_of_LMS_Adaptive_Algorithm.

[9] KEEL Knowledge Extraction based on Evolutionary Learning, "Sets, Standard Classification data." https://sci2s.ugr.es/keel/category.php?cat=clas.

[10] J. Kulicka, "The fuzzy intuitive sets in the decision-making," no. May, 2015, doi: 10.13140/RG.2.1.3608.4646.

[11] K. T. Atanassov, *On intuitionistic fuzzy sets theory*, vol. 283. 2012.

[12] Y. Park and J. Ghosh, "Ensembles of α-trees for imbalanced classification problems," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 131–143, 2014, doi: 10.1109/TKDE.2012.255.

[13] C. L. Liu and P. Y. Hsieh, "Model-Based Synthetic Sampling for Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1543–1556, 2020, doi: 10.1109/TKDE.2019.2905559.

[14] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 39, no. 2, pp. 539–550, 2009, doi: 10.1109/TSMCB.2008.2007853.

[15] Z. Kong, L. Wang, and Z. Wu, "Application of fuzzy soft set in decision making problems based on grey theory," *J. Comput. Appl. Math.*, vol. 236, no. 6, pp. 1521–1530, 2011, doi: 10.1016/j.cam.2011.09.016.

[16] K. Atanassov and P. Vassilev, "Intuitionistic fuzzy sets and other fuzzy sets extensions

representable by them," *J. Intell. Fuzzy Syst.*, vol. 38, no. 1, pp. 525–530, 2020, doi: 10.3233/JIFS-179426.
[17] K. Atanassov, G. Pasi, and R. Yager, "Intuitionistic fuzzy interpretations of multi-criteria multi-person and multi-measurement tool decision making," *Int. J. Syst. Sci.*, vol. 36, no. 14, pp. 859–868, 2005, doi: 10.1080/00207720500382365.
[18] R. Parvathi, C. Malathi, M. Akram, and K. T. Atanassov, "Intuitionistic fuzzy linear regression analysis," *Fuzzy Optim. Decis. Mak.*, vol. 12, no. 2, pp. 215–229, 2013, doi: 10.1007/s10700-012-9150-9.
[19] P. Y. Ekel, "Fuzzy sets and models of decision making," *Comput. Math. with Appl.*, vol. 44, no. 7, pp. 863–875, Oct. 2002, doi: 10.1016/S0898-1221(02)00199-2.
[20] P. Nair and I. Kashyap, "Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier," *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019*, pp. 460–464, 2019, doi: 10.1109/COMITCon.2019.8862250.
[21] J. R. Zeidler, "Performance Analysis of LMS Adaptive Prediction Filters," *Proc. IEEE*, vol. 78, no. 12, pp. 1781–1806, 1990, doi: 10.1109/5.60921.
[22] S. Gazor, "Prediction in LMS-type adaptive algorithms for smoothly time varying environments," *IEEE Trans. Signal Process.*, vol. 47, no. 6, pp. 1735–1739, 1999, doi: 10.1109/78.765152.
[23] S. Prasetyowati, "Hybrid FLC-LMS Algorithm for Predicting Sediment Volume in the River," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 2, pp. 395–409, 2021, doi: 10.22266/ijies2021.0430.36.
[24] A. Cano, A. Zafra, S. Ventura, and S. Member, "Weighted Data Gravitation Classification for Standard and Imbalanced Data Alberto," vol. 43, no. 6, pp. 1–16, 2012.
[25] B. R. J, "Cognito - Intuitive Auto Data Exploratory Toolkit," pp. 1–6, 2020.
[26] Z. Lv, "A Data Fusion and Data Cleaning System for Smart Grids Big Data," pp. 802–807, 2019, doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00119.
[27] V. Kumar and C. Khosla, "Data Cleaning – A thorough analysis and survey on Unstructured data," pp. 305–309, 2018.
[28] A. Almudhi, "Covid_1.pdf," *Saudi J. Biol. Sci. homepage www.sciencedirect.com*, vol. 28, pp. 3110–3116, 2021, [Online]. Available: https://reader.elsevier.com/reader/sd/pii/.
[29] I. C. I. and stat Virus, "Indonesia Coronavirus Information and Stats," Indonesia, 2020. [Online]. Available: https://corona.help/country/indonesia#charts-nav.
[30] B. Kominfo, "Menkominfo: Indikator Angka Kematian Covid-19 Tak Dihapus Tapi Dirapikan," *website*, 2021. https://www.kominfo.go.id/content/detail/36319/menkominfo-indikator-angka-kematian-covid-19-tak-dihapus-tapi-dirapikan/0/berita_satker (accessed Aug. 12, 2021).

Sri Arttini Dwi Prasetyowati received a Bachelor's degree (1989) and a Master's degree (1998) from the Department of Mathematics at Universitas Gadjah Mada, Indonesia. Doctoral Degree (2010) was received from Electrical Engineering Universitas Gadjah Mada Indonesia. She is lecturer at Electrical Engineering, Universitas Islam Sultan Agung, Indonesia. Research Interest: Signal Processing and data mining.

Munaf Ismail received a Bachelor’s degree (2003) from Electrical Engineering Universitas Diponegoro and a Master’s degree (2015) from Electrical Engineering Universitas Islam Sultan Agung. He is lecturer at Electrical Engineering, Universitas Islam Sultan Agung, Indonesia. Research interest: Sensors, Classification, Optimization.

Eka Nuryanto Budisusila received a Bachelor’s degree (1997) from Universitas Muhammadiyah Malang (UMM) in major of electrical engineering. And he received a Master degree (2007) in Electrical Enginering of Universitas Brawijaya Malang. And in recent year he takes the doctoral program of Electrical Engineering in Universitas Sriwijaya Palembang. He is a lecturer of Electrical Engineering of Universitas Islam Sultan Agung, Indonesia. He concerns in sensors and actuators, medical instrumentation, and digital image processing.

De Rosal Ignatius Moses Setiadi, currently works as a lecturer at the Department of Informatics Engineering and Doctor Candidate at the Dian Nuswantoro University, Indonesia. He received his master's degree in Informatics Engineering from Dian Nuswantoro University, Indonesia in 2012. He received his bachelor's degree in Informatics Engineering was obtained at the Catholic University Soegijapranata, Indonesia in 2010. Research interest is in image processing, machine learning, and data mining.

Mauridhi Hery Purnomo (Senior Member, IEEE) received the B.S. degree from the Electrical Engineering Institut Teknologi Sepuluh Nopember (ITS), Surabaya in 1984, the M.S. degree in 1995 and the Ph.D degree from the Electrical Engineering Osaka City University, Japan in 1998. He is a Professor with the ITS and involved in teaching on research philosophy, artificial intelligence, neural network, and image processing. His research interests include smart grid, renewable energy, artificial intelligent application on healthcare and power systems. He is the Chair of the IEEE Industrial Electronics Society Indonesia Section.